Identification of Overlapping Features in Time Series Data

by

Shilpy Sharma

A Thesis
Presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Doctor of Philosophy
in
Computer Science

Guelph, Ontario, Canada
© Shilpy Sharma, August, 2015
Abstract

Identification of Overlapping Features in Time Series Data

Shilpy Sharma
University of Guelph, 2015

Advisor:
Dr. Charlie Obimbo
Dr. David Swayne

Time series data can be defined as sample observed sequentially over time and occurs in every field such as finance, stock, bioinformatics and many more. Often time series analysis is the problem of trying to differentiate and extract meaningful statistical patterns that changes in a logical way. It is important to learn the different overlapping patterns in the long term and short term time series data. Conventional methods are based on analysing time series data using only one time series analysis technique.

The principal aim of this Thesis is to examine the concurrent overlapping features of time-series data using an improved version of three important time-series analysis techniques such as Locally Weighted Scatterplot Smoothing (LOWESS and the related algorithm LOESS), change point, trend, and control chart patterns. All of the algorithms are tested on real data and data which have been computer-generated.

We first examine smoothing techniques for visualization techniques such as LO(W)ESS. Various techniques are compared for their utility and, for LO(W)ESS, we examine and extend automatic means for smoothing and for outlier. We next analyze the masking effects of change points which can disguise themselves as false trends. By looking at
the possible overlap of both phenomena it is possible to obtain a composite picture of the time series when both change point and trend are present. Control charts patterns can be associated with certain assignable causes and recognition of such patterns can accelerate the frequently exhibit variations. Each pattern has special statistical characteristics which differentiate one pattern from another. In our simulations, and in real data, we illustrate that presence of more than one pattern may exist and identification of concurrent pattern is important. Finally, we demonstrate that, for the optimal choice of LO(W)ESS smoothing parameter, attention must be paid to the possibility of a local minimum-maximum error property, and that this focus on the best choice of smoothing parameter may prescribe introducing artificial change points.
Acknowledgments

First and foremost, I want to thank my advisors, Dr. Charlie Obimbo and Dr. David Swayne, for their patience and expert guidance during the course of my graduate studies. They are very enthusiastic and knowledgeable, especially Dr. Swayne, as he guided me all the time during the research and helped me with the writing of the Thesis.

I would like to thank the Heidelberg University for permitting me to access the data for my research analysis.

Lastly, I would like to thank my family for all their love and encouragement. For my parents, Mohinder Paul Sharma and Vijay Sharma, for giving birth to me, and my in-laws for all further support during my life. And most of all for my loving, supportive, encouraging, and patient husband Sachin for continued support during this Ph.D. is so appreciated. My gratitude also goes to my daughter, Sarah.
## Contents

List of Tables vii  
List of Figures viii  
List of Symbols x  
List of Abbreviations xi  

1 Introduction 1  
1.1 Scope and Motivation 4  
1.2 Outline 15  
1.3 Problem Statement 18  
1.4 Aims and Objectives 20  
1.5 Data for Methodology 21  
1.6 Contributions 22  
1.7 Thesis Structure 23  

2 Related Work 25  
2.1 LOWESS 26  
2.1.1 Selection of $\lambda$, $W$, $t$ and $\alpha$ 27  
2.1.2 Computational Details 28  
2.1.3 Automatic Selection of $\alpha$ and $\lambda$ 29  
2.1.4 Limitations in Existing Systems 31  
2.2 Control Charts 32  
2.2.1 Feature Extraction Techniques 38  
2.2.2 Limitation in Existing Systems 40  
2.3 Change point detection and Trend analysis 41  
2.3.1 Change point 41  
2.3.2 Trend detection 46  
2.3.3 Limitations in Existing Systems 48  
2.4 Summary 49  

3 LO(W)ESS 50  
3.1 Locally Weighted Scatterplot Smoothing 50  
3.2 Research Objective 52  
3.2.1 Proposed Methodology 52  
3.2.2 Experimental Design 56
## List of Tables

<table>
<thead>
<tr>
<th></th>
<th>Table Number</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Parameters for Control Chart Patterns</td>
<td>35</td>
</tr>
<tr>
<td>2.2</td>
<td>Basic Questionnaire for CCPs [38]</td>
<td>36</td>
</tr>
<tr>
<td>2.3</td>
<td>Parameters for Change point</td>
<td>42</td>
</tr>
<tr>
<td>3.1</td>
<td>Correlation Coefficient</td>
<td>55</td>
</tr>
<tr>
<td>3.2</td>
<td>Robust AIC score for different values of α</td>
<td>58</td>
</tr>
<tr>
<td>3.3</td>
<td>Best selected model using Akaike Weights</td>
<td>59</td>
</tr>
<tr>
<td>3.4</td>
<td>Robust AIC score for different values of α</td>
<td>61</td>
</tr>
<tr>
<td>3.5</td>
<td>Best selected model using Akaike Weights</td>
<td>61</td>
</tr>
<tr>
<td>4.1</td>
<td>Change in Mean</td>
<td>68</td>
</tr>
<tr>
<td>4.2</td>
<td>Step data with constant mean and variance</td>
<td>70</td>
</tr>
<tr>
<td>4.3</td>
<td>Change in variance</td>
<td>72</td>
</tr>
<tr>
<td>4.4</td>
<td>Sine Wave data</td>
<td>74</td>
</tr>
<tr>
<td>4.5</td>
<td>Non AR</td>
<td>76</td>
</tr>
<tr>
<td>4.6</td>
<td>Summary: Trend detection</td>
<td>77</td>
</tr>
<tr>
<td>4.7</td>
<td>Experimental Procedure</td>
<td>81</td>
</tr>
<tr>
<td>4.8</td>
<td>Summary: Trend detection</td>
<td>86</td>
</tr>
<tr>
<td>5.1</td>
<td>Parameters for Control Chart Patterns</td>
<td>95</td>
</tr>
<tr>
<td>5.2</td>
<td>Concurrent Control Chart Patterns</td>
<td>102</td>
</tr>
<tr>
<td>5.3</td>
<td>Pattern Identification, percent success from 100 experiments</td>
<td>105</td>
</tr>
<tr>
<td>6.1</td>
<td>Percentage of points around the centre with weight function close to unity</td>
<td>119</td>
</tr>
<tr>
<td>A.1</td>
<td>Detail of Data</td>
<td>135</td>
</tr>
<tr>
<td>B.1</td>
<td>Parameters for Control Chart Patterns</td>
<td>136</td>
</tr>
<tr>
<td>B.2</td>
<td>Parameters for Change Point</td>
<td>137</td>
</tr>
</tbody>
</table>
# List of Figures

1.1 WHO report: Cholera cases from contaminated water (1989-2013)\textsuperscript{[68]} \hspace{1em} 2
1.2 Presidential Elections 2012 \textsuperscript{[76]} \hspace{1em} 6
1.3 LO(W)ESS fit with different values of smoothing parameter \hspace{1em} 7
1.4 Example: Change point analysis \hspace{1em} 8
1.5 Change point detection \hspace{1em} 9
1.6 Control Chart Pattern: WildFire analysis \hspace{1em} 10
1.7 Control Chart Pattern: Flooding Analysis \hspace{1em} 11
1.8 Control Chart Pattern: Control Limits \hspace{1em} 11
1.9 Example: Daily concentration of CO\textsubscript{2} shows mixture pattern \hspace{1em} 12
1.10 Trend due to polluted plume \hspace{1em} 13
1.11 Example: Minimum maximum error property \hspace{1em} 15
1.12 Basic Control Chart Patterns \hspace{1em} 17
1.13 Structure of the Thesis \hspace{1em} 23
2.1 Control Chart \hspace{1em} 33
2.2 Mixture Patterns \hspace{1em} 38
3.1 LOWESS with different values of $\alpha$ \hspace{1em} 51
3.2 Steps for proposed methodology: LO(W)ESS \hspace{1em} 53
3.3 rLOWESS fit with $\alpha=0.3$ \hspace{1em} 58
3.4 rLOWESS fit with $\alpha=0.3$ \hspace{1em} 59
3.5 rLOWESS fit with $\alpha=0.8$ \hspace{1em} 62
4.1 Change point detection \hspace{1em} 67
4.2 Change point detection \hspace{1em} 69
4.3 Change point detection \hspace{1em} 71
4.4 Change point detection \hspace{1em} 73
4.5 Change point detection \hspace{1em} 75
4.6 Trend Analysis: MK \hspace{1em} 79
4.7 Hidden Change Point and Trend \hspace{1em} 80
4.8 Change Point location $\delta = 251$ \hspace{1em} 81
4.9 Trend for complete series \hspace{1em} 82
4.10 Trend analysis before and after change point \hspace{1em} 82
4.11 Combined Trend analysis before and after change point from Figure 4.10 \hspace{1em} 83
4.12 Change points using $wbs$ and $E$-divisive \hspace{1em} 84
4.13 Mann-Kendall \hspace{1em} 85
4.14 Change points using wbs on simulated dataset ................. 87
4.15 Change point, Trend and Piecewise Trend ....................... 88

5.1 Normal Pattern ................................................. 91
5.2 Shift up and Shift down Pattern .................................. 91
5.3 Stratification Pattern ............................................. 92
5.4 Cyclic Pattern .................................................. 93
5.5 Systematic Pattern .............................................. 93
5.6 Trend up and Trend down Pattern .................................. 94
5.7 Proposed Model: Concurrent Pattern Recognition ................... 95
5.8 ALSPI values for different patterns ............................... 100
5.9 Decision Tree .................................................. 102
5.10 Mixed patterns separated using RobustICA ....................... 103

6.1 Minimax property ................................................ 109
6.2 Minimax property: Different $\alpha$ values .......................... 110
6.3 Example: Minimax property with equal and opposite sign ......... 112
6.4 Example: Minimax property with equal and opposite sign ......... 112
6.5 Change point masking trend and locations of change points ........ 113
6.6 Example: Minimax property with equal and opposite sign ......... 113
6.7 Example: Minimax property with equal and opposite sign ......... 114
6.8 Pattern detection using change point, LO(W)ESS, and control chart pattern ...................................................... 116
# List of Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>LOWESS Polynomial Order</td>
<td>27</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>LOWESS Smoothing Parameter</td>
<td>27</td>
</tr>
<tr>
<td>$W$</td>
<td>Weight Function</td>
<td>27</td>
</tr>
<tr>
<td>$t$</td>
<td>Iteration Count</td>
<td>27</td>
</tr>
<tr>
<td>$w(x)$</td>
<td>Tricube Weight Function</td>
<td>28</td>
</tr>
<tr>
<td>$\delta_i$</td>
<td>Robust Weight Function</td>
<td>29</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>Standard Deviation</td>
<td>32</td>
</tr>
<tr>
<td>$S$</td>
<td>Mann-Kendall Statistics</td>
<td>47</td>
</tr>
<tr>
<td>$T$</td>
<td>Cox-Stuart Statistics</td>
<td>48</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Mean of sample</td>
<td>66</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>Variance of sample</td>
<td>66</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>Added noise</td>
<td>66</td>
</tr>
</tbody>
</table>
List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
<td>2</td>
</tr>
<tr>
<td>GEMS</td>
<td>Global Environment Monitoring System</td>
<td>2</td>
</tr>
<tr>
<td>LOWESS</td>
<td>Locally Weighted Scatterplot Smoothing</td>
<td>4</td>
</tr>
<tr>
<td>WMO</td>
<td>World Meteorological Organization</td>
<td>7</td>
</tr>
<tr>
<td>PMD</td>
<td>Pakistan Meteorological Department</td>
<td>10</td>
</tr>
<tr>
<td>EPA</td>
<td>Environmental Protection Agency</td>
<td>11</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike Information Criteria</td>
<td>14</td>
</tr>
<tr>
<td>LOESS</td>
<td>Locally Estimated Scatterplot Smoothing</td>
<td>15</td>
</tr>
<tr>
<td>CCP</td>
<td>Control Chart Pattern</td>
<td>16</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
<td>16</td>
</tr>
<tr>
<td>BN</td>
<td>Bayesian Network</td>
<td>16</td>
</tr>
<tr>
<td>ES</td>
<td>Expert System</td>
<td>16</td>
</tr>
<tr>
<td>DT</td>
<td>Decision Tree</td>
<td>16</td>
</tr>
<tr>
<td>GCV</td>
<td>Generalized Cross Validation</td>
<td>30</td>
</tr>
<tr>
<td>ARL</td>
<td>Average Run Length</td>
<td>36</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
<td>38</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent Component Analysis</td>
<td>38</td>
</tr>
<tr>
<td>IC</td>
<td>Independent Component</td>
<td>39</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>CUSUM</td>
<td>Cumulative Sum</td>
<td>42</td>
</tr>
<tr>
<td>WBS</td>
<td>Wild Binary Segmentation</td>
<td>44</td>
</tr>
<tr>
<td>SBS</td>
<td>Standard Binary Segmentation</td>
<td>44</td>
</tr>
<tr>
<td>BCP</td>
<td>Bayesian Analysis of Change Point</td>
<td>44</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov Chain Monte Carlo</td>
<td>44</td>
</tr>
<tr>
<td>OLS</td>
<td>Ordinary Least Square</td>
<td>46</td>
</tr>
<tr>
<td>MK</td>
<td>Mann-Kendall</td>
<td>47</td>
</tr>
<tr>
<td>rLOWESS</td>
<td>Robust Locally Weighted Scatterplot Smoothing</td>
<td>53</td>
</tr>
<tr>
<td>rLOESS</td>
<td>Robust Locally Estimated Scatterplot Smoothing</td>
<td>53</td>
</tr>
<tr>
<td>HI</td>
<td>Hample Identifier</td>
<td>53</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
<td>54</td>
</tr>
<tr>
<td>FWMC</td>
<td>Flow Weighted Mean Concentration</td>
<td>60</td>
</tr>
<tr>
<td>E – Aggl.</td>
<td>E-Agglomerative</td>
<td>65</td>
</tr>
<tr>
<td>IRD</td>
<td>Iterative Robust Detection</td>
<td>65</td>
</tr>
<tr>
<td>TPR</td>
<td>True Positive Rate</td>
<td>65</td>
</tr>
<tr>
<td>N</td>
<td>Normal</td>
<td>90</td>
</tr>
<tr>
<td>UCL</td>
<td>Upper Control Limit</td>
<td>90</td>
</tr>
<tr>
<td>LCL</td>
<td>Lower Control Limit</td>
<td>90</td>
</tr>
<tr>
<td>SD</td>
<td>Shift Down</td>
<td>90</td>
</tr>
<tr>
<td>CL</td>
<td>Center Line</td>
<td>90</td>
</tr>
<tr>
<td>SU</td>
<td>Shift Up</td>
<td>90</td>
</tr>
<tr>
<td>SR</td>
<td>Stratification</td>
<td>91</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>-------------</td>
<td>------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>C</td>
<td>Cyclic</td>
<td>92</td>
</tr>
<tr>
<td>SY</td>
<td>Systematic</td>
<td>93</td>
</tr>
<tr>
<td>TU</td>
<td>Trend Up</td>
<td>93</td>
</tr>
<tr>
<td>TD</td>
<td>Trend Down</td>
<td>93</td>
</tr>
<tr>
<td>UWL</td>
<td>Upper Warning Limit</td>
<td>109</td>
</tr>
<tr>
<td>LWL</td>
<td>Lower Warning Limit</td>
<td>109</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Water is the most critical resource issue of our lifetime and our children’s lifetime. Health of our water is the principal measure of how we live on the land.

Luna Leopold

A human body contains 60% water. Water plays an essential role in the transport of nutrient, removal of toxins, digestion and the functionality of organs [71]. Of all the environmental concerns, the lack of good quality water is of crucial concern. Every year five million people die from waterborne disease, most of these are children [85, 1]. While in developing countries shortage of water is also a serious problem, the problem of degradation of water quality is world-wide.

The quality of water is degraded by inputs such as sewage waste, factory effluent, and agricultural run-offs [1]. Water quality differs from one region to another based on seasonal, climatic changes and depending on soil type, surfaces and rocks. Water quality affects the quality of drinking water and the capacity of the water sources to support wildlife and healthy ecosystems. User contribution is significant to maintain
the quality of water.

Many organizations such as the World Health Organization (WHO), United Nations International Children’s Emergency Fund, Global Environment Monitoring System (GEMS) and Water Quality Association assist developing countries in procurement and administration of water quality information collected from rivers, watersheds, fresh water lakes, reservoirs, and ground waters.

That information is used to improve the water quality and find the state and trends of global water resources. According to WHO statistics, 783 million people do not use improved source of drinking water and approximately, 6 to 8 million deaths worldwide are due to calamities and waterborne diseases [72]. Figure 1.1 shows statistics of cholera disease occurred from bad quality of water worldwide [68].

Figure 1.1: WHO report: Cholera cases from contaminated water (1989-2013)[68]
From last few years there has been increased awareness of water pollution internationally. Broadly, it is considered that new methodologies are required for the analysis of water quality data.

The GEMS Water Programme maintains database and creates modeling tools for monitoring the improvement and calculates trends in water quality data. The focus of this Thesis is to contribute to the development of a framework for identification and classification of overlapping features in large datasets. Such a framework will greatly help in prediction and forecasting of state of quality of water or any type of time series data.

Water Quality data, a kind of time series, can be defined as samples observed over time. Often time series analysis comprises different techniques to differentiate significant statistics and other characteristics of the data. The extracted meaningful data help predicting future data based on the observed data. Time series analysis has a wide range of application area such as forecasting, projection, analysis and much more. It has two important goals:

1. Identification of different components represented by the sequence of observations such as abrupt or gradual changes, pattern detection, and seasonality.

2. Prediction about future events based on statistical properties

The different features in data are identified using different statistical methods. As an example, to maintain the quality of data, concentration of different parameters is measured over time using different analytical techniques. For example, trend analysis is performed to see if there is any increase or decrease in the parameter con-
centration, data smoothing is performed to remove the noise and identify the clear pattern, and pattern detection is performed to identify hidden patterns. This analysis helps to determine what factors from the past are currently impacting the quality of data and what would be the condition of water in the future [23, 45]. It may, or may not, be known whether a physical event has occurred when, for example, a measurement indicates some sort of disturbance or long-term change. The likelihood of corroboration of measurements with events may alter with time between the events happening and the examination of the data.

Analysing the presence of only one statistical property and ignoring other features may lead to erroneous results. For example, looking for the trend in a time series and ignoring the effect of change points leads to inaccurate interpretation.

Using a single method provides limited support for analysis and decision making process. The principal research question of interest is to propose a framework to identify overlapping features in a time series.

1.1 Scope and Motivation

Analysis of time series has grabbed the increased research interest from academic and scientific point of view. This interest has led to the development of a framework to analyze different characteristics of time series data. There are three important analytical components used for the analysis of a time series:

1. LOWESS: Locally Weighted Scatterplot Smoothing

2. Change point detection
3. Control chart patterns

Before using these techniques, it is important to know the sequential order in which they need to be used.

1. **LO(W)ESS**: It is one of the most widely used method for data smoothing and trend estimation. LO(W)ESS graphical representation helps to visualize the overall trends and identify times of changes in the time series. In 2012 presidential elections [76], LO(W)ESS fit is used to predict the presidential candidate: Obama, Romney and Gingrich. Figure 1.2 clearly shows that the winning likelihood of Obama (upper line) is more than 50% from Jan 1, 2012 and Jan 22, 2012 whereas Gingrich (lower line) and Romney (middle line) likelihood is below 50%. Moreover, Niblett [66] used LOWESS fit for the evolution of legal rules. There are many more other examples where LOWESS fit has been used for trend estimation. This is the easiest way to communicate trends especially for the non-technical people.
The smoothing parameter is one of the most important parameter for the LOWESS fit. Different value of smoothing parameter provides the different shapes of the LOWESS fit as shown in Figure 1.3. This shows that the random selection of smoothing parameter may cause the problem of over-smoothing or under-smoothing. Figure 1.3 (a) shows an example of under-smoothing whereas Figure 1.3 (d) shows an example of over-smoothing. Despite the wide application area, the values of LOWESS smoothing parameter are selected on trial and error basis. It is a biggest challenge to estimate the correct value of smoothing parameter in the presence of outliers (extreme data point from other values).
To allow the intelligent use of LOWESS fit in the smoothing and trend estimation, it is required to select the value of smoothing parameter based on the characteristics of data. This is something that has been overlooked in current methodologies.

2. Change point detection: It is a topic of interest from last 5 decades. It is used for detection of any physical change or change in statistical properties of a process. For example, World Meteorological Organization (WMO) used change point detection technique to analyze changes in the climate [27]. Based on WMO prediction for year 2014 to be the warmest year, Cahil performed a change point analysis for RealClimate to show that the warming is real [21].
The analysis was done on dataset from 1880-2014 as shown in Figure 1.4.

![Example: Change point analysis](image)

**Figure 1.4: Example: Change point analysis**

There are some other examples in the area of medical, financial, economic, sales, budgetary analysis, stock market, and many more where change point detection methods have been used to detect change in the behaviour of a process. Different change point methods are available for detection of unnatural variations in the time series data [10, 24, 26, 34, 49, 56, 64, 73]. These techniques are capable of detecting more than one change point. It is found that different methods show different numbers and locations of change points as shown in Figure 1.5. In this figure, the original data shows four change points, whereas Bayesian analysis of change point (BCP) method could only detect three change point locations. Although change point model (CPM) and ECP (ecp: R Package used to detect multiple change point) detected four change points, the locations of
change points are different. Believing in the result of one method can mislead the decision making process. It is extremely important to know which change point locations are accurate and how many change points are real change points, and whether some change points are artifacts.

![Change Point Detection](image)

*Figure 1.5: Change point detection*

It is required to compare different methods for accuracy and computational efficiency. These techniques only focus on detection of abrupt changes in the process and ignore the fact of gradual changes. Since time series can have different characteristics, focusing on detection of only one feature, abrupt change, and ignoring the presence of gradual change, can mislead the results.

3. **Control Chart Patterns**: In a time series, a parameter often undergoes changes due to some internal or external factors. For example, extreme weather
conditions such as drought, flooding, heat-waves and wildfires may lead the process to change its behaviour. Such conditions can be detected when analysing the data using control charts. Figure 1.6 shows an example of control chart pattern for wildfire [13]. Figure 1.6 shows a combination of different patterns, with more or less variations, where values lie either above or below the $3\sigma$ limit.

![Control Chart Pattern: WildFire analysis](image)

Figure 1.6: Control Chart Pattern: WildFire analysis

Similarly, in another example, the Pakistan Meteorological Department (PMD) studied the rainfall behaviour in 2010 using control charts shown in Figure 1.7 [3].
In this figure, monsoon period, July and August, is targeted to analyze the floods. If the data point goes beyond the control limit, it indicates the danger of flooding. July 2010 indicates the danger of flooding. An unnatural variation is detected by control charts if a process goes above or below the control limits as shown in Figure 1.8 [87].

Environmental Protection Agency (EPA) practices control charts approach to
detect the concentration of $CO_2$ [14]. When analyzing a long-term series, it is possible that a process can have multiple features present. For example, Figure 1.9 shows the daily concentration of $CO_2$ using control charts [14]. In this figure, from June to November the data points are fluctuating between upper control limit (UCL: shown using arrow), upper warning limit (UWL: shown using arrow), lower control limit (LCL: shown using arrow), and lower warning limit (UWL: shown using arrow). Most of the data points are above the mean line.

Figure 1.9: Example: Daily concentration of $CO_2$ shows mixture pattern

From April to beginning of May the data points are going beyond the LCL and show a decreasing trend. Again, from June to September the points are fluctuating between UCL and LCL and going away from the mean line which signify increasing trend. Identifying only one pattern may mislead the results. Thus, it is required to identify mixture patterns available in the long term or short term series for decision making process.
Time series can be affected by different factors such as gradual change, abrupt change or any complex form of change. Such factors may affect the statistical aspect of the time series data. For example, if a discharge has occurred and a polluted plume is spreading, this shift appears more likely as an increasing trend. Likewise, after taking corrective actions, this shift appears as decreasing trend. Figure 1.10 shows an example, where an increasing trend is obvious. In reality, this increasing trend is due to contamination not the actual behaviour. The plume regain its original form after corrective actions.

![Figure 1.10: Trend due to polluted plume](image)

If the data in a time series is not stable, relying on a single statistical method might give misleading or false results. This highlights the need of a framework that is capable of extracting all the important characteristics of a time series.
From the literature it is found that Akaike information criteria (AIC) selects a best model based on minimizing the value of the mean square error of prediction/estimation [86]. We also observe that AIC exhibits minimum-maximum property, such that the best selected model has values confined exactly up and below following the minimum maximum property. Similar results have been visualized from the experimental results by drawing the dimensionless control chart.

In Chapter 6, we observe that LO(W)ESS error with AIC satisfies choice of smoothing parameter as minimum maximum error property, unless special event such as change-point (either real or inferred from signal behaviour), intervene to change the underlying signal. We analyze LO(W)ESS error by transforming the curve into a dimensionless form, and analyzing the effects of various phenomenon such as change points and other special cases. The LO(W)ESS error is analyzed using the perfect value of smoothing parameter. It is found that the best value of smoothing parameter conform the minimum maximum error property as shown in Figure 1.11. Based on the experimental dataset, smoothing parameter, $\alpha = 0.25$, is the best value.
1.2 Outline

We investigate three different methodologies used to access time series databases:

1. LO(W)ESS,

2. Control Chart Patterns, and

3. Change point detection and Trend analysis

Firstly, a key aspect of data smoothing is to “filter out” noise (undesired data points). LOWESS is a powerful non parametric technique for fitting a smoothed line for a given dataset either through univariate or multivariate smoothing [19]. It implements regression on a collection of points in a moving range, weighted according to distance, around abscissa values in order to calculate smoothed ordinal values. LOWESS and LOESS are an acronym for Locally Weighted Scatterplot Smoothing because, for data smoothing, it is locally weighted regression that is used. Further-
more, a robust weight function can be used to compensate for undue influence of extreme points. The regression model is used based on: a linear polynomial for LOWESS and quadratic polynomial for LOESS [83]. From the literature review, most authors consider LOWESS/LOESS to be the same. If outliers are present in the dataset, so called robust LOWESS/ LOESS procedure is used to overcome the problem of distorted values. LO(W)ESS is widely used in different application areas such as normalization and accessing non-linear relationships between variable. It is unfortunate that despite of its wide application area, the important parameters are selected on trial-and-error basis. This raises the need for selecting smoothing parameter and degree of polynomial automatically based on the dataset.

Secondly, control charts have been used as a tool to detect unnatural variations in a time series. There are eight basic control chart patterns (CCPs), as shown in the Figure 1.12, such as natural, stratification, systematic, cyclic, trend and shift can be detected from a time series.

Traditionally, Shewart’s rules, Nelson rules, AIAG rules, Boeing AQS rules, and Trietsch rules have been used to interpret control chart patterns manually [65, 67]. However, analysing control chart patterns manually requires expert knowledge and experience. Otherwise; it can lead to false or inaccurate analysis. For the last two decades many different approaches (artificial intelligence, artificial neural networks (ANN), expert systems, support vector machines, decision trees (DT), hybrid techniques and Bayesian networks (BN)) have been used for automatic detection of control chart patterns [6, 29, 30, 31, 32, 33, 38, 54, 78]. Expert systems (ES) were created to overcome the problem of manual techniques, but the method requires explicit rules
for pattern recognition. Moreover, ES has problem of false recognition of the shift and trend [50]. ANN is considered as the solution to ES problems and the majority of work has been done using ANN but it still has drawbacks. It has a complex network topology and training process, which is very time-consuming [90]. Although BN is superior to ANN in the way that it uses prior knowledge for analysis and being considered as decision theoretic and does not require a training phase, nevertheless it requires to update belief after every step which increase the complexity of the software system. Most of the existing CCPs methods focus on the recognition of one abnormal pattern.

![Basic Control Chart Patterns](image)

**Figure 1.12: Basic Control Chart Patterns**

From the literature, we find that auto-tuning parameter for LO(W)ESS, concurrent detection of control chart patterns and the relationship between change point detection and trend analysis are overlooked. This is something that has not been taken into consideration for important methods of time series analysis in the research, so far. We, therefore, endeavour to fill this gap by proposing a method that
automatically selects the smoothing parameter and the degree of polynomial for the
calculation of LOWESS fit based on the dataset. We also propose a method for the
concurrent detection of basic control chart patterns. Moreover, we proposed a method
for the interaction between the change point detection and trend analysis. The Thesis
then follows a statistical approach to conduct experiments that prove the merits of
proposed methods.

1.3 Problem Statement

Time series analysis comprises different techniques to differentiate significant
statistics and other characteristics of the data. The extracted meaningful data help
predicting future data based on the observed data. Time series analysis has a wide
range of application area such as forecasting, projections, analysis and much more.
The time series analysis has two important goals:

1. Identification of patterns represented by the sequence of observations and

2. Predicting future events based on observed data.

The modern analytical components for trend analysis and forecasting in the time
series data are

1. Data smoothing techniques using LO(W)ESS

2. Change point detection and Trend analysis

3. Control chart pattern detection
(a) **LO(W)ESS**: There is no perfect automatic method used for identification of trend in the time series. Despite the numerous advantages of LO(W)ESS, the tuning parameters are selected on trial-and-error basis. The selection of the appropriate value of the smoothing parameter and degree of polynomial is extremely important to find the perfect fit for LOWESS. Improper selection of the smoothing parameter may lead to erroneous interpretation as shown in figure. The smoothing parameter decides the number of variables to participate in a particular span, whereas the degree of polynomial shows whether linear or quadratic fits the data. Random selection of smoothing parameter may cause a problem of over smoothing or under-smoothing.

(b) **Control Chart Patterns**: Most of the existing CCPs methods focus on the recognition of one abnormal pattern. There are very few papers in the literature which identify concurrent patterns [59]. However, a time series may have mixture where more than one pattern may exist. Detection of concurrent patterns from time series is extremely important to predict the future events.

(c) **Change point detection and Trend analysis**: Change point measures the abrupt changes, whereas trend measures the gradual changes in the data. Analyzing only change point and overlooking trend or vice-versa may mislead the calculation. Therefore, attempt to detect change point and calculate trend at the same time are equally important for analysis. It could be possible that the trend masks abrupt changes. For example, estimating trend could imply that the trend is either increasing or decreasing, but not why and when it is increas-
ing or decreasing. This concern is answered when both change point detection and trends are analyzed in parallel in the time series. Analyzing the data this way can make improved prediction for future events.

1.4 Aims and Objectives

The principal objective of this Thesis is to identify overlapping features in a time series. A time series can have a combination of different features which cannot be detected by one method. Exploring time series using a combination of different techniques help in gaining deep understanding of data and therefore facilitate informed decision making. The main objective of this Thesis is to answer the following question.

**Question:** What is the best way to use combination of different modelling techniques for identifying overlapping features in a time series?

The main objective of this Thesis is achieved by answering the following specific research questions. Each question has its own importance and it is elaborated in the organization of this Thesis.

1. Which method is both computationally efficient and accurate in identifying correct numbers and locations of change points?

2. Is there really a change point or a trend?

3. Has trend masked the change point or has change point masked the trend?

4. How does one identify mixture patterns based on the statistical characteristics?
5. How can the smoothing parameters be selected, based on the statistical properties of a dataset?

6. What is the best order to use analytical techniques to identify overlapping features?

Question 1-3 are answered in Chapter 4. Chapter 5 answers Question 4. Question 5 is answered in Chapter 3 and Question 6 is answered in Chapter 6.

1.5 Data for Methodology

GEMS has four million records for over 100 parameters from 3,000 stations (River, Lake, Groundwater and Wetlands). Parameters include hydrologic and sampling variables, major ions, metals, microbiology, nutrients, organic contaminants and organic matter. The dataset is already cleaned for analysis [84].

The principle dataset used for our research analysis is from Heidelberg University [7]. It is collected from 16 stations: Blanchard, Chickasaw, Cuyahoga, Grand, Great Miami, Honey Creek, Lost Creek, Maumee, Muskingum, Portage, Raisin, Rock Creek, Sandusky, Scioto, Tiffin, and Vermilion. The data have been collected monthly but randomly from 1974 - 2013 for 10 parameters as dependent variables: Flow, SS, TP, SRP, NO\textsubscript{23}, TKN, Chloride, Sulfate, Silica, and Conductivity. In the raw data, information on some parameters is missing or incomplete measured. Details of data including parameters is given in Appendix, Table A.1.

The efficacy of the proposed system is tested using the synthetic data. For our analysis, we have generated artificial change points in the data set to test if the existing
systems are capable of detecting the exact numbers and locations of change points. Moreover, to differentiate trend and change points from each other we intentionally simulated data with change point and trend together.

Sample patterns are needed to differentiate different patterns and validating control chart pattern recognizer. For our analysis, we have used synthetic data generated by Monte Carlo simulation. The equations used for generating Control chart patterns and Change point are given in Appendix, Table B.1 and Table B.2 respectively.

1.6 Contributions

The literature survey shows that only one analytical technique at a time is used for identifying trend or patterns in a time series data. A time series can be affected by different factors such as gradual change, abrupt change or any complex form of change. Such factors may affect the statistical aspects of the time series data. In such situations, believing that there is only one type of event for which statistical test is performed may lead to misleading or inaccurate decisions. Therefore, it is necessary to have a system that detects different patterns which possibly interact or mask each other, based on the statistical characteristics of the data set. We have proposed a framework in which three different statistical components can be used simultaneously to analyze overlapping patterns. The order of application of those components is important for analysis. Initially, change point detection is used to analyse any abrupt change in the behaviour of a process and then data is segregated based on change point locations. Data is smoothed using LO(W)ESS taking into account the change
point locations and finally, patterns are analysed for the particular dataset. This way the problem of masking effects can be minimized.

### 1.7 Thesis Structure

Figure 1.13 shows the structure of the Thesis.

![Diagram showing the structure of the Thesis](image)

- Chapter 1: Introduction
- Chapter 2: Literature Review
- Chapter 3: LO(W)ESS
  - Proposed methodology and experimental results
- Chapter 4: Changepoint and Trend
  - Proposed methodology and experimental results
- Chapter 5: Control Chart Patterns
  - Proposed methodology and experimental results
- Chapter 6: Minimum-Maximum Property of LO(W)ESS fit for analysis
  - Proposed methodology and experimental results
- Chapter 7: Conclusions and Future Work

Figure 1.13: Structure of the Thesis

This Thesis consists of seven chapters. The brief description for each chapter is given below:

- Chapter two represents literature review on modern techniques such as data
smoothing, detection of abrupt and gradual changes, and control chart patterns being used for time series analysis. This chapter also demonstrates the gaps within the existing research.

Chapter three explores the methodology for auto-tuning LO(W)ESS smoothing parameters. The focus of this chapter is to explore experimentally the selection of smoothing parameter and degree of polynomial on synthetic as well as on real data. It also focuses on combination of techniques to analyze a time series.

Chapter four assesses change-point and trend techniques. This chapter analyzes the current and the best approaches of change point detection and trend. Also, it focuses on interaction between change point and trend.

Chapter five focuses on identification of concurrent patterns based on their statistical properties. The aim of this chapter is to show experimentally the concurrent pattern detection using combination of RobustICA and decision trees.

Chapter six explores the minimum maximum properties of LO(W)ESS fit. The aim of this chapter is to experimentally show the use of combination of techniques to identify statistical features of the time series.

Chapter seven summarizes the conclusions of Thesis and gives recommendations. Future work based on the outcomes of this research is also mentioned.
Chapter 2

Related Work

Time series analysis has always been a hot topic. It has wide application, such as economic forecasting, sales forecasting, stock market analysis, budgetary analysis, and census analysis. In this chapter, some modern methodologies are discussed with the purpose of performing a gap analysis and delimiting the work of this Thesis.

The goal is to give an overview of related work that has been done for LOWESS, control chart patterns and change point analysis. Firstly, the data smoothing method, LO(W)ESS (Section 2.1) is discussed. The focus here is on different smoothing parameters of LOWESS along with the computational details, parameter selection, and limitations of current systems.

In Section 2.2, the control charts and different methods for the identification of control charts patterns, in the dataset, as well as limitations of the existing system are discussed.

In Section 2.3, the focus is on current methodologies for change point analysis and trend. The techniques and methodologies used for trend and change point detection are discussed.

Finally, the chapter summarizes on what still needs to be done with respect to LOWESS, control chart patterns and change point analysis.
2.1 LOWESS

LOWESS, an algorithm proposed by Cleveland [19], is an outlier resistant smoothing method based on local polynomial fit [18]. The improved version was published by Cleveland and Susan J. Devlin in 1988. LOWESS implements a regression on points in a moving range, weighted according to distance around \( X \) values in order to calculate \( Y \) values. “LOWESS” and “LOESS” come from “Locally Weighted Scatterplot Smoothing” because locally weighted linear regression is used for data smoothing. As the smoothed values are calculated using nearest data points, and the regression weight function is biased towards these data points, the process is called local and weighted respectively. This arises from assumption that the process is unaffected by extreme points. Differentiation of a regression model depends on the way it is used: a linear polynomial is used for LOWESS whereas a quadratic polynomial is used for LOESS. From the literature review, most authors consider LOWESS/LOESS same, but they are different. LOWESS is derived from term “Locally Weighted Scatterplot Smoothing” whereas in [81] LOESS stands for “Locally Estimated Scatterplot Smoothing”.

LO(W)ESS regression follows the smoothing procedure in which focus is on the local points (i.e. the smoothed values of \( y \) are obtained based on the data points within the certain range, based on the smoothing parameter). Once the regression function values are calculated with flexible weights and polynomial degree, LOESS fit is complete [48].
2.1.1 Selection of $\lambda$, $W$, $t$ and $\alpha$

LOWESS calculations require the following four items:

* $\lambda$: the polynomial order, where $\lambda \in \{0, 1, 2\}$
* $W$: the weight function
* $t$: count of iterations
* $\alpha$: smoothing parameter

The weight function and iteration can be selected regardless of the properties of the data but $\alpha$ and $\lambda$ should be chosen carefully based on the properties of the data on the scatterplot [48].

Selecting $\lambda$

Linear least squares regression is used to fit a local polynomial for every point in a local region [18]. There are 3 cases for $\lambda$:

$$\lambda = \begin{cases} 
0 & \text{if the target to be plotted is co-related variable} \\
1 & \text{suitable smoothed points are provided with simplicity of computation} \\
2 & \text{computational considerations are on less priority to achieve flexibility}
\end{cases}$$

Selecting $W$

The purpose of the weight function is to create LOWESS a local polynomial fit, considering the nearest points of the point. $W$ is inversely proportional to $x$ to produce the smoothed appearance for points. To enhance a chi-squared distribution of an estimate of the error variance, tricube is calculated to achieve an adequate
smooth for \( W \) that decreases to 0 in all cases\[18\]. Equation 2.1 shows the formula for Tricube function.

\[
w(x) = \begin{cases} 
(1 - |x|^3)^3, & \text{if } |x| < 1; \\
0, & \text{otherwise}
\end{cases}
\] (2.1)

Selecting \( t \)

One can either set a convergence criterion for iteration or limiting \( t = 2 \) for a large number of datasets \[18\].

Choosing alpha (\( \alpha \))

The smoothness of LO(W)ESS curve is directly proportional to \( \alpha \) whereas the variability is inversely proportional (i.e. as the value of \( \alpha \) increases the smoothness of the fit increases and the variability decreases). \( \alpha = 0.5 \) is considered as a good starting point, when the situation is unclear, otherwise, the value of \( \alpha \) can vary anywhere between 0.2 to 0.8 \[19\]. Additionally, in practice, one needs to often try a few choices of \( \alpha \), say between 0.4 to 0.8. Smoothing effects increase with the increased value of \( \alpha \), but it may not be able to give detailed information about the curve \[18\]. This is the major drawback of selecting high value of \( \alpha \).

2.1.2 Computational Details

In the literature, the selection of smoothing parameter, \( \alpha \), is often entirely based on trial and error. Some researchers argue that it should be between 0.2 and 0.8 while other considers 0.5 as an ideal starting point \[18\]. There is no specific technique for
selection of the exact value of $\alpha$. Random selection of $\alpha$ may lead to over-smoothing or under-smoothing of data. The LOWESS/LOESS fit which follows almost all of the data-points is called “under-smoothing” or “over-fitting” whereas if does not follow the data and produces a smooth line is called “lack of fit” or “under-smoothing”.

The step by step calculation of LOWESS/LOESS and rLOWESS/rLOESS are as follows [18, 19, 53].

1. Compute tricube weights, Equation 2.1, using scaled distance. These weights are calculated for a set of numbers in local neighbourhood

2. Run weighted least squares regression for those set of numbers

3. If outliers are present in data; calculate residuals, median of residuals and robust weight, Equation 2.2, using robust weight function

$$
\delta_i = \begin{cases} 
[1 - \left(\frac{e_i}{6s}\right)^2]^2, & \text{if } |e_i| < 6s; \\
0, & \text{otherwise}
\end{cases}
$$

(2.2)

4. Run weighted least regression using robust weights

5. Repeat steps 3 and 4 until convergence criteria is met

### 2.1.3 Automatic Selection of $\alpha$ and $\lambda$

The selection of appropriate values for $\alpha$ and $\lambda$ is extremely important to find the perfect fit for LOWESS. Improper selection of smoothing parameter may lead to erroneous interpretation. The literature review showed that there were 9 different
methods used for selection of smoothing parameter in non-parametric regression [4, 25, 28].

1. Akaike Information Criteria (AIC),

2. an improved version of $AIC_c$,

3. generalized cross validation (GCV),

4. robustified cross validation,

5. average predictive square error,

6. parallel AIC,

7. cross validation,

8. Mallows Cp criteria,

9. risk estimation using classical pilots, and

10. local risk estimation

According to Aydin [4], improved AIC and GCV are the best methods for the selection of smoothing parameter based on small and large datasets. The purpose is to choose smoothing parameter with the smallest AIC score. According to a survey of the literature, $AIC_{c1}$ is an unbiased estimator whereas $AIC_c$ is bias corrected. According to Hurvich, $AIC_c$ is less vulnerable to under-smoothing and over-smoothing compared to AIC and GCV [46]. Moreover, GCV does not work well for large values of smoothing parameter. The Equations 6.1, 2.4, and 2.5 shows formulas for $AIC_{c1}$,
$AIC_c$ and GCV respectively.

\[
AIC_{c1} = \left\{ n \log(\hat{\sigma}^2) + n \frac{\delta_1(n+v_1)}{\delta_2^2} \right\} \quad (2.3)
\]

\[
AIC_c = \left\{ \log(\hat{\sigma}^2) + 1 + \frac{2(\text{Trace}(L)+1)}{n-\text{Trace}(L)-2} \right\} \quad (2.4)
\]

\[
GCV = \left\{ \frac{n\hat{\sigma}^2}{(n-\text{Trace}(L))^2} \right\} \quad (2.5)
\]

\(n=\) number of observations, \(\delta_1=\text{Trace}(I-L)'(I-L)\), \(\delta_2=\text{Trace}((I-L)'(I-L))^2\), \(v_1=\) equivalent number of parameters = Trace\((L'L)\).

Here, \(L\) is hat matrix or smoothing matrix and \(I\) is Identity matrix. Another important parameter is selection of degree of polynomial. If \(X\) and \(Y\) show a monotonic relationship, then \(\lambda = 1\) otherwise \(\lambda = 2\) [48].

### 2.1.4 Limitations in Existing Systems

The problem with current methods being used for model selection is that they do not work for the data with outliers. According to Tharmaratnam [82], non-robust AIC has a tendency to under-fit or wrongly fit in the presence of outliers, so, AIC with \(m\), \(s\) and \(mm\)-estimator provide the accurate fit. Tharmaratnam’s experimental results show that AIC with \(s\) and \(mm\)-estimator is the best method for model selection with or without outliers.

The literature review survey also shows that there is disagreement on distinguishing the term LOWESS and LOESS. Most of the authors consider LOESS as the updated version of LOWESS. Moreover, none of the literature discusses the selection of degree of polynomial for the LOWESS fit. The selection is done arbitrarily, based
on user’s experience for the selection of \( \lambda \). Since data may have presence of outliers, fitting the LO(W)ESS in presence of outliers requires a method for detection of these outliers.

### 2.2 Control Charts

Control charts are statistical tools to find out variations occurring from common causes and special causes in the process. It needs an upper, a lower control limit and a centerline based on the data being plotted. Control charts signify “in-control” (common cause) or “out-of-control” (special cause) process. Control charts have their origins in modeling industrial processes for consistency. The term “in control” is equivalent to saying that the behavior is within the expectation for natural variation, and “out of control” means otherwise. Similarly, this does not mean that when something is “in control” or “out of control” that there is something that is controllable. Instead, these terms are used to explain variation that would expect (in control), as opposed to variation that is deemed excessive (out of control).

A control chart is a run chart that includes statistical limits [44, 74]. According to Shewart, a process is “in-control” if data points lie between \( \pm 3\sigma \). Otherwise, if points lie outside \( 3\sigma \) the process is said to be “out-of-control” [65].

If the data points lie between the \( \pm 2\sigma \), it signifies the normal behavior of the process. Otherwise, if the data points goes beyond \( \pm 2\sigma \), then the process may have some random error or variations, but if the data point(s) goes beyond the \( \pm 3\sigma \) then it indicate that the process is behaving abnormally and require remedial actions. Figure
2.1 shows the basic concept of control chart with control limits.

With a normal data dispersion data is expected to be included within $\sigma$ [40]:

1. $1\sigma$ - 68.3%
2. $2\sigma$ - 95.5%
3. $3\sigma$ - 99.7%

It is expected that 99.7% of process outcome is within the $3\sigma$ limit. If it goes beyond this limit a so-called, *special cause* of variation is considered [40]. Control charts monitor variations over time [12]. The behavior of the process can be monitored for a specific period. In order to plot control charts, it is usual to check for normality, stability, amount of data and correlated data. If data do not conform to
these characteristics then control charts give wrong signals [2]. It is better to use values from the t-table of 0.05/0.003 probability level to set control limits rather than $\pm 2\sigma$, $3\sigma$ as it does not exactly include 95% and 99.7% of data [36]. Action limits from the average, probability error is 0.26% (100% - 99.7%), whereas for warning limits, it is 4.56% (100% - 95.44%) [65].

Control charts have been used as a tool to detect unnatural variations in a time series. Recognition of unnatural variations using control chart patterns (CCPs) is important in water quality to find out the unnatural variations in water quality parameters. There are eight basic CCPs: natural, stratification, systematic, cyclic, trend, and shift. Control chart patterns can be generated using the equation given in Table B.1.

Very few authors have addressed the issue of feature extraction from concurrent patterns [59]. The problem of identifying mixture patterns is a non-trivial one. Hachicha proposed a classification scheme for differentiating classification techniques [38]. In order to classify different methods for CCPs one should answer model of control chart pattern, type of patterns, identification of number of patterns, type of pattern recognition i.e. raw data-based or feature based, type of approach being used, validation of approach and performance criteria for methods. Hachicha developed a basic questionnaire required for detection of control chart patterns, shown in Table 2.2 [38].

Different authors followed varying approaches based on the model and type of patterns to be discovered. Some authors emphasize raw-based techniques, whereas others focus on feature-based techniques. Feature-based techniques are more efficient
Table 2.1: Parameters for Control Chart Patterns

<table>
<thead>
<tr>
<th>Pattern type</th>
<th>Parameters /Values</th>
<th>Pattern Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal (NOR)</td>
<td>Mean ($\mu$) = 80, Standard deviation ($\sigma$) = 5</td>
<td>$y_i = \mu + r_i \sigma$</td>
</tr>
<tr>
<td>Stratification (STA)</td>
<td>Random noise($\sigma'$) = 0.2$\sigma$ to 0.4$\sigma$</td>
<td>$y_i = \mu + r_i \sigma'$</td>
</tr>
<tr>
<td>Systematic(SYS)</td>
<td>Systematic Departure(d) = 1$\sigma$ to 3$\sigma$</td>
<td>$y_i = \mu + r_i \sigma + d \times (-1)^i$</td>
</tr>
<tr>
<td>Cyclic (CYC)</td>
<td>Amplitude (a)= 1.5$\sigma$ to 2.5$\sigma$, Period (T) = 8 and 16</td>
<td>$y_i = \mu + r_i \sigma + a \sin\left(\frac{2\pi i}{T}\right)$</td>
</tr>
<tr>
<td>Increasing Trend(UT)</td>
<td>Gradient (g)=0.05($\sigma$) to 0.1($\sigma$)</td>
<td>$y_i = \mu + r_i \sigma + ig$</td>
</tr>
<tr>
<td>Decreasing Trend(DT)</td>
<td>Gradient (g) = 0.05($\sigma$) to 0.1($\sigma$)</td>
<td>$y_i = \mu + r_i \sigma - ig$</td>
</tr>
<tr>
<td>Upward Shift (US)</td>
<td>Shift magnitude(s)=1.5($\sigma$) to 2.5($\sigma$), Shift position (P) = 9, 17,25</td>
<td>$y_i = \mu + r_i \sigma + ks$</td>
</tr>
<tr>
<td>Downward Shift (DS)</td>
<td>Shift magnitude(s)=1.5($\sigma$) to 2.5($\sigma$), Shift position (P) = 9, 17,25</td>
<td>$y_i = \mu + r_i \sigma - ks$</td>
</tr>
</tbody>
</table>

compared to raw data-based and most of the work on feature based techniques is
done by Gauri [6].

In 1920s Shewart rules came into existence for pattern recognition. With tech-
nological advancement, machine learning techniques have been developed to over-
come the problems of manual techniques. Traditionally, Shewart’s rules, Nelson rules,
AIAG rules, Boeing AQS rules, and Trietsch rules have been used to interpret con-
trol chart patterns manually [65, 67]. However, analyzing control chart patterns

35
Table 2.2: Basic Questionnaire for CCPs [38]

<table>
<thead>
<tr>
<th>Question</th>
<th>Possible Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of data model</td>
<td>(a) i.i.d process</td>
</tr>
<tr>
<td></td>
<td>(b) Non-normality distribution</td>
</tr>
<tr>
<td></td>
<td>(c) Auto-correlated process</td>
</tr>
<tr>
<td></td>
<td>(d) Multivariate</td>
</tr>
<tr>
<td></td>
<td>(e) Multivariate and auto-correlated</td>
</tr>
<tr>
<td>Types of CCPs</td>
<td>(a) Basic patterns</td>
</tr>
<tr>
<td></td>
<td>(b) Special patterns</td>
</tr>
<tr>
<td>Number of patterns detected</td>
<td>(a) Single</td>
</tr>
<tr>
<td></td>
<td>(b) Concurrent</td>
</tr>
<tr>
<td>Data representation technique</td>
<td>(a) Raw-based</td>
</tr>
<tr>
<td></td>
<td>(b) Feature-based</td>
</tr>
<tr>
<td>Type of approach for detecting CCPs</td>
<td>(a) Manual</td>
</tr>
<tr>
<td></td>
<td>(b) Automatic</td>
</tr>
<tr>
<td>Techniques for validation of the approach</td>
<td>(a) Monte-Carlo</td>
</tr>
<tr>
<td></td>
<td>(b) Real-process data</td>
</tr>
<tr>
<td>Performance criteria</td>
<td>(a) No performance measures</td>
</tr>
<tr>
<td></td>
<td>(b) Non-ARLs (Average Run Length)</td>
</tr>
<tr>
<td></td>
<td>(c) Traditional ARLs (ARL0 and ARL1)</td>
</tr>
<tr>
<td></td>
<td>(d) Advanced ARLs (pattern distinction capability)</td>
</tr>
</tbody>
</table>
manually requires expert knowledge and experience. Otherwise, it can lead to false or inaccurate analysis. From last two decades different approaches (artificial intelligence, artificial neural network, expert systems, support vector machines, hybrid techniques and Bayesian networks) have been used for automatic detection of CCPs [6, 29, 30, 31, 32, 33, 38, 54, 78].

Expert systems (ES) were created to overcome the problem of manual techniques but it requires explicit rules for pattern recognition. Moreover, ES has a problem for false recognition of shift and trend. Artificial Neural Networks (ANN) are considered as a solution to ES problem and the majority of work has been done using ANN. The BN approach still has drawbacks. It involves complex network topology and a lengthy training process, which is very time-consuming [90]. ANN requires large amount of training data, getting into local extremes easily, difficulty in obtaining a stable solution, over-fitting of model, and bad generalization ability. These weaknesses restrict the applications of ANN. Although Bayesian Networks are superior to ANN in the way that they use prior knowledge for analysis and are considered as decision theoretic, not requiring training phase, BN require belief updating after every step, increasing the complexity of a system.

Most of the existing CCP methods focus on the recognition of one abnormal pattern. There are very few papers in the literature which identify concurrent patterns [59]. However, a water quality time series may have a mixture where more than one pattern exists as shown in Figure 2.2. The feature based approach first extract features and then use those features for control chart pattern detection [90].

Guh and Tannock used a back-propagation neural network whereas Chen et al.
used integrated wavelet method and back propagation neural network for online recognition of concurrent patterns [17, 37]. Wang et al. proposed an integrated approach by combining wavelet analysis and a neural network for the recognition of mixture patterns for detection of concurrent patterns. Masood and Hassan [63] discussed the issues of using ANN for detection of control chart patterns. Lu et al. [59] proposed a method ICA-SVM (Support Vector Machine) to differentiate mixture patterns. The major drawbacks of using SVM are the memory and the complex algorithm structure.

2.2.1 Feature Extraction Techniques

Independent Component Analysis

Independent Component Analysis (ICA) is a powerful feature extraction technique. The purpose is to identify independent sources from their mixture, without knowing the process of mixing or any specific information of the sources. In other
words, linearly mixed sources. The independent components generated from mixture patterns are used as the independent sources of the mixture patterns. The hidden basic patterns of the mixture patterns could be discovered in these independent components (ICs). ICA has been used in several fields of multivariate data processing, face recognition, image processing and time series analysis. Non-Gaussian properties such as kurtosis and negentropy is used to measure the statistical independence of independent components [60, 61].

Kurtosis is important factor in ICA theory because of its algebraic and statistical properties. It is generalization of variance with high degree of polynomial [47]. Negentropy is another important non-Gaussian measure in ICA theory. It is based on the methodology of information theory, differential of entropy [47]. The advantage of using negentropy is its statistical property; it is acceptable by statistical theory.

ICA has applications in brain imaging (basic blind source separation model), econometric models (decomposition of parallel time series) and feature extraction (extraction of independent features). It uses the following three step of processing the data [47]:

1. **Centering**: the foremost step is to center the vector by subtracting its mean. This step helps to simplify the algorithm.

2. **Whitening**: another useful step is to whiten the data. In this step, the observed vector is transformed to new vector which is uncorrelated and variance in unity.

3. **Mixing matrix**: in this step, filtering is done by multiplying another matrix from the right.
FASTICA

FastICA is an improved version of ICA [60, 61, 62]. Compared to ICA, FastICA has significantly improved properties: firstly, the convergence criterion for FastICA is cubic whereas ICA is linear. Secondly, it is easy to use since there is no parameter selection. Thirdly, it is memory and computationally efficient, distributed and easy to use. Based on these properties of FastICA it has been used for control chart pattern feature extraction. The basic form of FastICA algorithm is as follows:

Algorithm 1 FastICA Algorithm [60, 61, 62]

1: procedure FASTICA
2: Choose an initial (e.g. random) weight vector \( w \)
3: Iterate: Let \( w^+ = E \{xg(w^T x)\} - E \{g'(w^T x)\} w \)
4: Normalize: Let \( w = w^+ / ||w^+|| \)
5: If not converged, go back to 3

2.2.2 Limitation in Existing Systems

The literature survey shows that very little work has been done to identify mixture patterns. So far, the techniques such as SVM, DT, hybrid scheme, engineering process control, and ANN use ICA for the identification of basic patterns [16, 17, 60, 61, 62, 89]. According to Baloch [9], Kurtosis and negentropy based methods suffer in the presence of outliers. Baloch’s experimental results show that RobustICA is efficient compared to Kurtosis and negentropy based ICA. RobustICA
is discussed in detail in Chapter 5, Section 5.3.1. Moreover, different techniques for feature extractions have drawbacks in their system. The results generated from ANN are difficult for naive user. Also, BN requires the updating of belief at every step. We would prefer a system that is easy to understand and do not require more user input to proceed further.

2.3 Change point detection and Trend analysis

2.3.1 Change point

Change point analysis is another tool to detect change(s) in a process over a period of time. It is simple and flexible to use and characterizes the changes by providing detailed information, confidence interval and confidence level. Change point analysis can be applied to large datasets with variable, attribute, non-normal, normal and data with outliers [80]. Change point analysis is efficient enough to detect changes overlooked by control charts. There are different methods to detect change-points [10, 24, 26, 34, 49, 56, 64, 73]:

1. Standard normal homogeneity test

2. non-parametric SNH test

3. two-phase regression of Wang

4. TPR of Lund and Reeves

5. method of Vincent
6. Akaike’s information criteria and Sawas Bayes criteria

7. CUSUM (Cumulative Sum)

8. wild binary segmentation

9. agglomerative algorithm for change point

10. iterative robust detection method and

11. Bayesian analysis.

Table 2.3: Parameters for Change point

<table>
<thead>
<tr>
<th>Pattern type</th>
<th>Pattern Equation</th>
<th>Parameters Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Step</td>
<td>$\mu_1 = \mu_0 + \delta \frac{\sigma_0}{\sqrt{n}}$</td>
<td>$\delta = 0.5, 1.1.5, 2, 2.5, 3$</td>
</tr>
<tr>
<td>Linear Trend</td>
<td>$\mu_1 = \mu_0 + \beta (i - \tau) \frac{\sigma_0}{\sqrt{n}}$</td>
<td>$\beta = 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$</td>
</tr>
<tr>
<td>Systematic</td>
<td>$\mu_1 = \mu_0 + l \cdot (-1)^i \cdot \sigma_0$</td>
<td>$l = 0.5, 1, 1.5, 2, 2.5, 3$</td>
</tr>
<tr>
<td>Cyclic</td>
<td>$\mu_1 = \mu_0 + l \cdot \sin(2\pi \frac{i}{5}) \cdot \sigma_0$</td>
<td>$l = 0.5, 1, 1.5, 2, 2.5, 3$</td>
</tr>
<tr>
<td>Mixture</td>
<td>$\mu_1 = \mu_0 + l \cdot (-1)^h \cdot \sigma_0$</td>
<td>$l = 0.5, 1, 1.5, 2, 2.5, 3$</td>
</tr>
</tbody>
</table>

where $\mu_1 = \text{new mean}$, $\mu_0 = \text{old mean}$, $\sigma_0= \text{old standard deviation}$, $n= \text{number of observations}$, $l= \text{cyclic amplitude}$, $\beta=\text{slope}$, $t=\text{sampling time}$

According to Reeves [73], the Sawas Bayes criteria also known as Bayesian Information Criteria, performs well compared to other methods. In order to detect change-point by shift in mean using synthetic data, consider using the equation given in Table B.2 [35].

42
Here is a glance of modern existing algorithm used to detect change points in a dataset:

**Review of existing methods**

◊ **E-Agglomerative**: This algorithm is used to detect multiple change points in the dataset. It is able to detect number and location of change points present in the dataset and it is considered computationally efficient [64]. The number of change points is estimated based on “goodness of fit” of clusters in the dataset. An agglomerative algorithm which is hierarchal in nature uses a bottom-up approach for change point detection. It preserves the time ordering of the observations. Although both divisive and agglomerative algorithms are hierarchical algorithm, they differ in approach. The divisive algorithm uses a top-down approach, division of data, whereas agglomerative algorithm uses a bottom-up approach, merging of data. The “goodness-of-fit” is used to merge the contiguous pair of clusters. The goodness-of-fit statistic is computed using “linkage-criterion” distance among two clusters [64]. The algorithm alters an object-to-variable table into an object-to-object distance matrix as an initial step.

Initially, each object is considered as a single element and is merged with the closest element, in every iteration. The process is repeated until a single cluster is obtained and results are organized as a tree like structure.

The R package, “ecp”, is available to calculate number of change points [49]. In
this package, two different methods, E-Divisive and E-Agglomerative algorithms are provided for change point detection in univariate as well as multivariate data. The divisive methods estimate change points using a bisection algorithm whereas the agglomerative method detects abrupt change using agglomeration [49]. According to Matteson, the agglomerative approach is preferred over the divisive approach [64].

♦ **Wild Binary Segmentation (WBS):** This change point detection method claims to detect the exact number and potential locations of change points without any prior assumption. It uses the idea of computing Cumulative Sum (CUSUM) from the randomly drawn intervals and considers the largest CUSUM as the first change point candidate to test against stopping criteria [26]. Then the data is divided into segments based on the first location of the detected change point. The process is repeated on the sub-segments.

The R package, WBS, is available with two methods, Standard Binary Segmentation (sbs) and Wild Binary Segmentation (wbs). According to Fryzlewicz [11], WBS overcomes the problems of the binary segmentation technique and is an improvement in terms of computation. Algorithm 2 shows the WBS algorithm.

♦ **Bayesian Analysis of Change points (BCP):** This algorithm works on a product partition model and compute the change point based on data and current partition. Tuning parameter values are provided for accurate analysis of change points [24]. The Markov chain Monte Carlo (MCMC) implementation of Bayesian analysis can be estimated in $O(n^2)$ time. We used the R package,

1: procedure WildBinSeg\((s, e, \zeta_\tau)\)

2: \hspace{1em} if \(e - s < 1\) then

3: \hspace{2em} STOP

4: \hspace{1em} else

5: \hspace{2em} \(M_{s,e} = \) set of those indices \(m\) for which \([s_m, e_m] \in F^M_T\) is such that \([s_m, e_m] \subseteq [s, e]\)

6: \hspace{2em} (Optional: augment \(M_{s,e} = M_{s,e} \cup 0\), where \([s_0, e_0] = [s, e]\))

7: \hspace{2em} \((m_0, b_0) = \arg\max_{m \in M_{s,e}, b \in [s_m, \ldots, e_{m-1}]} \left| \tilde{X}^b_{s_m, e_m} \right|\)

8: \hspace{2em} if \(\left| \tilde{X}^{b_0}_{s_{m_0}, e_{m_0}} \right| > \zeta_\tau\) then

9: \hspace{3em} add \(b_0\) to the set of estimated change-points

10: \hspace{2em} WildBinSeg \(s, b_0, \zeta_\tau\)

11: \hspace{2em} WildBinSeg \(b_0 + 1, e, \zeta_\tau\)

12: \hspace{2em} else

13: \hspace{3em} STOP

14: \hspace{2em} ENDIF

15: \hspace{1em} ENDIF
“bcp”, for our experiments [24].

Robust detection of Change point: This methods works well when outliers are present but it prompts false alarms since it works on “hierarchical binary splitting” techniques. Least squares regression does not perform well in the presence of outliers, and therefore, robust detection overcomes that problem [34]. Robust detection is modification of two-phase regression model method. The two-phase regression model decomposes the time series into linear segments using ordinary least squares (OLS). Applying OLS to the data means that dataset is not serially correlated and has the same statistical properties. Moreover, experimental results show that two-phase regression is not capable of detecting accurate locations of change points. It detects only one change point at a time. Also, this method does not work well in the presence of outliers. This method cannot be applied to time series because of constant variance error [34]. Robust detection overcomes of all drawbacks of two-phase regression model. There is no R package available; so, we ran iterative robust detection method using Python implementation, provided by Stooksbury [34].

2.3.2 Trend detection

Trend analysis can be defined as a process of estimating gradual change in future events from past data. Different parametric and non-parametric techniques are used to estimate trends. The Mann-Kendall, Spearman rho, seasonal Kendall, the Sen’s t
and Cox Stuart are the most widely used methods for trend analyses [39, 43, 70, 91].

Mann-Kendall

The Mann-Kendall (MK) test is a non-parametric technique to detect trends in a time series. Most of the researchers consider this test as an excellent method to detect trend [15]. This test works well even in the presence of missing values.

$H_0$: No trend exists in the series

$H_A$: Trend exists

This test compares the difference of values rather than data values. Initially, the data values are ranked and then sum of the difference of consecutive values is computed using the MK statistics. The MK test statistics is given by the following Equation 2.6

$$S = \left\{ \sum_{i=1}^{n-1} \sum_{j=i}^{n} Sgn(X_j - X_i) \right\}$$

where $X_i$ and $X_j$ are the sequential data values, $n$ is the length of the dataset and

$$Sgn(\Theta) = \begin{cases} 
+1 & \text{if } \Theta > 0 \\
0 & \text{if } \Theta = 0 \\
-1 & \text{if } \Theta < 0
\end{cases}$$

Cox-Stuart

Cox-Stuart test is robust and less powerful method with power = 0.78 for trend analysis. This test comes under the umbrella of nonparametric test. The hypothesis $H_a$ shows the trend is monotonic: increasing trend or decreasing trend whereas $H_0$ shows no trend. In order to test the trend using Cox-Stuart method, the observations
should be even. If the observations are odd then delete the observation at location \((\frac{N+1}{2})\). Now, subdivide the even dataset into two vectors and compute the difference of vectors. Then count the number of positive and negative signs and select the number with small magnitude. Finally, use that magnitude to calculate the binomial distribution. If the obtained value is less than significance level 0.05 then accept the \(H_a\) for trend. Therefore, it is used in wide application areas to know the development of obtained values [77].

\(H_o\): No trend exists in the series

\(H_A\): Trend exists

\[
T = \left\{ \sum_{i=1}^{m} I(C_{yt} < C_{yi+c}) \right\}
\]  

(2.8)

2.3.3 Limitations in Existing Systems

Interaction between Change point detection and trend analysis is extremely important. Analyzing only change point and overseeing trend or vice-versa may mislead the calculation. There exists different school of thoughts on these perspectives. For example, is it a trend or change point? Are there one or multiple change points? Are the locations and number of change points accurate? These scenarios necessitate the separation of these concerns and advocate exploring the interaction between change point and trend.
2.4 Summary

It is appropriate to summarize this chapter to illustrate the important concepts of LOWESS, control chart patterns and change point and trend. We found that the importance of outliers is ignored in time series analysis using LOWESS and control chart patterns. Moreover, the importance of the relationship between abrupt and gradual changes is ignored when both are present in the time series.

In this chapter we explored the detailed literature of LOWESS, control chart patterns and change point and trend. From literature, the current system to auto-tune smoothing parameters is not capable to obtain accurate results for LOWESS fit. It either produces an over-smoothed or an under-smoothed curve.

A literature review has also been done in this chapter on CCPs and automatic techniques being used for identification of CCPs. Pitfalls of the current system have also been highlighted. It is not capable of generating accurate independent components in the presence of outliers.

Finally, the chapter examines the literature on change point and trend analysis. Our literature survey shows that importance of discovering the relationship between the change point and the trend when both are present in a time series.
Chapter 3

LO(W)ESS

This chapter outlines and explores the proposed method for automatic selection of smoothing parameters for LO(W)ESS. The methodology tests two types of datasets: an artificially generated linear dataset and an actual dataset from Rock Creek River [7]. It is hypothesized that the automatic selection of smoothing parameters for LO(W)ESS will outperform the existing methods.

In Section 3.1, different types of LO(W)ESS models have been discussed. Section 3.2.1 illustrates the proposed methodology and Section 3.2.2 shows the experimental results using synthetically generated data as well as actual dataset. Finally, in Section 3.3, the results are summarized.

3.1 Locally Weighted Scatterplot Smoothing

LOWESS is a heavily-used non-parametric technique for fitting a smoothed line for a given dataset, either through univariate or multivariate smoothing [19]. It implements regression on a collection of points in a moving range and weighted according to distance from a central abscissa in order to calculate an ordinal value. The regression model is used based on a linear polynomial for LOWESS and quadratic polynomial for
LOESS [83]. Many authors consider LOWESS and LOESS to be the same algorithm but this is not the case.

The selection of smoothing parameter, $\alpha$, is often entirely based on a “repeated trial” basis. Some researchers argue that it should lie between 0.2 and 0.8 while others consider 0.5 as an ideal parameter value [18]. There is no specific technique for selection of the exact value of $\alpha$. Selection of $\alpha$ may lead to “over-smoothing” or “under-smoothing” of data, does not necessarily provide good information for LO(W)ESS fit. Figure 3.1 shows a sample LO (W)ESS fit using different smoothing parameters. The LO(W)ESS fit which follows the almost all the data points is called under smoothing or over-fitting whereas if the fit does not follow the data and produce a smooth line is called lack of fit or under-smoothing.

![Figure 3.1: LOWESS with different values of $\alpha$](image)

There are numerous techniques for data smoothing: splines, Bezier curves, kernel and polynomial regression[5, 20]. LO(W)ESS fit is very informative when the dataset is large [20]. Moreover, it is used to solve the problems of precision, noise filtering,
and outliers and is known to adapt well to bias problems, as opposed to the other named methods. Also, LO(W)ESS is computationally efficient [42].

3.2 Research Objective

The objective is to develop a model that auto-tunes the important smoothing parameters. To achieve this objective, RobustAIC is used to select the of best smoothing parameter, $\alpha$, and weighted Pearson correlation for to select the degree of polynomial, $\lambda$.

The specific objectives are the following:

(a) Identify outliers in the dataset

(b) Identify the degree of polynomial

(c) Select the LO(W)ESS method based on the data

(d) Select the value of $\alpha$

(e) Plot the LO(W)ESS fit

3.2.1 Proposed Methodology

The proposed algorithm for selection of smoothing parameters for LO(W)ESS as given in the Algorithm 3. The steps for appropriate selection of smoothing parameters for LO(W)ESS are shown in Figure 3.2.

The experimental steps for selection of $\alpha$ and $\lambda$ are as follows:
Algorithm 3 LO(W)ESS

1: procedure LO(W)ESS($\alpha$, $\lambda$)

2: Identify the presence of outliers using Hample Identifier

3: if outlier(s) detected then

4: rLO(W)ESS

5: else

6: LO(W)ESS

7: Evaluate the degree of polynomial, $\lambda$, using Pearson Correlation

8: if monotonic relationship then

9: LOWESS or rLOWESS

10: else

11: LOESS or rLOESS

12: Identify the value of smoothing parameter, $\alpha$, using RobustAIC

1. Analyze the presence of outliers in the data: In order to detect the presence of outliers a Hample Identifier (HI) is employed [55]. HI is one of the
most robust methods to detect accurate number of outliers [55]. It computes the location scale using median and median absolute deviation (MAD).

2. **Differentiate LO(W)ESS from rLO(W)ESS**: If outliers are present in the dataset, rLO(W)ESS is used; otherwise, LO(W)ESS is used for analysis.

3. **Examine the presence of a monotonic relationship**: This step is extremely important to identify which degree of polynomial is used, linear or quadratic. If X and Y show a monotonic relationship, then $\lambda = 1$; otherwise, $\lambda = 2$ [48]. There are three techniques, parametric and non-parametric, to check the monotonic relationship [41].

   (a) Kendall Tau,
   
   (b) Pearson Correlation, and
   
   (c) Spearman Rho

The strength of correlation can be measured based on values given in Table 3.1. For automatic selection of $\lambda$, weighted Pearson correlation, $r_w$, or Mean Squared Error (MSE) can be used [51, 79]. The Pearson Correlation coefficient is calculated for the express purpose of identifying whether the data is monotonic. The best fit polynomial is considered to be the one for which MSE is less or having high $r_w$ value. The $r_w$ is more suitable for testing correlation compared to Pearson correlation [51]. Equation (3.1) shows the formula to calculate MSE and Equation (3.2) for $r_w$.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$$ (3.1)
Table 3.1: Correlation Coefficient

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Very Strong</th>
<th>Strong</th>
<th>Moderate</th>
<th>Weak</th>
<th>Very Weak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.80-1.0</td>
<td>0.60-0.79</td>
<td>0.40-0.59</td>
<td>0.20-0.39</td>
<td>0.00-0.19</td>
</tr>
</tbody>
</table>

\[
r_w = \frac{\sum_{i=1}^{n} w_i(x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sqrt{\sum_{i=1}^{n} w_i(x_i - \bar{x}_w)^2 \sum_{i=1}^{n} w_i(y_i - \bar{y}_w)^2}}
\] (3.2)

4. **Differentiating LOWESS, LOESS, rLOWESS and rLOESS:** The selection of best LO(W)ESS model can be performed as follows:

(a) Monotonic relationship without outliers: LOWESS

(b) Monotonic relationship with outliers: rLOWESS

(c) No monotonic relationship without outlier: LOESS

(d) No monotonic relationship with outlier: rLOESS

5. **Select the best smoothing parameter using RobustAIC:** Tharmaratnam [82] proposed the robust version of AIC, which produces promising results for model selection in the presence of outliers compared to non-robust AIC. The best selected model can be estimated using Akaike weights [88]. The algorithm for calculating Akaike weights is as follows:

(a) calculate AIC for all the models and identify the best model, \( AIC_{\text{min}} \)

(b) calculate the difference between AIC of every model and \( AIC_{\text{min}} \)

\[
\Delta_i(AIC) = AIC_i - AIC_{\text{min}}
\] (3.3)
(c) compute Akaike weights for each model and normalized relative likelihoods

\[ w_i = \frac{exp[-0.5 \times \Delta_i(AIC)]}{\sum_{i=1}^{n} exp[-0.5 \times \Delta_i(AIC)]} \] (3.4)

The model is considered as the best model for the fit when the value for robustAIC is lowest.

3.2.2 Experimental Design

Experiment 1: LOWESS analysis for noisy synthetic linear data

This experiment explores how LO(W)ESS performs when some dataset with or without outliers are present. In order to perform this experiment, steps from procedure 1 are used. Noisy linear data has been generated using (3.5). It is known in advance that data is linear and outliers are present and the experiment is designed to detect whether the proposed methodology works as per expectations.

\[ Y = X + e \] (3.5)

An artificial dataset with 50 data points is generated for the experiment. Normally distributed noise, \( e \), with zero mean and 0.01 variance has been added to produce outliers in the linear data, whereas \( X \) is generated uniformly. The selection of appropriate \( \alpha \) and \( \lambda \) is extremely important to find the perfect fit for LO(W)ESS as previously noted. Improper selection of the smoothing parameter may lead to erroneous interpretation. The \( \alpha \) decides the number of variables to participate in a particular span, whereas \( \lambda \) shows whether a linear or quadratic fits the data.
**Step 1 Presence of Outliers:** The Hample Identifier (HI) detected two outliers at index 21 and 46.

**Step 2 LO(W)ESS vs. rLO(W)ESS:** Since two outliers are detected in the dataset, rLO(W)ESS is employed. The computational procedure for LO(W)ESS and rLO(W)ESS is given in Chapter 2.

**Step 3 Identify monotonic relationships:** The selection of appropriate $\alpha$ and $\lambda$ is extremely important. Based on experimental results for linear data, the value of $r_w = 0.896$ whereas $r = 0.582$. This indicates that there is strong relationship; therefore, $\lambda=1$ should be used for analysis. Similarly, the calculation value for MSE for $\lambda=1$ is 0.00184 whereas for $\lambda=2$ is 0.00188. Again, MSE confirms that $\lambda=1$ is the best for analysis.

**Step 4 Differentiating LOWESS, LOESS, rLOWESS and rLOESS:** Since a monotonic relationship exists and presence of outliers is detected in Step 1, rLOWESS is chosen for the analysis.

**Step 5 RobustAIC for parameter selection:** RobustAIC selects the best value for $\alpha$ based on the data. Based on Tharmaratnam robust AIC [82], Table 3.2 illustrates that $\alpha = 0.3$ is the smallest AIC score. Table 3.3 shows the best selected model using Akaike weights.

Figure 3.3 shows the rLOWESS fit based on $\alpha=0.3$. In this Figure, original data, rLOWESS fit and data with added noise is showed. In this situation, the original data to which the noise was “added” is not recovered. This is an artificial dataset.

The criteria do not recover the noise-free data, but rather they smooth the dataset according to the noisy data as presented, some of which began as an added noise, and
Table 3.2: Robust AIC score for different values of $\alpha$

<table>
<thead>
<tr>
<th>Smoothing Parameter</th>
<th>Robust AIC Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>-323.061</td>
</tr>
<tr>
<td>0.2</td>
<td>-340.792</td>
</tr>
<tr>
<td>0.3</td>
<td>-347.394</td>
</tr>
<tr>
<td>0.4</td>
<td>-325.645</td>
</tr>
<tr>
<td>0.5</td>
<td>-320.392</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoothing Parameter</th>
<th>Robust AIC Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>-311.830</td>
</tr>
<tr>
<td>0.7</td>
<td>-309.811</td>
</tr>
<tr>
<td>0.8</td>
<td>-317.420</td>
</tr>
<tr>
<td>0.9</td>
<td>-306.575</td>
</tr>
<tr>
<td>1.0</td>
<td>-310.225</td>
</tr>
</tbody>
</table>

Figure 3.3: rLOWESS fit with $\alpha=0.3$

which is retained in the smoothed line. Figure 3.4 shows the rLOWESS fit based on $\alpha=0.3$. 

58
Table 3.3: Best selected model using Akaike Weights

<table>
<thead>
<tr>
<th>Robust AIC</th>
<th>$\delta_{AIC}$</th>
<th>$w_i(AIC)$</th>
<th>Robust AIC</th>
<th>$\delta_{AIC}$</th>
<th>$w_i(AIC)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-323.061</td>
<td>24.333</td>
<td>5.01685E-06</td>
<td>-311.830</td>
<td>35.564</td>
<td>1.82663E-08</td>
</tr>
<tr>
<td>-340.792</td>
<td>6.602</td>
<td>0.035536013</td>
<td>-309.811</td>
<td>37.583</td>
<td>6.65626E-09</td>
</tr>
<tr>
<td>-347.394</td>
<td>0</td>
<td>0.964439054</td>
<td>-317.420</td>
<td>29.974</td>
<td>2.98884E-07</td>
</tr>
<tr>
<td>-325.645</td>
<td>21.749</td>
<td>1.82616E-05</td>
<td>-306.575</td>
<td>40.819</td>
<td>1.3199E-09</td>
</tr>
<tr>
<td>-320.392</td>
<td>27.002</td>
<td>1.32088E-06</td>
<td>-310.225</td>
<td>37.169</td>
<td>8.18708E-09</td>
</tr>
</tbody>
</table>

Figure 3.4: rLOWESS fit with $\alpha=0.3$

Experiment 2: LO(W)ESS analysis for Rock Creek River

In this experiment, the proposed method is tested on naturally occurring data. The dataset used has been obtained from Heidelberg University [7]. It is collected from Rock Creek River for 10 different parameters and we will analyze “Suspended
“Solid” data collected hourly but intermittently between 1982-2013.

Calculating the mean of stratified samples does not necessarily provide an accurate picture of the data. The data readings vary from hourly to daily data and therefore, weights should be considered to calculate average. The number of data points is reduced flow-weighted mean concentration (FWMC) for accuracy and consistency which should give a clear picture of actual loadings.

The flow-weighted mean is considered to be an accurate method for use in calculating average for stratified samples in which the readings have unequal time intervals, varying from hourly to daily readings [8, 22]. Calculating FWMC does not lead to the effect of missing data [22]. Equation (3.6) shows the formula to calculate flow-weighted average is [8, 22].

\[
FWMC = \frac{\sum_{i}^{n}(c_i \cdot t_i \cdot q_i)}{\sum_{i}^{n}(t_i \cdot q_i)}
\]  

(3.6)

where \( q_i \) = flow in the \( i^{th} \) sample, \( c_i \) = concentration of the \( i^{th} \) sample, \( t_i \) = time window for the \( i^{th} \) sample

**Step 1 Presence of Outliers:** Based on dataset, HI detected 28 outliers in the dataset

**Step 2 LO(W)ESS vs. rLO(W)ESS:** Since outliers are detected in the dataset, rLO(W)ESS is chosen.

**Step 3 Identify monotonic relationships:** Based on the analysis, the value of \( r_w = 0.0191 \) shows a very weak relationship; rLOESS is indicated.

**Step 4 Differentiating LOWESS, LOESS, rLOWESS and rLOESS:** Since monotonic relationship exists and presence of outliers is detected in Step 1, rLOESS
is chosen for the analysis.

**Step 5 RobustAIC for parameter selection:** RobustAIC selects the best value for $\alpha$ based on the data. Based on Tharmaratnam robust AIC [82], Table 3.4 shows that $\alpha = 0.8$ is the smallest AIC score. Table 3.5 shows the best selected model using Akaike weights.

**Table 3.4: Robust AIC score for different values of $\alpha$**

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>Robust AIC Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>3190.12</td>
</tr>
<tr>
<td>0.2</td>
<td>2801.927</td>
</tr>
<tr>
<td>0.3</td>
<td>2475.304</td>
</tr>
<tr>
<td>0.4</td>
<td>2496.164</td>
</tr>
<tr>
<td>0.5</td>
<td>2479.504</td>
</tr>
<tr>
<td>0.6</td>
<td>2452.204</td>
</tr>
<tr>
<td>0.7</td>
<td>2193.253</td>
</tr>
<tr>
<td>0.8</td>
<td>1930.11</td>
</tr>
<tr>
<td>0.9</td>
<td>1963.701</td>
</tr>
<tr>
<td>1.0</td>
<td>2097.753</td>
</tr>
</tbody>
</table>

**Table 3.5: Best selected model using Akaike Weights**

<table>
<thead>
<tr>
<th>Robust AIC</th>
<th>$\delta_{AIC}$</th>
<th>$w_i(AIC)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3190.12</td>
<td>1260.01</td>
<td>2.47E-274</td>
</tr>
<tr>
<td>2801.927</td>
<td>871.817</td>
<td>4.87E-190</td>
</tr>
<tr>
<td>2475.304</td>
<td>545.194</td>
<td>4.10E-119</td>
</tr>
<tr>
<td>2496.164</td>
<td>566.054</td>
<td>1.21E-123</td>
</tr>
<tr>
<td>2479.504</td>
<td>549.394</td>
<td>5.02E-120</td>
</tr>
<tr>
<td>2452.204</td>
<td>522.094</td>
<td>4.25E-114</td>
</tr>
<tr>
<td>2193.253</td>
<td>263.143</td>
<td>7.23E-58</td>
</tr>
<tr>
<td>1930.11</td>
<td>0</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>1963.701</td>
<td>33.591</td>
<td>5.08E-08</td>
</tr>
<tr>
<td>2097.753</td>
<td>167.643</td>
<td>3.95E-37</td>
</tr>
</tbody>
</table>
The Akaike weights can be calculated from AIC values and provide the information about which model is the best. As shown in Table 3.4, the $\alpha=0.8$ has the lowest AIC score. Similarly, Table 3.5 shows that, based on robust AIC score, the Akaike weight for $\alpha=0.8$ appears to be the best model compared to other $\alpha$ values.

![Figure 3.5: rLOWESS fit with $\alpha=0.8$](image)

### 3.3 Summary

LO(W)ESS is widely used in different application areas such as for normalization and accessing non-linear relationships between variables and considered as one of the important member of non-parametric regression in statistical circle. LO(W)ESS fit is significant technique for investigating short term and long term structures in the dataset.

As a scatterplot smoother LO(W)ESS fit is helpful in exploring bivariate and
multivariate data, discovering complex relationship between variables and regression analysis within the empirical data. Therefore, it can be considered as a fairly comprehensive technique for modelling functional dependences in variety of data analysis.

It is unfortunate that despite its wide application area, the important parameters are selected on trial and error basis. Over-smoothing and under-smoothing is neither acceptable nor desirable in such situations. Over-smoothing divulges trend but ignores local variations whereas under-smoothing results in too many local variations.

An automatic approach for selection of smoothing parameters for LO(W)ESS fit has been proposed and tested. The degree of polynomial and presence of outliers is used to select the type of LO(W)ESS. Also, the best value of smoothing parameter is chosen based on the least value of AIC values. AIC with mm-estimator is employed for the selection of best model for smoothing parameters that works well in the presence of outliers. Also, weighted MSE is used for the estimation of best degree of polynomial. The accuracy of the aforementioned methodology has been tested and demonstrated using experimental results.

In the first experiment, an artificial dataset has been generated to test if the proposed method works as per expectations. It is known in advance that the data is linear with the presence of outliers in it. A real dataset from Rock Creek River is examined using the proposed method. Our proposed method is able to automate the smoothing parameter and degree of polynomial. At the same time, it eliminates the problem of over-smoothing and under-smoothing of data. The approach is flexible and easy to implement in a variety of situations.
Chapter 4

Experimental Research: Change point detection and Trend analysis

In Chapter 2, we surveyed the current literature on three techniques: LOWESS, change point analysis and trend, and control chart patterns.

This chapter analyzes the current and the best approaches of change point detection and trend, along with the interaction between change point and trend. It begins by comparing modern methods of change point detection, in Section 4.1, as well as trend analysis, in Section 4.2. Finally, in Section 4.3, the importance of the relationship between change point and trend is illustrated.

Flawed results are the main scientific problems that stipulate the research hypothesis and either prove or disprove the hypotheses. These problems then help in answering these research questions of Thesis: which method is both computationally efficient and accurate in identifying correct numbers and locations of change points? Is there really a change point or trend? Has trend masked the change point or has change point masked the trend?
4.1 Change point

Change point detection is the identification of abrupt variations in behavior due to distributional or structural changes. As mentioned in Chapter 2, the research has focused on the latest and the best techniques being used in the detection of change points which by virtue of current literature, appear to be the most widely used and the current algorithms: Bayesian Change point Analysis (BCP), Wild Binary Segmentation (WBS), E-Agglomerative algorithm (E-Agglo.), and Iterative Robust Detection (IRD). The motive is to study, categorize, and examine the above mentioned change point methods and draw inference on their functionality and effectiveness. Literature analysis shows that different methods often give different numbers and locations of change points. The intention here is to analyze which methods are the most capable of finding the exact number and locations of change points. The power and accuracy of these methods are measured using simulated data. Conclusions are drawn on the functionality and usefulness of each method.

4.1.1 Experimental Validation

The effectiveness and accuracy of the methods can be estimated using true positive rate (TPR) [52, 57]. Equation 4.1 shows the methods for calculation of TPR.

\[
TPR = \begin{cases} 
\frac{n_{cr}}{n_{cp}} 
\end{cases} 
\]

where \(n_{cr}\) = the number of times change points are correctly detected, \(n_{cp}\) = total number of change points.
4.1.2 Experimental Design

To demonstrate the performance and effectiveness of change point detection methods, synthetic data are generated with 1000 data points. The reason for generating artificial data is that the number and locations of change points can be positioned in advance, which greatly simplifies the experiments. In order to test the above mentioned change point techniques, five different type of data has been generated and the methods tested it to see if they are able to detect the exact numbers and locations of change points.

The following five synthetically generated datasets with manually inserted change points are used.

◊ **Simulation 1**: Auto-regressive model is used to generate the dataset

\[
y(t) = \begin{cases} 
0.7y(t - 1) - 0.4y(t - 2) + \varepsilon; 
\end{cases} \quad (4.2)
\]

where \(\varepsilon\) is a noise with varying \(\mu\) values and constant \(\sigma^2\). The change points are inserted at step 100 (\(\mu=1.5, \sigma^2=1.5\)), 200 steps later at location 300 (\(\mu=2.5\) and \(\sigma^2=1.5\)), 300 steps later at location 600 (\(\mu=1, \sigma^2=1.5\)), 200 steps later at location 800 (\(\mu=2.5, \sigma^2=1.5\)) and 200 steps later at location 1000 (\(\mu=1.5, \sigma^2=1.5\)).

**Analysis**: The results from Simulation 1 in Table 4.1(a) show that all the methods are capable of detecting change in the mean of data. In order to test the computational efficiency, WBS, BCP and E-Agglomerative methods are
tested using 100 random iterations, as shown in Table 4.1(b) using R packages.

Table 4.1(b) 10% and 25% represents the number of variations from data points considered to be near the actual change points. 10% and 25% represents the number of variations from data points considered to be near the actual change points. For example, if a change point is at location 200 then the nearest locations of the actual change points are considered as 200 ± 10. Any number lies between this ranges i.e. 180 to 210 is considered as the actual change point. Similarly, a number lies between the range 175 to 225 is considered to be the part of 25%. The number to calculate the 10% and 25% is the difference
between the change point locations. For example, the first change point is at 100 so; the 10% of 100 is calculated. Second change point range is considered as $300 - 100 = 200$ and 10% of 200 is calculated.

Table 4.1: Change in Mean

(a) Change point Locations

<table>
<thead>
<tr>
<th>Method</th>
<th>Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBS</td>
<td>99, 299, 601, 808</td>
</tr>
<tr>
<td>$\alpha=1$</td>
<td>100, 300, 602, 809</td>
</tr>
<tr>
<td>$\alpha=2$</td>
<td>100, 300, 602, 809</td>
</tr>
<tr>
<td>BCP</td>
<td>99, 298, 601, 808</td>
</tr>
<tr>
<td>Robust Detection</td>
<td>100, 300, 602, 808</td>
</tr>
</tbody>
</table>

(b) Results from 100 iterations

<table>
<thead>
<tr>
<th>Method</th>
<th>10%</th>
<th>25%</th>
<th>Missed</th>
<th>Extra</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBS</td>
<td>67</td>
<td>97</td>
<td>8</td>
<td>1</td>
<td>32.53</td>
</tr>
<tr>
<td>$\alpha=1$</td>
<td>70</td>
<td>91</td>
<td>11</td>
<td>10</td>
<td>36769.08</td>
</tr>
<tr>
<td>$\alpha=2$</td>
<td>72</td>
<td>92</td>
<td>7</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>BCP</td>
<td>80</td>
<td>100</td>
<td>10</td>
<td>80</td>
<td>270.53</td>
</tr>
</tbody>
</table>

The variations are computed using TRP, using Equation 4.1; “missed” represents the number of times actual change points locations are missed, and “extra” represents the number of times phantom change points are detected. From our experiments, it was observed that WBS appears to have the best combination of accuracy and speed. BCP, because of its sensitivity, generates unnecessary
locations and numbers. The E-Agglomerative produces accurate locations and numbers but it is computationally cumbersome.

♦ Simulation 2: A change point is added at every 250 steps between 1 and 1000. Finally, noise with mean 0 and variance 4 is added. The change points are at steps 250, 500, and 750.

![E-Agglomerative](image1.png) ![WBS](image2.png) ![BCP](image3.png)

(a) E-Agglomerative (b) WBS (c) BCP

Figure 4.2: Change point detection

Analysis: Figure 4.2 shows the change points based on iteration, given in Table 4.2(a). The results in Table 4.2(a) from Simulation 2 show that all the methods are capable of detecting step change in the data while Table 4.2(b) shows the results from 100 iterations of WBS, BCP and E-Agglomerative methods. In Simulation 2, WBS appears again to be computationally faster than
BCP and E-Agglomerative.

Table 4.2: Step data with constant mean and variance

(a) Change point Locations

<table>
<thead>
<tr>
<th>Method</th>
<th>Change point Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBS</td>
<td>250, 501, 758</td>
</tr>
<tr>
<td>α=1 Divisive</td>
<td>251, 502, 759</td>
</tr>
<tr>
<td>α=2 Divisive</td>
<td>251, 502, 759</td>
</tr>
<tr>
<td>BCP</td>
<td>240, 501, 752</td>
</tr>
<tr>
<td>Robust</td>
<td>250, 501, 758</td>
</tr>
</tbody>
</table>

(b) Results from 100 iterations

<table>
<thead>
<tr>
<th>Method</th>
<th>10%</th>
<th>25%</th>
<th>Missed</th>
<th>Extra</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBS</td>
<td>86</td>
<td>98</td>
<td>22</td>
<td>10</td>
<td>35.78</td>
</tr>
<tr>
<td>α=1 Divisive</td>
<td>84</td>
<td>97</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>α=2 Divisive</td>
<td>85</td>
<td>98</td>
<td>3</td>
<td>4</td>
<td>27350.66</td>
</tr>
<tr>
<td>BCP</td>
<td>80</td>
<td>87</td>
<td>33</td>
<td>27</td>
<td>267.2</td>
</tr>
</tbody>
</table>

As described earlier, 10% and 25% represent the ranges of variations based on the change points. As the change point locations are 250, 500, and 750 the 10% range for first change point is 225 to 275, for second change point it is 475 to 525 and for third change point it is 725 to 775. Similarly, the ranges for 25% are 187.5 to 312.5 for first change point, second change point range is 437.5 to 562.5, and the range for third change point is 687.5 to 812.5.

♦ Simulation 3: In this test, data points are generated, from Equation 4.2. A
change point is added at 200, 300, 400, 100 steps. The noise is added with mean 0, and a different variance at steps 200($\mu=1.5$ and $\sigma^2=2.5$), 300($\mu=1.5$, $\sigma^2=3.5$), 400($\mu=1.5$, $\sigma^2=4.5$), 100($\mu=1.5$, $\sigma^2=5.5$) and locations are 200, 500, 900 and 1000.

(a) E-Agglomerative  
(b) WBS  
(c) BCP

Figure 4.3: Change point detection

Analysis: Figure 4.3 shows the change points based on iteration, given in Table 4.3(a). The results, in Table 4.3(a), from Simulation 3 show that none of the methods are able to detect change in the variance of data. While performing this experiment, it was attempted to change the variance value but no method has the ability to detect change in the variance. The change points ranges for 10% are 180 to 220, 470 to 530, 860 to 940, and 990 to 1010. Similarly,
ranges for 25% are 150 to 250, 425 to 575, 800 to 1000, and 975 to 1025. In order to reduce the computation time for E-agglomerative method, the code was parallelized using “doParallel” package in R.

Table 4.3: Change in variance

(a) Change point Locations

<table>
<thead>
<tr>
<th>Method</th>
<th>Change point Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBS</td>
<td>No change point detected</td>
</tr>
<tr>
<td>α=1</td>
<td>574</td>
</tr>
<tr>
<td>α=2</td>
<td>No change point detected</td>
</tr>
<tr>
<td>BCP</td>
<td>No change point detected</td>
</tr>
<tr>
<td>Robust detection</td>
<td>No change point detected</td>
</tr>
</tbody>
</table>

(b) Results from 100 iterations

<table>
<thead>
<tr>
<th>Method</th>
<th>10%</th>
<th>25%</th>
<th>Missed</th>
<th>Extra</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBS</td>
<td>8</td>
<td>20</td>
<td>81</td>
<td>59</td>
<td>32.71</td>
</tr>
<tr>
<td>α=1</td>
<td>4</td>
<td>10</td>
<td>99</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>α=2</td>
<td>1</td>
<td>2</td>
<td>11</td>
<td>6</td>
<td>14009.17</td>
</tr>
<tr>
<td>BCP</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>221.1</td>
</tr>
</tbody>
</table>

◊ Simulation 4: In this test, data points are generated using

\[
y(t) = \begin{cases} 
sin(w) + \varepsilon; \\
\end{cases}
\]

where \( w \) is randomly generated data and \( \varepsilon \) is noise with \( \sigma^2 = 0.8 \).

Analysis: Figure 4.4 shows the change points based on iteration, given in
Table 4.4(a). The change points are added at every 250 steps i.e. at location 250, 500 and 750. The results, in Table 4.4(a), from Simulation 4 show that robust detection is not able to detect change points. The change point’s ranges for 10% are 225 to 275, 475 to 525, and 725 to 775 for first, second, and third change points respectively. Similarly, ranges for 25% are 187.5 to 312.5, 437.5 to 562.5, and 687.5 to 812.5 for first, second, and third change points respectively. The results, in Table 4.4(b), show that WBS again outperforms.

◊ **Simulation 5**: In this test, the data points are generated using the non auto-regressive (AR) model. The dataset was generated using exponential steep
Table 4.4: Sine Wave data

(a) Change point Locations

<table>
<thead>
<tr>
<th>Method</th>
<th>Change point Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBS</td>
<td>272, 499, 739</td>
</tr>
<tr>
<td>α=1 Divisive</td>
<td>273, 524, 739</td>
</tr>
<tr>
<td>α=2 Divisive</td>
<td>273, 500, 739</td>
</tr>
<tr>
<td>BCP</td>
<td>272, 523, 743</td>
</tr>
<tr>
<td>Robust detection</td>
<td>No change point detected</td>
</tr>
</tbody>
</table>

(b) Results from 100 iterations

<table>
<thead>
<tr>
<th>Method</th>
<th>10%</th>
<th>25%</th>
<th>Missed</th>
<th>Extra</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBS</td>
<td>83</td>
<td>96</td>
<td>19</td>
<td>4</td>
<td>36.24</td>
</tr>
<tr>
<td>α=1 Divisive</td>
<td>85</td>
<td>94</td>
<td>9</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>α=2 Divisive</td>
<td>88</td>
<td>95</td>
<td>9</td>
<td>5</td>
<td>20768.36</td>
</tr>
<tr>
<td>BCP</td>
<td>60</td>
<td>70</td>
<td>70</td>
<td>90</td>
<td>221.1</td>
</tr>
</tbody>
</table>
decay, Equation 4.4.

\[
y = \begin{cases} 
c x^n; & (4.4) 
\end{cases}
\]

Here, n is generated randomly. If \( c > 1 \) then it generates exponential growth whereas if \( c < 1 \) then it generates exponential decay. Finally, noise was added to generate exponential steep decay to test change points for non AR data. The change points are added at 300, 100, 200, and 250 steps. The noise is added with mean 0 and variance 2. The exact locations of change points are 300, 400, 600 and 850.

![Figures](image_url)

(a) E-Agglomerative

(b) WBS

(c) BCP

Figure 4.5: Change point detection

**Analysis:** Figure 4.5 shows the change points based on iteration, given in Table 4.5(a). The change point’s ranges for 10% are 270 to 330, 390 to 410,
580 to 620, and 825 to 875 for first, second, third, and fourth change points respectively. Similarly, ranges for 25% are 225 to 375, 375 to 425, 550 to 650, and 787.5 to 912.5 for first, second, third and fourth change points respectively.

The results in Table 4.5(a), from Simulation 5, show that \textit{wbs, E-divisive and bcp} shows change points at closest locations whereas \textit{robust detection} shows the different locations of change points.

Table 4.5: Non AR

(a) Change point Locations

<table>
<thead>
<tr>
<th>Method</th>
<th>Change point Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBS</td>
<td>288, 400, 595, 848</td>
</tr>
<tr>
<td>(\alpha=1) Divisive</td>
<td>288, 401, 596, 849</td>
</tr>
<tr>
<td>(\alpha=2) Divisive</td>
<td>288, 401, 596, 849</td>
</tr>
<tr>
<td>BCP</td>
<td>285, 400, 596, 848</td>
</tr>
<tr>
<td>Robust detection</td>
<td>206, 399, 560, 594</td>
</tr>
</tbody>
</table>

(b) Results from 100 iterations

<table>
<thead>
<tr>
<th>Method</th>
<th>10%</th>
<th>25%</th>
<th>Missed</th>
<th>Extra</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBS</td>
<td>91</td>
<td>98</td>
<td>10</td>
<td>9</td>
<td>36.24</td>
</tr>
<tr>
<td>(\alpha=1) Divisive</td>
<td>87</td>
<td>94</td>
<td>15</td>
<td>17</td>
<td>20690.36</td>
</tr>
<tr>
<td>(\alpha=2) Divisive</td>
<td>89</td>
<td>95</td>
<td>15</td>
<td>8</td>
<td>20690.36</td>
</tr>
<tr>
<td>BCP</td>
<td>70</td>
<td>75</td>
<td>65</td>
<td>90</td>
<td>221.1</td>
</tr>
</tbody>
</table>
4.2 Trend Analysis

Trend analysis is defined as a process of estimating gradual change in future even from past data. Two non-parametric methods, i.e. Mann-Kendall (MK) test and Cox-Stuart tests are used to detect trend in the time series. For experiment, we have used the same dataset generated in Section 4.1 was used. In this Section 4.2, the trend on the synthetically generated data is tested.

4.2.1 Experimental Design

The trend detection method, MK, is applied to the data simulated in Section 4.1. Figure 4.9 shows the LOWESS fit using MK, from all the simulations and the summary of results are given in Table 4.6.

Table 4.6: Summary: Trend detection

<table>
<thead>
<tr>
<th>Experiment</th>
<th>τ</th>
<th>P-Value</th>
<th>α</th>
<th>Test Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation 1</td>
<td>0.382</td>
<td>$\leq 2.22e-16$</td>
<td>0.05</td>
<td>No Trend</td>
</tr>
<tr>
<td>Simulation 2</td>
<td>-0.0665</td>
<td>$= 0.0016443$</td>
<td>0.05</td>
<td>No Trend</td>
</tr>
<tr>
<td>Simulation 3</td>
<td>0.0118</td>
<td>$=0.57692$</td>
<td>0.05</td>
<td>Increasing Trend</td>
</tr>
<tr>
<td>Simulation 4</td>
<td>-0.00544</td>
<td>$=0.79689$</td>
<td>0.05</td>
<td>Decreasing Trend</td>
</tr>
<tr>
<td>Simulation 5</td>
<td>-0.296</td>
<td>$\leq 2.22e-16$</td>
<td>0.05</td>
<td>No Trend</td>
</tr>
</tbody>
</table>

**Simulation 1**: As the computed p-value is lower than the significance level $\alpha = 0.05$, one fail to reject the null hypothesis $H_0$, and reject the alternative hypothesis $H_a$.

Figure 4.9(a) shows the MK LOWESS fit with p-value and $\tau$. 
Simulation 2: As the computed p-value is lower than the significance level $\alpha = 0.05$, one fail to reject the null hypothesis $H_0$, and reject the alternative hypothesis $H_a$. Figure 4.9(b) shows the MK LOWESS fit with p-value and $\tau$.

Simulation 3: As the computed p-value is greater than the significance level $\alpha = 0.05$, one should accept alternative hypothesis $H_a$, and reject the null hypothesis $H_0$. Since the value of $\tau$ is positive, this means the trend is increasing. Figure 4.9(c) shows the MK LOWESS fit with p-value and $\tau$.

Simulation 4: As the computed p-value is greater than the significance level $\alpha = 0.05$, one should accept alternative hypothesis $H_a$, and reject the null hypothesis $H_0$. Since the value of $\tau$ is negative this means that the trend is decreasing. Figure 4.9(d) shows the MK LOWESS fit with p-value and $\tau$.

Simulation 5: As the computed p-value is lower than the significance level $\alpha = 0.05$, one fail to reject the null hypothesis $H_0$, and reject the alternative hypothesis $H_a$. Figure 4.9(e) shows the MK LOWESS fit with p-value and and $\tau$.

4.3 Interaction between Change point and Trend

As the gradual and the abrupt change may exists together, sometimes, it is hard to distinguish them using statistical tests. For example, in hydrological time series data, if a sudden release of water from an impoundment followed by a return to a prior regime, it appears as if the sharp downward trend may be misinterpreted as a series of change points. Analyzing only change point and overlooking trend or vice-
versa may mislead the calculations. For the improved prediction for future events, it is extremely important to analyze the relationship between the trend and the change point.
4.3.1 Trend masked Change point

Experimental Design

There may be some situations in a time series when a gradual change has covered-up the abrupt changes. For example, if both trend and change point are present in the time series and estimating only trend could imply that trend is either increasing or decreasing, but not why and when it is increasing. Analyzing both trend and change point can make improved predictions in the time series analysis. This will assist in verifying the stated hypothesis. The purpose of this simulation was to demonstrate the possibility of a change point masking trend, even though that the original Figure 4.7 could have been resolved visually.

Figure 4.7: Hidden Change Point and Trend

For assessing the relationship between trend and change point, experimental steps are outlined in Table 4.7.

Step 1: Presence of Change points

wbs, bcp methods have been used for detection of change points in our artificial
Table 4.7: Experimental Procedure

<table>
<thead>
<tr>
<th>Procedure 1</th>
<th>Experimental Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Detect the presence of change points</td>
</tr>
<tr>
<td>2</td>
<td>Analyze the dataset for trend, separately</td>
</tr>
<tr>
<td>3</td>
<td>Divide the data into segments based on change points</td>
</tr>
<tr>
<td>4</td>
<td>Analyze those segments for the presence of trend</td>
</tr>
</tbody>
</table>

dataset. According to these methods, a change point appeared at location 251 as shown in Figure 4.8.

**Step 2: Analysis of Trend**

![Figure 4.8: Change Point location $\delta = 251$](image)

The data for trend was analyzed as shown in Figure 4.9. According to MK and Cox Stuart method, a downward trend is predicted when the complete dataset is analysed. If it is considered as the final conclusion, it is a misleading result.
Step 3: Segment data based on Change points

When divided the dataset into segments and testing them before and after the change point for trend analysis one gets the actual picture. There is only one change point at location 250 in the dataset.

Step 4: Trend analysis on Segmentation

When the first part is analyzed, according to MK and Cox Stuart, an upward trend is predicted, as shown in Figure 4.10(a) whereas the second half shows no trend in the dataset, in the Figure 4.10(b). For optimal analysis of data, a piecewise LOWESS curve was used as shown in Figure 4.11.

From this example it is clear that there was an upward trend before the change
Due to some abrupt changes, where there could be a change point, the trend changes. That is, instead of considering the whole trend as decreasing, one should conclude that change point has moved the direction of trend. Using repeated Monte-Carlo simulations, it was found that in 72% of the cases the trend has affected the location of change points.

### 4.3.2 Change point masked Trend

As described in Section 4.3, there may be situations when trend mask change point. Similarly, there may be situation when change point mask trend. The same step as described in Table 4.7 is used to show a situation where change point masks trend. In order to test if change point can mask trend in the dataset 200 data points were generated using exponential steep decay.
Experimental Design

Step 1: Presence of Change points

*wbs, E-Divisive* methods were used for the detection of change points in our artificial dataset. According to these methods, change points appeared at location 67, 94, 113, 160, and 175 as shown in Figure 4.12.

Step 2: Analysis of Trend

The data for trend was analyzed as shown in Figure 4.13. In this dataset, MK and Cox-Stuart could not detect actual trend. According to Mann-Kendall, the p-value is \( \leq 2.22 \times 10^{-16} \). Since the p-value is less than significance level \( \alpha = 0.05 \), it shows that we fail to reject the null hypothesis that there is no evidence of a trend.

Step 3: Segment data based on Change points

When the dataset is divided into segments and analysing those segments for trend before and after the change point we get the actual picture.

Step 4: Trend analysis on Segmentation

The data is analysed using piece-wise MK based on presence of change points before
and after. According to “wbs” and “E-Divisive”, there are six locations of change points and the summary of the trends based on those change points are shown in Table 4.8.

Figure 4.14 shows the segmented change-points in the data. According to the step 4 in Table 4.7, the data was tested before and after each change point. Figure 4.15 shows a typical result trend, change points, piece-wise trend with LOWESS, and piece-wise LOWESS. Using the Monte-Carlo simulations, it was found that 78% of cases the change points has marked the change in direction of the trend

4.4 Summary

In this chapter we first formulated the comparisons of four prominent change point detection methods using five different simulations studies. From the experimental results of Simulation 1 and Simulation 2, it is observed that all the methods are capable of detecting correct numbers and locations of change points. Testing the
data from the Simulation 1 and Simulation 2 on 100 random iterations, WBS appears to be the best combination of accuracy and speed. BCP generates unnecessary change points because of its sensitivity. On the other hand, the E-agglomerative produces accurate numbers and locations but it is computationally inefficient.

In Simulation 3, none of the chosen methods is able to detect change in the variance of the data. In Simulation 4, all the methods except Robust detect are able to detect the specified change point locations in the dataset. Testing on the random iterations, again WBS works better compared to BCP and E-agglomerative for this dataset. The results from Simulation 5, it is perceived that all the methods except robust detection are able to detect correct locations of change points. Robust detection method detected the location of change points but that do not cover the criteria of 10% or 25%. Again the random iteration in Simulation 5 shows that WBS

<table>
<thead>
<tr>
<th>Experiment</th>
<th>$\tau$</th>
<th>P-Value</th>
<th>$\alpha$</th>
<th>Test Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change point Location 1</td>
<td>0.262</td>
<td>$=0.0014981$</td>
<td>0.05</td>
<td>Increasing Trend</td>
</tr>
<tr>
<td>Change point Location 2</td>
<td>0.259</td>
<td>$=0.060625$</td>
<td>0.05</td>
<td>Increasing Trend</td>
</tr>
<tr>
<td>Change point Location 3</td>
<td>0.31</td>
<td>$=0.068873$</td>
<td>0.05</td>
<td>Increasing Trend</td>
</tr>
<tr>
<td>Change point Location 4</td>
<td>0.0254</td>
<td>$=0.83811$</td>
<td>0.05</td>
<td>Increasing Trend</td>
</tr>
<tr>
<td>Change point Location 5</td>
<td>-0.273</td>
<td>$=0.27576$</td>
<td>0.05</td>
<td>Decreasing Trend</td>
</tr>
<tr>
<td>Change point Location 6</td>
<td>-0.0667</td>
<td>$=0.76653$</td>
<td>0.05</td>
<td>Decreasing Trend</td>
</tr>
<tr>
<td>Change point Location 7</td>
<td>-0.107</td>
<td>$=0.49229$</td>
<td>0.05</td>
<td>Decreasing Trend</td>
</tr>
</tbody>
</table>
The chapter also explored the trend detection methods on different simulated methods. Mann-Kendall and Cox-Stuart methods were used for analysis.

Finally, this chapter articulate the statistical issues involved when both abrupt and gradual changes are present. Keeping this in mind, different case studies using synthetically generated dataset were generated and tested for abrupt changes. Even though experimental examples given are simple ones which could be noticed by a visual assessment, these demonstrate that the automatic techniques are far from
foolproof. A LO(W)ESS fit is used for predicting the presence of trend. Through the experimental examples, the hypothesis is accepted that trend masks the change point and change point masks the trend. Using repeated Monte-Carlo simulations, it is found that 72% of times the trend has affected the location of change points and 78% of cases the change points has marked the change in direction of the trend.
Chapter 5

Experimental Research: Control Chart Patterns

The focus is to assess the proposed methodology for control chart patterns. Pattern detection using control charts is a statistical process that determines whether a natural/unnatural pattern exits. In this chapter, we discuss control chart patterns, methodology for the detection of concurrent patterns, and experimental results of the proposed methods.

Section 5.2, focuses on different types of natural/unnatural patterns along with their statistical characteristics for differentiation. In Section 5.3, the proposed method for the detection of concurrent patterns using control charts is discussed. In Section 5.4, demonstrates the experimental results using the proposed methodology and finally, in Section 5.5, summarizes this chapter.

5.1 Research Objective

This Thesis gives consideration to the detection of concurrent patterns in a time series. The objective is to develop a model that accurately identifies mixture patterns and handles mixed type data. To achieve this objective, RobustICA is employed along with decision trees (DT). The specific objectives are the following:
(a) Extraction of patterns from mixed pattern using RobustICA, and

(b) Application of DT for pattern identification

5.2 Control Chart Patterns

Control chart patterns are a useful tool to detect usual or unusual patterns in a time series [30, 90, 31]. In hydrological time series, heavy rain, snowfall, drought or flood like situations may arise which lead to unnatural patterns such as trend up or down, shift up or down, cyclic, stratification, and systematic. The natural/unnatural patterns can be differentiated by the feature based approach [30]. It is considered that decisions from the feature based approach are straightforward and understandable by the users. The feature based approach first extracts features and then uses those features for control chart pattern detection [90]. Different patterns have the following characteristics.

1. **Natural/normal pattern (N):** this pattern shows that the nature of the natural process where points lies within $\pm 3\sigma$, defined as upper control limit (UCL) and lower control limit (LCL). In this pattern, points alter at random and most of the points are close to the mean, center line (CL), and few points fall near the control limits. The distribution is smooth and unimodal, neither flat nor skewed. Figure 5.1 shows an example of a normal pattern.

2. **Sudden Shift (SD or SU):** A sudden shift in the process level is revealed by either positive or negative change in only one direction, where data points
start to appear on either upper side (positive) or lower side (negative) of the chart only. With the new process mean value, the process behaves randomly or naturally. Plotting the two periods, with old process mean and new process mean, different distribution for both periods is detected. Figure 5.2 shows an example of Shift pattern.

3. **Stratification (SR):** Stratification can be defined as mixture of data point with small variations and lies within the control limits. This pattern seems the
presence of set of more data point hugging the centre line and absence of data points near the control limits. Based on these characteristics, it appears that control limits are very wide and the variations between the samples are very less. This can also be analysed as a substantial decrease in the variance of the control points. Figure 5.3 shows an example of Stratification pattern.

4. **Cyclic (C):** Cyclic pattern is formed when most of the points are present near the upper or lower end of the control chart. There is absence of points near the centre line. The pattern appears as crest and trough of a wave.
Figure 5.4: Cyclic Pattern

Figure 5.4 shows an example of Cyclic pattern.

5. **Systematic (SY)**: The systematic pattern is observed when data points vary above and below the center line. It is comparatively flat-topped with larger variations compared to normal pattern. Figure 5.5 shows an example of Systematic pattern.

Figure 5.5: Systematic Pattern

6. **Trend up/down (TD or TU)**: The trend up can be defined as constant
rise of successive points in positive direction whereas the trend down can be
defined as constant fall of successive points in negative direction. In both cases,
trend up and trend down, the sample points appear on one side of the chart
only. Figure 5.6 shows an example of trend up and trend down where points
are going away from the CL. For trend up, the points rise towards the UCL
whereas for trend down, the points fall towards the LCL.

5.3 Proposed Methodology

The Algorithm 4 shows the steps for the detection of different unnatural patterns.

**Algorithm 4** Pattern Detection Algorithm

1: procedure PATTERNDETECT

2: Determine the sample length $N$

3: Input the sample, mixed pattern data, to RobustICA to identify ICs

4: Extract features from the control chart pattern feature identification method

5: Input the extracted features to DT for the classification of individual pattern
Figure 5.7 shows the proposed flow diagram for detection of concurrent patterns.

Figure 5.7: Proposed Model: Concurrent Pattern Recognition

The control chart patterns are generated using the equations in Table 5.1.

Table 5.1: Parameters for Control Chart Patterns

<table>
<thead>
<tr>
<th>Pattern type</th>
<th>Parameters /Values</th>
<th>Pattern Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal (N)</td>
<td>Mean ((\mu) = 80) Standard deviation ((\sigma) = 5)</td>
<td>(y_i = \mu + r_i\sigma)</td>
</tr>
<tr>
<td>Stratification (SR)</td>
<td>Random noise ((\sigma') = 0.2\sigma \text{ to } 0.4\sigma)</td>
<td>(y_i = \mu + r_i\sigma')</td>
</tr>
<tr>
<td>Systematic(SY)</td>
<td>Systematic Departure (d) = (1\sigma \text{ to } 3\sigma)</td>
<td>(y_i = \mu + r_i\sigma + d \ast (-1)^i)</td>
</tr>
<tr>
<td>Cyclic (C)</td>
<td>Amplitude ((a) = 1.5\sigma \text{ to } 2.5\sigma) Period ((T) = 8 \text{ and } 16)</td>
<td>(y_i = \mu + r_i\sigma + a \sin\left(\frac{2\pi i}{T}\right))</td>
</tr>
<tr>
<td>Increasing Trend(TU)</td>
<td>Gradient ((g) = 0.05(\sigma) \text{ to } 0.1(\sigma))</td>
<td>(y_i = \mu + r_i\sigma + ig)</td>
</tr>
<tr>
<td>Decreasing Trend(TD)</td>
<td>Gradient ((g) = 0.05(\sigma) \text{ to } 0.1(\sigma))</td>
<td>(y_i = \mu + r_i\sigma - ig)</td>
</tr>
<tr>
<td>Upward Shift (SU)</td>
<td>Shift magnitude ((s) = 1.5(\sigma) \text{ to } 2.5(\sigma)) Shift position ((P) = 9, 17, 25)</td>
<td>(y_i = \mu + r_i\sigma + ks)</td>
</tr>
<tr>
<td>Downward Shift (SD)</td>
<td>Shift magnitude ((s) = 1.5(\sigma) \text{ to } 2.5(\sigma)) Shift position ((P) = 9, 17, 25)</td>
<td>(y_i = \mu + r_i\sigma - ks)</td>
</tr>
</tbody>
</table>

In this table, \(\mu\) is the nominal mean value, \(\sigma\) is the standard deviation, \(\sigma'\) is
a function for normally distributed random numbers, $a$ is the amplitude of cyclic variations and $T$ is the period of a cycle, $g$ is the gradient of an increasing or a decreasing trend pattern, $p$ indicates the shift position in an upward or a downward shift pattern and $s$ is the magnitude of the shift, and $i$ is the discrete time for sample, and $y_i$ is the value of the sampled data point at time $i$.

### 5.3.1 RobustICA

Independent Component Analysis (ICA) is a technique to decompose statistically independent sources from each other. An ICA has wide application area such as image processing, face recognition, multivariate data processing, and time series prediction [92]. It works on the criteria of “Blind source separation” which means separating patterns from mixed patterns without any knowledge of the mixing process. The ICs extracted from the mixture patterns work as independent sources that can be used for pattern detection. Patterns/signals are extricated based on non-Gaussian statistical properties: kurtosis and negentropy. Different versions of ICA: ICA, fastICA and RobustICA are available for analysis [92, 93].

#### Why RobustICA

RobustICA is a powerful algorithm for independent component analysis. It is a modified version of fastICA and is kurtosis based. Compared to ICA, fastICA and other kurtosis based algorithms, RobustICA outperforms as the pre-whitening step is not done, which in turn make it computationally efficient [92, 93]. In the pre-whitening, a new vector is created from the observed vector which is uncorrelated
and the variance is unity. Moreover, it is cost-effective, robust and widely used in real world applications.

5.3.2 Feature Extraction

As described in Section 5.2, the unusual and natural patterns can be distinguished from each other based on feature characteristics. According to Gauri [32], the performance of the feature selection technique decreases if feature selection methods are highly correlated. In this work, correlation analysis is not required because we have employed features selection method which is known to have low correlation with one another. Based on feature characteristics of control chart patterns, four feature candidates for pattern identification are listed below:

1. **Slope of least square regression and Ratio between variance of the observation (RVE))** [33]: If the value of the slope of least square regression line is 0 then it signifies $N$, $SR$, $SY$, and $C$ pattern otherwise, if the slope of least square regression line is greater than 0 then it signifies $TU$, $SU$, $TD$, and $SD$ pattern. The positive and negative sign signifies whether the trend/shift is upward or downward. According to Gauri, RVE is more powerful compared to the absolute value [32]. Positive value shows that trend or shift is upward whereas negative values shows shift or trend is downward. Equation 5.1 shows
the mathematical expression of calculation of RVE.

\[ RVE = \left\{ \frac{\sum_{j=1}^{N} (y_{i} - \bar{y})^2}{(N-1)} \right\} \left\{ \frac{\sum_{j=1}^{N} (y_{i} - \bar{y})^2 - \left( \sum_{j=1}^{N} y_{i} (t_{i} - \bar{t}) \right)^2}{\sum_{j=1}^{N} (t_{i} - \bar{t})^2} \right\} \] (5.1)

2. **Sum of mean regression Error** [89, 90]: In order to differentiate trend from shift, the trend pattern have intermediate mean error while shift patterns have highest mean error. This method can be used to differentiate upward trend from upward shift and downward trend from downward shift. Similarly, this rule can be used to differentiate systematic pattern from cyclic pattern. For systematic pattern the value is high whereas for cyclic it is low. Equation 5.2 shows the mathematical expression of calculation of sum of mean regression error (MRE).

\[ MRE = \sum_{j=1}^{N} (y_{t} - \hat{y}_{MN}) \] (5.2)

3. **ACLPI** is defined as Area between the pattern and mean line per interval in terms of \( SD^2 \) [33]: to differentiate \( N \), \( SR \), \( C \) and \( SY \) pattern from each other this method can be applied. If the computed value of ACLPI is lowest then it signifies \( SY \) pattern, if the value is intermediate then it signifies \( SR \) and \( N \) pattern otherwise, the highest value signifies cyclic pattern. ACLPI can be calculated using Equation 5.3.

\[ ACLPI = \left\{ \frac{ACL}{(N-1)} \right\} \left\{ \frac{SD^2}{(N-1)} \right\} \] (5.3)

ACL is defined as the area between the centre line and the pattern. The value of the ACL is computed by adding the areas of the triangles and trapezia formed
by the overall pattern and the centre line [30]. $SD^2$ can be calculated using Equation 5.4.

$$SD^2 = \left\{ \frac{\sum_{j=1}^{N} (y_j - \bar{y})^2}{(N - 1)} \right\}^{1/2}$$ (5.4)

4. **ALSPI** is defined as Area between the overall pattern and the LS line per interval in terms of $SD^2$ [33]: to differentiate SY, N and SR pattern from each other this method can be used. If the computed value of ALSPI is lowest then it signifies SY pattern, if the value is intermediate then it signifies N pattern whereas the highest value of ALSPI signifies SR pattern. ALSPI can be calculated using Equation 5.5.

$$ALSPI = \left\{ \frac{ALS}{(N - 1)} \right\} \frac{SD^2}{(N - 1)}$$ (5.5)

ALS is the area between the fitted line and the overall pattern. Similar to ACL, the value of the ALS is computed by adding the areas of the triangles and trapezia formed by the overall pattern and the LS line [31]. $SD^2$ can be calculated using Equation 5.4.

Here the highest, lowest and intermediate values are in the relative sense. For example, firstly, calculate the ALSPI values for different patterns and then plot a scatter diagram for the calculated values, as shown in Figure 5.8. As described earlier, the four patterns which lie near the centre line can be differentiated using ALSPI. The ALSPI for the systematic pattern has lowest value, normal and stratification patterns have intermediate values and cyclic pattern has highest values. Therefore, the magnitude of ALSPI can help to differentiate systematic, cyclic, normal and
stratification patterns. It is hard to decide manually the boundaries for highest, lowest and intermediate values, so, decision trees can be used for analysis.

![Graph showing ALSPI values for different patterns](image)

Figure 5.8: ALSPI values for different patterns

5.3.3 Decision Tree for pattern recognition

Decision trees are the simplest and most successful method for decision problems [75, 69]. They have been extensively used in different area for analysis, due to their various characteristics: simplicity, comprehensibility, ability to handle mixed-type data and using few parameters [69]. A decision tree describes formally the decisions to be made, the events that may occur, and the outcomes associated with combinations of decisions and events. Decision tree models take as input an object or situation described by a set of properties, and as outputs provides yes/no decision.
They use “divide and conquer” and “top down” approaches for analysis, that can handle a large amount of data in a cost effective way of classification [75]. They start from the root node, that represent the classification problem, and split the tree into branches, which represent the discrete value classifier. According to Othman’s experimental results[69], DT has more detection accuracy when compared to support vector machine and multilayer perceptron neural networks.

As shown in Figure 5.7, the ICs generated by RobustICA will be provided to decision trees for the pattern recognition. The basic characteristics of various patterns used are grouped into different pattern class in the tree structure as shown in Figure 5.9. A judicious ordered selection of features, in this process, can facilitate a good starting point for selection prevention of redundant inclusion of highly correlated features.

The logic behind the defined hierarchical structure can be explained as follows. The eight most commonly observed patterns are grouped into one broad class; where the patterns move away from the centre line, and into another class where the patterns are built around the centre line.

5.4 Experimental Results

To demonstrate the performance and effectiveness of concurrent pattern detection method, synthetic data are generated with 500 data points. The data points are generated using equations given in Table 5.2.
Table 5.2: Concurrent Control Chart Patterns

<table>
<thead>
<tr>
<th>Mixed Pattern type</th>
<th>Pattern Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR+SY</td>
<td>$y_i = \mu + r_i\sigma' + d \ast (-1)^i$</td>
</tr>
<tr>
<td>SY+C</td>
<td>$y_i = \mu + r_i\sigma + d \ast (-1)^i + a\sin\left(\frac{2\pi i}{T}\right)$</td>
</tr>
<tr>
<td>C+T(D/U)</td>
<td>$y_i = \mu + r_i\sigma + a\sin\left(\frac{2\pi i}{T}\right) \pm ig$</td>
</tr>
<tr>
<td>C+S(D/U)</td>
<td>$y_i = \mu + r_i\sigma + a\sin\left(\frac{2\pi i}{T}\right) \pm ks$</td>
</tr>
<tr>
<td>T(D/U)+S(D/U)</td>
<td>$y_i = \mu + r_i\sigma \pm ig \pm ks$</td>
</tr>
<tr>
<td>SY+S(D/U)</td>
<td>$y_i = \mu + r_i\sigma + d \ast (-1)^i \pm ks$</td>
</tr>
<tr>
<td>SR+S(D/U)</td>
<td>$y_i = \mu + r_i\sigma' \pm ks$</td>
</tr>
</tbody>
</table>
Step 1: Determine the sample length $N$

For simplicity of the experiment, 120 data points of mixed pattern were generated, 60 points for stratification pattern and 60 data points for increasing trend. This example demonstrates the steps to detect the concurrent patterns from the proposed method.

Step 2: Input the sample, mixed pattern data, to RobustICA to identify ICs

The artificially generated mixed pattern is input to the RobustICA for the generation of ICs. Figure 5.10 shows the ICs generated by the RobustICA from the input data.

![Figure 5.10: Mixed patterns separated using RobustICA](image)

Step 3: Extract features from the control chart pattern feature identification method

Based on feature characteristics of the patterns, we started from the root node as shown in Figure 5.9. In order to differentiate stratification from increasing trend, RVE is used. Computing RVE value from Equation 5.1 for stratification, the value
for RVE is -0.039. Similarly, the computed value for RVE for increasing trend is 1.648.

**Step 4: Input the extracted features to DT for the classification of individual pattern**

Inputting the computed values to the DT, RVE value is less than 0 signifies that the pattern is stratification and RVE value for increasing trend pattern is greater than 0.

Based on the domain knowledge, the pattern identification is done using four prominent feature detection methods for recognizing basic eight control chart patterns. In short, each pattern has special characteristics which separate one pattern from another. Based on these characteristics, the best feature to differentiate the patterns is selected as shown in Figure 5.9. At the first stage, we differentiate the patterns, using RVE, which go away from the center line, either increasing or decreasing, from the ones which stay near the centre line and warning limits. These patterns, near the center line and warning limit, can be segregated using ACLPI. This helps to separate the patterns such as $C$, $ST$, $SR$ and $N$. Similarly, $SR$ and $N$ can be discriminated using ALSPI, which assist in separating patterns that hugs the center line from the one in which data-points lies between warning limits. Table 5.3 shows the pattern recognition success results.
<table>
<thead>
<tr>
<th>Concurrent Pattern</th>
<th>Accuracy Rate</th>
<th>Concurrent Pattern</th>
<th>Accuracy Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>N + SR</td>
<td>99.50</td>
<td>N + SY</td>
<td>98.20</td>
</tr>
<tr>
<td>N + TU</td>
<td>98.16</td>
<td>N + TD</td>
<td>97.78</td>
</tr>
<tr>
<td>N + SU</td>
<td>97.80</td>
<td>N + SD</td>
<td>98.95</td>
</tr>
<tr>
<td>N + C</td>
<td>100</td>
<td>SR + TU</td>
<td>96.07</td>
</tr>
<tr>
<td>SR + TD</td>
<td>97.29</td>
<td>SR + SU</td>
<td>98.54</td>
</tr>
<tr>
<td>TU + C</td>
<td>98.99</td>
<td>TU + SU</td>
<td>98.61</td>
</tr>
<tr>
<td>TU + SD</td>
<td>98.78</td>
<td>TU + TD</td>
<td>96.48</td>
</tr>
<tr>
<td>SY + C</td>
<td>97.23</td>
<td>SY + SD</td>
<td>100</td>
</tr>
<tr>
<td>SY + SU</td>
<td>99.92</td>
<td>SY + TD</td>
<td>98.12</td>
</tr>
<tr>
<td>SY + TU</td>
<td>96.53</td>
<td>SR + C</td>
<td>97.29</td>
</tr>
<tr>
<td>SR + SY</td>
<td>99.22</td>
<td>SR + SD</td>
<td>99.69</td>
</tr>
</tbody>
</table>

### 5.5 Summary

Control chart patterns (CCPs) are important statistical tool for detecting the presence of unusual patterns. Hydrological data may encounter unusual patterns due to heavy rainfall, snow, or drought like conditions. Correct and timely identification of CCPs is significant since these unusual patterns are associated with the concentration of chemical or metal in the hydrological data. Effective recognition of correct mixture CCPs is a thought provoking task. However, most of the existing methods focus on
the classification of single pattern only which in turn gives erroneous results if mixture patterns exits.

RobustICA and a decision tree based approach have been used to detect eight CCPS. Feature based technique is employed to segregate one pattern from another and can be applied to area application area to assess different patterns. Conversely, some of the existing methods, which focus on concurrent pattern detection, use ICA to differentiate concurrent patterns. ICA has drawbacks compared to RobustICA.

The proposed method first uses RobustICA to generate independent patterns and then computes the features and finally, uses decision tree for pattern recognition. Based on specific feature of pattern, the induction rules of decision tree are generated for specific pattern recognition. The process of feature extraction does not require any experience and skill thus this proposed method is completely automated.

Besides, experimental results, from 100 simulations, indicate that the proposed scheme has achieved more that 98% classification accuracy. Hence, the proposed method is quite promising for recognition of mixture patterns. Also, the proposed scheme can efficiently analyze mixture patterns in time series of medical, financial and any other applications compared to support vector machines, neural networks, and Bayesian network.
Chapter 6

Minimum Maximum properties of LO(W)ESS fit

In 1970s, Hirotugu Akaike used information theory, Akaike Information Criteria (AIC), to build numerical equivalent of Occams razor [86]. The AIC is used to select the better of the two models. Suppose, the true distribution is F and there are two models G1 and G2, the best model can be selected with lower Kullback-Leibler divergence distance from F. Moreover, AIC has two important properties: consistency and a loss function [86]. For assessment of regression function in square error type loss, AIC is considered as the ideal method for parametric and non-parametric situations [58, 86]. The AIC can be derived from following formula:

\[ AIC_{c1} = \left\{ n \log(\hat{\sigma}^2) + n \frac{\delta_2(n + v_1)}{\hat{\sigma}^2 - 2} \right\} \] (6.1)

\( \delta_1 = \text{Trace}((I - L)^T(I - L)) \)

\( \delta_2 = \text{Trace}(((I - L)^T(I - L))^2) \)

\( v_1 = \text{equivalent number of parameters} = \text{Trace}(L^T L), \)

if \( \hat{y} \) is the fitted value then \( v_1 = \sum_{i=1}^{n} \frac{\text{variance}(\hat{y}_i)}{\sigma^2} \)

\( L = \text{hat matrix or smoothing matrix and can be calculated using} = X(X^T X)^{-1}X^T \)

\( I = \text{Identity matrix} \)
\( \hat{\sigma}^2 \) = error mean square

\( n \) = number of parameters

AIC is derived from a model's likelihood function and resulting maximum likelihood estimate. An important feature of AIC is that it adds bias correction as penalty and removes common term to all models. This then makes sum of the squared bias and the estimation errors of the same order as criterion value. This leads to compare the sum of squared bias and the estimation error over the models. This compromise between squared bias and estimation error produces minimax optimal rate.

6.1 Estimation of Features using LO(W)ESS fit: The Minimum Maximum Property

Considering a problem of data smoothing using a LO(W)ESS fit where the selection of smoothing parameter is done using AIC criteria. The AIC helps in estimating the parsimonious model among a set of models using maximum likelihood principle. Yang [86] proved that AIC holds the minimum-maximum (minimax) property. A mini-max property can be defined as an absolute minimum (or maximum) at \( x = c \) provided \( f(c) \) is the smallest (or largest) value that the function will ever take on a domain as shown in Figure 6.1.

In this chapter a similar behaviour of mini-max property of LO(W)ESS fit is observed using the experimental examples in the earlier chapters.

In order to assess the performance of LO(W)ESS as a smoothed time series which
represents reality, we propose visualising the results in a transformed, dimensional less control chart.

The control chart limits are decided using outlier resistant summary statistics. The mid-line, center line (CL), is itself the LO(W)ESS line and is calculated using the median of LOWESS fit residuals. The x-axis units are the sampling interval delta T. The y-axis units are $\pm 1\sigma$, $\pm 2\sigma$, and $\pm 3\sigma$. In the control chart, the UCL, UWL, LWL and LCL is computed using the following method:

CL = median of LO(W)ESS residual

UCL(3$\sigma$) = median of LO(W)ESS residual + 2.66* (median absolute deviation of LO(W)ESS residuals)

LCL(3$\sigma$) = median of LO(W)ESS residual - 2.66* (median absolute deviation of LO(W)ESS residuals)

UWL(2$\sigma$) = median of LO(W)ESS residual + 1.77* (median absolute deviation of LO(W)ESS residuals)
LWL(2\sigma) = \text{median of LO(W)ESS residual-1.77}^* \text{ (median absolute deviation of LO(W)ESS residuals)}

The advantage of using median and median absolute deviation is that it is an outlier resistant method and the LOWESS fit is unaffected by the presence of outliers.

From all the experimental examples in the earlier chapters, the following properties are observed:

1. In this coordinate system, all LO(W)ESS lines, (CL), are of the same length.

   (a) $\alpha = 0.25$
   
   (b) $\alpha = 0.5$
   
   (c) $\alpha = 0.75$
   
   (d) $\alpha = 1$

   Figure 6.2: Minimax property: Different $\alpha$ values

2. The random selection of $\alpha$ arises the problem of over-smoothing or under-smoothing of data and the good LO(W)ESS fit provides the accurate infor-
mation about the data. Similar behaviour has been observed i.e. the best value of $\alpha$ follows the mini-max property as shown in Figure 1.11 with $\alpha=0.25$, in Chapter 1. As the $\alpha$ is varied, the data points move up or down (or remain static) and does not essentially follow the mini-max property. Figure 6.2 shows an example of the variations of minimax properties with varying value of $\alpha$. Here, $\alpha=0.25$ follows the mini-max property whereas $\alpha=0.5$, $\alpha=0.75$ and $\alpha=1$ do not really follow the mini-max property.

3. If the $\alpha$ value is optimum, we are at or near, minimum-maximum (minimax), where the positive error and neighbouring negative error values are very nearly equal and opposite in sign as shown in Figure 6.3. The linear dataset, used in Chapter 3 Section 3.2.2 Experiment 1, is used in Figure 6.3. Similarly, for Figure 6.4 the synthetically generated linear dataset is used with mean 0. Also, a noise component is added to the dataset.

Unlike the simulation for approximation by continuous functions, we cannot have a minimax approximation, shown in Figure 6.4. This problem renders a theoretical minimax theorem (equal in magnitude, opposite in sign) virtually impossible.

However, we can see from a fourth observation.

4. Taking Figure 6.6 into consideration, reducing the error in panel $i$, increases the error in panel $i−1$ and $i+1$. This is the dataset used in Chapter 4 Section 4.3.2 to show the situation where change point masked the trend. Figure 6.5 shows a situation of change point masking the trend and the number of change points.
WBS detected 6 change points at location 28, 66, 91, 116, 155, and 177 as shown in Figure 6.5(b).
Figure 6.5: Change point masking trend and locations of change points

Figure 6.6: Example: Minimax property with equal and opposite sign

Figure 6.6(a) shows the original dataset for change point masking trend and Figure 6.6(b) shows the LO(W)ESS analysis of change point masking trend. In Figure 6.7, the dataset is segmented based on the locations of change points. Figure 6.7(a), (b), (c), (d), (e), (f), and (g) shows the LO(W)ESS analysis based on the change point location, before and after the change points. For example,
Figure 6.7(a) shows the LO(W)ESS analysis before the first change point i.e. data subset from 1 to 28, Figure 6.7(b) shows the LO(W)ESS analysis after change point location 28 and before change point location 66 and so on.

![Figure 6.7: Example: Minimax property with equal and opposite sign](image)

This leads to a conjecture.

Conjecture 1: the optimal AIC fit for a LO(W)ESS curve exhibits a minimax error curve.

The next question:

What situations can affect the minimax error in the LOWESS curve?

We provide one example of such behaviour where the introduction of (possibly spurious) change-points alters the AIC within segments of the curve, thereby restoring the minimax property.
6.2 Relationship between Change point detection, LO(W)ESS and Control Chart Patterns

In Section 6.1, the behaviour of LO(W)ESS fit is observed using control chart. As described earlier, a time series can have a combination of different features which cannot be detected by one method. Keeping the properties of LO(W)ESS fit in Section 6.1, to identify all the hidden characteristics of a time series, it is proposed to identify and classify different possible features, and decision making process using the similar characteristics.

The purpose of showing the relationship between the analytical techniques is to analyse a time series component with specified sequential order to identify different characteristics of data. It is extremely important to use the analytical techniques in specified sequential order otherwise it may lead to erroneous results. As we have seen in Chapter 4 some situations where either change point is masking trend or trend is masking change point, it is important to identify if the time series is behaving abnormally and having some hidden characteristics such as change point. Once change point is detected, the time series is smoothed before and after the change point. Finally, those chunks are used to identify statistical characteristics of the time series.

The proposed method is capable of detecting the important features by three steps. It is important to use analytical techniques in the specified sequential order. Algorithm 5 shows the order in which the analytical techniques can be used.

Change point detection helps to detect if the properties of a process change
Algorithm 5 Relationship between analytical techniques

1: procedure Change point, LO(W)ESS and control chart

2: Change point detection

3: Data Smoothing

4: Pattern Detection

over time. Data smoothing eliminates “noise” and extracts real trend in the process. Finally, pattern detection helps to find different patterns in a time series.

Figure 6.8: Pattern detection using change point, LO(W)ESS, and control chart pattern

For demonstration, a sample of 120 data points is considered as an example. As described in Chapter 3, experimental results show the importance of detection of abrupt changes before performing data smoothing. Firstly, detection of change point is performed. Based on the data points, it shows that the change point occurs at
location 60, as shown in Figure 6.8. Next, the data smoothing is performed using LO(W)ESS before and after the change point. Finally, in order to detect the hidden patterns, the characteristics of different patterns are analysed. The LOWESS fit, before the change point, is close to the centre line and signifies a stratification pattern, described in Chapter 3. According to the characteristics of stratification pattern, the data points hug the centre line. Similarly, the LOWESS fit, after the change point, shows large variations and the points are above and below the centre line. This clearly signifies that the pattern is systematic. Similar behaviour is observed in the first chunk, before the change point, and it can be concluded that the pattern before the change point is stratification whereas the pattern after the change point is systematic.

Using the Monte-Carlo simulations in 100 repeated experiments, it was found that 99% of cases show accurate results if analysis is done using the steps explained in Algorithm 5.

### 6.3 Interpretation and Summary

Entities must not be multiplied beyond necessity.

*Incorrectly attributed to William of Ockham*

LO(W)ESS is widely used in explaining behaviour in many situations, both simple and very complex. Without care as to what LO(W)ESS is replacing, erroneous conclusions are possible (in fact, likely).
It is tempting to look for a theoretical framework for a proof of the behaviour of the LO(W)ESS curve which is chosen by AIC to be “optimal”.

The numerical explanation is easy to understand, and leads to a realistic view of the meaning of the LO(W)ESS interpolation.

The example in Figure 6.2 has a noise component with mean zero, and the period of fluctuations is a single step. This minimax property has exactly the expected behaviour as seen in Figure 6.2(a) as opposed to the remaining three Figure 6.2(b), (c), (d) where the minimax property is not observed.

In Figure 6.3 the frequency of oscillations is not quite as clear-cut as compared with the “window size”, but the minimax property is still holding quite well.

In Figure 6.4, the minimax property is holding with some difficulty, even though the alpha parameter is being optimized.

In Figure 6.5, the curve between 100 and 200 is being “stressed” by an event with a long period compared to the sampling interval. The minimax property is still approximately holding but the total time domain must be broken up into separate components.

The tricubic weight function is very “flat”. The following Table 6.1 illustrates the fraction of points between the mid-point and the extreme edge of the interval [-1,1] for the weight to fall below 0.99, 0.95, 0.50.
Table 6.1: Percentage of points around the centre with weight function close to unity

<table>
<thead>
<tr>
<th>Point</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>15%</td>
</tr>
<tr>
<td>0.95</td>
<td>25%</td>
</tr>
<tr>
<td>0.10</td>
<td>32%</td>
</tr>
<tr>
<td>0.50</td>
<td>55%</td>
</tr>
</tbody>
</table>

Provided enough points fall inside this threshold the weight function can be arbitrarily close to unity and therefore, the regression for the mid-point is numerically close to the unit weight function.

Treating the sequence of linear regression as operating in a sliding window, we can expect as the regression line to remain very similar between step $i$ and step $i+1$. Since, we are restricting the regression line to be linear (LOWESS) or quadratic (LOESS), then the LO(W)ESS value in the $(i+1)^{st}$ step will not have changed significantly from the $i^{th}$. Then either

1. We will have an error of opposite sign or

2. We will experience a step trend or step upwards or downwards.

In our simulations, we have added a noise component with mean 0. Therefore it is reasonable to expect to see some form of the minimax property. If, however, we are dealing with censored data eg. detection limits the LOWESS line using the AIC criteria will still assume the noise as mean 0. This is likely to be masked by a shift in
the mean of the LOWESS line away from the censored data. For example, a cluster of data points at the limit of detection and limit of quantification would push the LOWESS line away from that cluster unless remedial steps are taken which recognize the detection limit.
Chapter 7

Summary and Future Work

The main research question and the main objective of this Thesis has been stated as, “What is the best way to use combination of different modelling techniques for identifying overlapping features in a time series?”

This objective has achieved by answering the specific research questions. Each question has its own importance based on scientific and academic point of view. The specific questions are as follows:

1. Which method is both computationally efficient and accurate in identifying correct numbers and locations of change points?

2. Is there really a change point or a trend?

3. Has trend masked the change point or has change point masked the trend?

4. How does one identify mixture patterns based on the statistical characteristics?

5. How can the smoothing parameters be selected, based on statistical properties of a dataset?

6. What is the best order to use analytical techniques to identify overlapping features?
The structure that has been followed so far is summarized as: establishment of research objective and motivation (Chapter 1); review of existing time series analysis techniques such as LOWESS, control chart patterns, and change points and trend (Chapter 2). The specific research questions relating to the Thesis have been addressed to achieve the main objective (Chapter 4, Chapter 5, and Chapter 3). These three chapters focus on the issues addressed in the existing techniques and attempts to fill the gaps and provide a set of solutions for improvements.

7.1 Overall Conclusions

The key accomplishments of this Thesis have been:

1. A statement for analysis of time series in the presence of multiple overlapping features

2. Informative inter-comparisons between the modern change point detection methods

3. Differentiation and identification of abrupt and gradual changes. This is relevant when either change point masks the actual characteristics of trend or trend masks the actual characteristics of change point

4. Identification and classification of basic mixture patterns from a time series using statistical properties of data. The patterns that can be identified are normal, shift up, shift down, trend up, trend down, stratification, systematic, and cyclic
5. Auto-tuning smoothing parameters of LO(W)ESS, data smoothing technique, based on statistical properties of data.

6. A framework to establish links between the three techniques, control chart patterns, change point analysis, and LO(W)ESS, to identify the overlapping features in a time series for decision making process and predictions.

7. An agenda for future work need to cope with persisting and emerging problems

7.1.1 Change point detection and Trend analysis

The research presented in Chapter 4 was motivated by the importance of being able to detect accurate numbers and locations of change points. The study was designed in order to provide an insight into the computational efficiency and accuracy of four modern techniques of change point detection: wild binary segmentation, iterative robust detection, Bayesian change point analysis and the E-agglomerative algorithm. The experimental results from this research indicate that wild binary segmentation potentially outperforms. In the other techniques a shorter time, it is able to detect most of the correct locations and numbers of change points in different simulations. Through these experiments, it is also concluded that none of the chosen methods is currently able to detect change in the variance of data.

This chapter also presents the statistical issues involved when both abrupt and gradual changes are present. The study provides an insight into situations when either change point covers the trend or trend covers the change point.

Thus, this Thesis may be regarded as a valuable resource to analyze a time series
with overlapping features. This will potentially enhance the prediction and decision making process.

### 7.1.2 Control Chart Patterns

The research presented in Chapter 5 was motivated by the importance of analysing water quality to avoid the possible short term and long term impacts on human being, aquatic, and wild life, and providing the required statistical measures for decision makers.

Analysing hydrological time series is essential for the prediction of water level, where statistical methods quantify the performance of data through analysis. Control chart detects special causes in the dataset by analysing the data on control limits such as upper control limit, lower control limit and center line. It predicts various types of patterns such as patterns that go along the center line, lies within the $1\sigma$, $2\sigma$, and $3\sigma$ limits, going away from center line either in positive or negative directions.

Water quality time series data exhibits combinations of different patterns. Conversely, most of the existing methods focus on the classification of single pattern only which in turn gives erroneous results if mixture patterns exist.

In this study, CCPs mixture recognition is done using a combination of RobustICA and decision trees. Potentially, four statistical features are presented which provides an opportunity for understanding the behaviour of unusual patterns in detail.

The proposed method includes two main modules:
1. Independent pattern identification from mixture patterns using RobustICA

2. Classification of patterns using decision tree induction rules

The pattern identification process does not require any efforts from user. It is truly automatic. In the experimental results indicate, the proposed scheme has achieved more than 98% of accuracy. Hence, it is promising for recognition of combination of unnatural patterns. This scheme can be applied to any time series applications such as financial, medical and many more.

7.1.3 LOWESS

The research presented in Chapter 3 was motivated by the wide application areas of LO(W)ESS fit. LO(W)ESS is a novel non-parametric regression technique to discover complex relationships between variables and explore bivariate and multivariate data.

Despite the wide application area it is unfortunate that the selection of smoothing parameter is done on trial and error. Over-smoothing and under-smoothing is neither acceptable nor desirable in such situations. Over-smoothing divulges trend but ignores local variations whereas under-smoothing results in too many local variations.

In this research a methodology is proposed to auto-tune smoothing parameters. The selection of degree of polynomial is done using weighted MSE and the type of the degree of polynomial and presence of outliers is used to select the type of LO(W)ESS, LOWESS, LOESS, rLOWESS and rLOESS. The AIC-mm estimator is used for the selection of best value of smoothing parameter.
An experimental study is done using simulated data and real data. The proposed model is able to auto-tune the smoothing parameter and degree of polynomial. At the same time, it eliminated the problem of over-smoothing and under-smoothing of fit and decides the best fit based on the statistical characteristics of data.

### 7.2 Future Work

There are several possible extensions to identify overlapping features discussed within this Thesis. The nature of possible future work involves not only direct extension of analysis of the problem presented, but also could involve additional statistical challenges.

One possible extension would include the analysis of limit of detection and limit of quantitation using control charts. This may help to discover relationship between control chart and limit of detection and limit of quantitation.

The proposed methodology within this Thesis helps to identify overlapping features in the data that are underpinned by statistical characteristics. The consequence of analysis using proposed methodology is that identification and classification of statistical characteristics will be improved reliability and accuracy.

Thesis develops improved version of existing time series analysis methods and explore the relationship between them to identify hidden and concurrent patterns for possibly accurate analysis and provides a platform for future work.
Bibliography


[76] D. Rothschild. As Gingrich’s fate rises, so does Obama’s, January 2012. [Online; posted 24-January-2012].


Appendix A

Data Description

In this appendix we present Data Description that we omitted in the body of the thesis.

A.1 Detail Data Description

We will use data from Honey Creek, Maumee, Rock creek and Sandusky for analysis. Details of these stations along with parameters and their measuring units are given in Table A.1. The details of a dataset are given below:

1. Datetime: Shows data and time sample was collected. The format is mm/dd/yy hh:mm
2. Days since 10/01/74: This is converted into Julian date i.e. 741001
3. Sample time window: This is calculated as difference of the time interval between following and preceding samples multiplied by 0.5.
4. Flow: This is measured in cubic feet per seconds. Data is converted into U.S.G.S rating table
5. SS: Suspended Solids concentration measured in mg/L
6. TP: Total Phosphorus concentration measured in mg/L (as P)
7. SRP: Soluble Reactive Phosphorus concentration measured in mg/L
8. No23: Nitrate plus nitrite concentration measured in mg/L
9. TKN: Total Kjeldahl Nitrogen concentration measured in mg/L
10. CL: Chloride concentration measured in mg/L

11. Sulfate: Concentration measured in mg/L

12. Silica: Concentration is measured in mg/L


<table>
<thead>
<tr>
<th>Parameters</th>
<th>Honey Creek</th>
<th>Maumee</th>
<th>Rock Creek</th>
<th>Sandusky</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Data</td>
<td>19,084</td>
<td>16,911</td>
<td>16,545</td>
<td>17,764</td>
</tr>
<tr>
<td>Flow, CFS</td>
<td>19,065</td>
<td>16,898</td>
<td>16,342</td>
<td>17,751</td>
</tr>
<tr>
<td>SS, mg/L</td>
<td>18,976</td>
<td>16,725</td>
<td>16,438</td>
<td>17,606</td>
</tr>
<tr>
<td>TP, mg/L</td>
<td>19,010</td>
<td>16,806</td>
<td>16,467</td>
<td>17,651</td>
</tr>
<tr>
<td>SRP, mg/L</td>
<td>18,961</td>
<td>16,771</td>
<td>16,484</td>
<td>17,603</td>
</tr>
<tr>
<td>$NO_{23}$, mg/L</td>
<td>19,010</td>
<td>16,799</td>
<td>16,478</td>
<td>17,629</td>
</tr>
<tr>
<td>TKN, mg/L</td>
<td>16,420</td>
<td>15,336</td>
<td>16,361</td>
<td>14,814</td>
</tr>
<tr>
<td>Chloride, mg/L</td>
<td>18,314</td>
<td>16,759</td>
<td>16,454</td>
<td>16,947</td>
</tr>
<tr>
<td>Sulfate, mg/L</td>
<td>6,830</td>
<td>6,479</td>
<td>6,703</td>
<td>6,362</td>
</tr>
<tr>
<td>Silica, mg/L</td>
<td>6,879</td>
<td>6,469</td>
<td>6,754</td>
<td>6,397</td>
</tr>
<tr>
<td>Conductivity, µmho</td>
<td>6,868</td>
<td>6,475</td>
<td>6,760</td>
<td>6,412</td>
</tr>
</tbody>
</table>
Appendix B

Parameters for Patterns

B.1 Parameters for Control Chart Patterns

Table B.1 shows different parameters and pattern equations to generate synthetic data.

Table B.1: Parameters for Control Chart Patterns

<table>
<thead>
<tr>
<th>Pattern type</th>
<th>Parameters /Values</th>
<th>Pattern Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal (NOR)</td>
<td>Mean ((\mu) = 80)</td>
<td>(y_i = \mu + r_i \sigma)</td>
</tr>
<tr>
<td></td>
<td>Standard deviation ((\sigma) = 5)</td>
<td></td>
</tr>
<tr>
<td>Stratification (STA)</td>
<td>Random noise((\sigma') = 0.2\sigma \text{ to } 0.4\sigma)</td>
<td>(y_i = \mu + r_i \sigma')</td>
</tr>
<tr>
<td>Systematic (SYS)</td>
<td>Systematic Departure((d) = 1\sigma \text{ to } 3\sigma)</td>
<td>(y_i = \mu + r_i \sigma + d \ast (-1)^i)</td>
</tr>
<tr>
<td>Cyclic (CYC)</td>
<td>Amplitude ((a)= 1.5\sigma \text{ to } 2.5\sigma)</td>
<td>(y_i = \mu + r_i \sigma + a \sin\left(\frac{2\pi i}{T}\right))</td>
</tr>
<tr>
<td></td>
<td>Period ((T) = 8 \text{ and } 16)</td>
<td></td>
</tr>
<tr>
<td>Increasing Trend (UT)</td>
<td>Gradient ((g)=0.05(\sigma) \text{ to } 0.1(\sigma))</td>
<td>(y_i = \mu + r_i \sigma + ig)</td>
</tr>
<tr>
<td>Decreasing Trend (DT)</td>
<td>Gradient ((g) = 0.05(\sigma) \text{ to } 0.1(\sigma))</td>
<td>(y_i = \mu + r_i \sigma - ig)</td>
</tr>
<tr>
<td>Upward Shift (US)</td>
<td>Shift magnitude((s)=1.5(\sigma) \text{ to } 2.5(\sigma))</td>
<td>(y_i = \mu + r_i \sigma + ks)</td>
</tr>
<tr>
<td></td>
<td>Shift position ((P) = 9, 17, 25)</td>
<td></td>
</tr>
<tr>
<td>Downward Shift (DS)</td>
<td>Shift magnitude((s)=1.5(\sigma) \text{ to } 2.5(\sigma))</td>
<td>(y_i = \mu + r_i \sigma - ks)</td>
</tr>
<tr>
<td></td>
<td>Shift position ((P) = 9, 17, 25)</td>
<td></td>
</tr>
</tbody>
</table>
Table B.2: Parameters for Change Point

<table>
<thead>
<tr>
<th>Pattern type</th>
<th>Pattern Equation</th>
<th>Parameters Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Step</td>
<td>$\mu_1 = \mu_0 + \delta \cdot \frac{\sigma_0}{\sqrt{n}}$</td>
<td>$\delta = 0.5, 1.1.5, 2, 2.5, 3$</td>
</tr>
<tr>
<td>Linear Trend</td>
<td>$\mu_1 = \mu_0 + \beta(i - \tau) \cdot \frac{\sigma_0}{\sqrt{n}}$</td>
<td>$\beta = 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$</td>
</tr>
<tr>
<td>Systematic</td>
<td>$\mu_1 = \mu_0 + l \cdot (-1)^i \cdot \sigma_0$</td>
<td>$l = 0.5, 1, 1.5, 2, 2.5, 3$</td>
</tr>
<tr>
<td>Cyclic</td>
<td>$\mu_1 = \mu_0 + l \cdot \sin(2\pi \frac{i}{8}) \cdot \sigma_0$</td>
<td>$l = 0.5, 1, 1.5, 2, 2.5, 3$</td>
</tr>
<tr>
<td>Mixture</td>
<td>$\mu_1 = \mu_0 + l \cdot (-1)^h \cdot \sigma_0$</td>
<td>$l = 0.5, 1, 1.5, 2, 2.5, 3$</td>
</tr>
</tbody>
</table>

h is randomly (0 or 1)

B.2 Parameters for Change Point

$\mu_1 =$ new mean, $\mu_0 =$ old mean, $\sigma_0 =$ old standard deviation, 
n= number of observations, l= cyclic amplitude, $\beta =$ slope, t= sampling time