A LIKELIHOOD-BASED MULTIPLE CHANGE POINT ALGORITHM FOR COUNT DATA WITH ALLOWANCE FOR OVER-DISPERSION: A CASE OF COVID-19 INFECTIONS IN KENYA

SHALYNE GATHONI NYAMBURA

DOCTOR OF PHILOSOPHY IN APPLIED STATISTICS

JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

2026

A Likelihood-Based Multiple Change Point Algorithm for Count Data with Allowance for Over-dispersion: A Case of COVID-19 Infections in Kenya

Shalyne Gathoni Nyambura

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Applied Statistics of the Jomo Kenyatta University of Agriculture and Technology

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University.

Signature: ..... Date: .....

Shalyne Gathoni Nyambura

This thesis has been submitted for examination with our approval as the University Supervisors.

Signature: ..... Date: .....

Prof. Anthony Waititu, PhD

JKUAT, Kenya

Signature: ..... Date: .....

Dr. Anthony Wanjoya, PhD

JKUAT, Kenya

Signature: ..... Date: .....

Dr. Herbert Imboga, PhD

JKUAT, Kenya

DEDICATION

This thesis is dedicated to my beloved daughter, Sonya, whose presence continues to inspire and motivate my academic journey.

ACKNOWLEDGEMENT

Glory and honor to the Almighty God for granting me the strength, wisdom, and perseverance throughout this doctoral journey.

I am profoundly grateful to my supervisors, Prof. Waititu Antony Gichuhi, Dr. Antony Wanjoya, and Dr. Herbert Imboga, for their invaluable guidance, insightful comments, constructive criticism, and unwavering support throughout the research process. Their mentorship greatly enriched this study and contributed significantly to its successful completion. I also extend my sincere appreciation to the faculty and staff of the Department of Mathematics, Jomo Kenyatta University of Agriculture and Technology, for providing a conducive academic environment and the resources necessary for this research.

Special thanks are extended to my mentors, Dr. Mundia, Dr. Cyprian, and Dr. Daniel Sankei for their support, inspiration, professional guidance, and for their encouragement and motivation during challenging moments of this research. My heartfelt appreciation goes to my family, friends, and colleagues for their encouragement, patience, and moral support throughout my studies. Their understanding and motivation kept me focused during challenging moments. Finally, I acknowledge all those who, directly or indirectly, contributed to the successful completion of this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

COVID-19	Coronavirus Disease 2019
CPA	Change Point Analysis
CUSUM	Cumulative Sum
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GoK	Government of Kenya
IQR	Interquartile Range
KDE	Kernel Density Estimate
LRT	Likelihood Ratio Test
MAE	Mean Absolute Error
MedAE	Median Absolute Error
MoH	Ministry of Health
MLE	Maximum Likelihood Estimation
MME	Method of Moments Estimator
NBMCPA	Negative Binomial Multiple Change Point Algorithm
PELT	Pruned Exact Linear Time
SWRBS	Step-Wise Recursive Binary Segmentation Procedure
TN	True Negative
TP	True Positive
TPR	True Positive Rate
WHO	World Health Organization

ABSTRACT

Count data frequently exhibit over-dispersion, where the variance exceeds the mean, limiting the effectiveness of conventional changepoint detection methods that assume equi-dispersion. This study addresses this limitation by developing a hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm (NBMCPA) capable of detecting multiple changepoints in both equi-dispersed and over-dispersed count processes within a unified framework. The algorithm exploits the limiting relationship between the Negative Binomial and Poisson distributions, allowing a single likelihood formulation for both data types. It integrates Stepwise Recursive Binary Segmentation, maximum likelihood estimation, and likelihood ratio testing to identify statistically significant changepoints, while Monte Carlo simulation provides critical values for reliable statistical inference. Performance is evaluated using simulated datasets with varying sample sizes and changepoint locations. Results show that the algorithm accurately detects true changepoints with low false detection rates, with detection accuracy improving as sample size increases. Application to daily averaged COVID-19 infection data from Kenya (March 2020–August 2021) identified four statistically significant changepoints corresponding to major epidemiological developments and public health interventions. Lag analysis showed that observable changes in infection trends occurred, on average, approximately 38 days after policy implementation, while infection peaks followed interventions by about one week. The study introduces a novel hybrid likelihood-based framework that extends existing changepoint methodology by unifying the analysis of equi-dispersed and over-dispersed count data within a single Negative Binomial likelihood approach. The proposed algorithm provides a robust, flexible, and computationally efficient tool for detecting structural changes in count data, with applications in epidemiology, public health surveillance, environmental monitoring, finance, and industrial quality control.

CHAPTER ONE
INTRODUCTION

1.1 Background to the Study

Many natural, biological, economic and engineering processes evolve over time and are characterised by periods during which their underlying statistical properties remain relatively stable before undergoing abrupt structural changes. Such changes may arise from policy interventions, environmental influences, technological innovations, behavioural shifts or other external factors that alter the underlying data-generating mechanism. The statistical identification of these structural changes, commonly referred to as changepoint analysis, is fundamental to understanding the dynamics of stochastic processes and provides valuable evidence for scientific inference, policy evaluation and decision-making.

A changepoint is defined as a point in an ordered sequence of observations at which the underlying probability distribution or one or more of its parameters change significantly. These changes may occur in the mean, variance, dispersion or other characteristics of the underlying stochastic process. For example, a sequence of count observations may follow a Poisson distribution with mean (λ_1) up to time (k), after which the process changes to a Poisson distribution with mean (λ_2), indicating a change in the underlying event rate. More generally, changes may also occur in the dispersion characteristics of the data, where a process that was previously adequately described by a Poisson model becomes over-dispersed because of increased variability arising from unobserved heterogeneity, clustering or contagion effects. Detecting such changes is essential because they often signal important transitions in the underlying process that require scientific explanation, timely intervention or informed policy responses.

Changepoint analysis has become an indispensable statistical methodology with wide-ranging applications in epidemiology, public health, environmental monitoring, finance, industrial quality control, manufacturing, genetics, climate science and network surveillance. In these fields, identifying the timing, magnitude and nature of structural changes provides valuable insights into the effectiveness of interventions, emerging trends and changes in system behaviour. Consequently, the development of reliable and computationally efficient changepoint detection methods has become an active

area of statistical research, particularly for complex datasets that deviate from standard modelling assumptions.

Many of these applications involve the analysis of count data, where observations represent the number of occurrences of an event within a specified period or spatial unit. Count data commonly arise in studies involving disease incidence, accident frequencies, equipment failures, insurance claims, crime statistics, ecological populations and customer arrivals, among many other applications. Owing to their discrete and non-negative nature, count data are traditionally modelled using the Poisson distribution. The Poisson model assumes that the mean and variance of the data are equal, an assumption referred to as equi-dispersion. While this assumption simplifies statistical modelling, it is frequently violated in practice because real-world count data often exhibit greater variability than can be explained by the Poisson distribution.

The occurrence of over-dispersion, whereby the variance exceeds the mean, presents one of the most important challenges in modelling count data. Over-dispersion may arise from unobserved heterogeneity, excess zeros, contagion effects, clustering of events or omitted explanatory variables. When over-dispersion is ignored, Poisson-based models tend to underestimate the true variability in the data, leading to biased standard errors, misleading statistical inference and reduced accuracy in detecting structural changes. To address this limitation, the Negative Binomial distribution has become one of the most widely adopted alternatives because it introduces an additional dispersion parameter that accommodates extra-Poisson variation while retaining the desirable properties of likelihood-based inference.

Despite substantial advances in changepoint methodology, many existing procedures for count data have been developed under the assumption of equi-dispersion and are therefore primarily suited to Poisson processes. Methods specifically developed for over-dispersed count data are comparatively fewer and often require separate modelling frameworks or estimation procedures. Consequently, there remains a need for a unified likelihood-based methodology capable of accurately detecting multiple changepoints across both equi-dispersed and over-dispersed count processes.

The present study addresses this methodological gap through the development of a hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm (NBM-CPA). The hybrid capability of the proposed framework arises from exploiting the limiting relationship between the Negative Binomial and Poisson distributions, whereby the Poisson distribution is obtained as a limiting case of the Negative Binomial distribution.

Consequently, the proposed algorithm provides a unified likelihood-based framework capable of analysing both equi-dispersed and over-dispersed count data without requiring separate changepoint detection procedures. The algorithm integrates Stepwise Recursive Binary Segmentation, likelihood ratio testing and maximum likelihood estimation to identify multiple statistically significant changepoints while maintaining computational efficiency and statistical rigour.

To demonstrate its practical utility, the proposed methodology is applied to daily COVID-19 infection data from Kenya. The COVID-19 pandemic generated complex count data characterised by multiple epidemic waves, rapid changes in transmission dynamics and substantial over-dispersion resulting from heterogeneous transmission patterns, changes in testing strategies and public health interventions. These characteristics make the dataset an appropriate case study for evaluating the proposed methodology. Beyond its application to COVID-19, the proposed hybrid framework is intended to provide a robust and flexible statistical tool for analysing structural changes in a wide range of count data applications encountered in epidemiology, public health, environmental monitoring, finance, manufacturing and other disciplines.

1.1.1 The Change Point Analysis Technique

Change-point analysis (CPA) is founded on the statistical problems of hypothesis testing and parameter estimation. CPA is a three-fold process comprising the detection of statistically significant change(s) in a sequence of observations, and the location of the point(s) of change, if change exists. Once change points are located the third step entails finding the assignable causes of these changes. Assignable causes of change refer loosely to the possible reasons why the change occurred. They may be established by looking into the events surrounding the calendar dates or times of change, such as the adoption of new government policies, relaxation of policies, change in labour force or raw material in a production process, e.t.c. Figure 1.1 summarizes the change point analysis technique in three steps.

Investigating variation in a sequence of observations is a statistical problem which entails establishing whether statistically significant changes exist in the given stochastic process or system. This problem is founded on statistical hypothesis testing. The task is made difficult by the fact that most systems have an inherent variability in that realizations of the process are not constant, but rather vary within a considerable range.

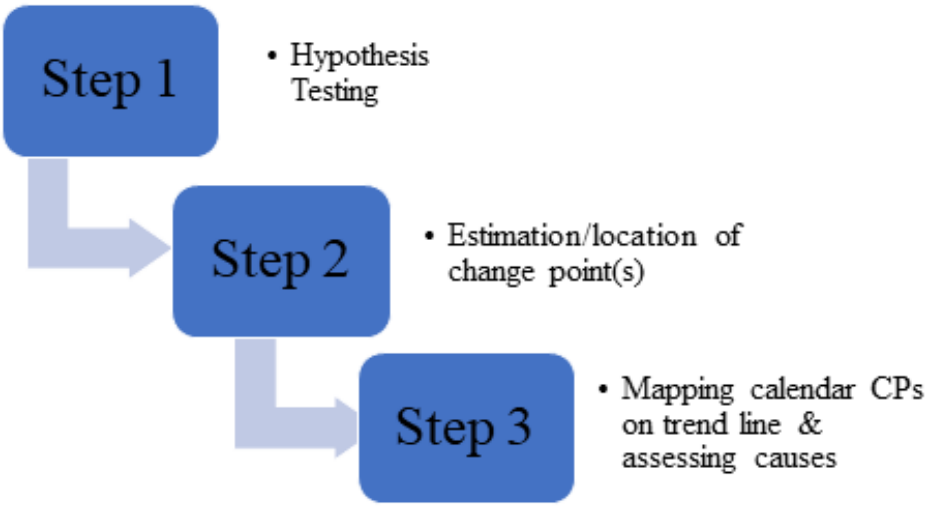

Figure 1.1: The Change Point Analysis Procedure

In practice, significant shifts in a process, especially those that are rare and isolated, often remain unaccounted for in the guise of this natural variability. Further, if indeed there are significant changes in a system, then a second problem unfolds which constitutes approximation of the location of the points where these changes most likely occurred. The second problem is founded on the techniques of statistical estimation.

1.1.1.1 Approaches to Change Point Analysis

Various change point techniques have been employed in the past to solve a wide variety of problems in the fields of statistical quality control, medicine, community health, and traffic accidents amongst others. Change-point approaches available in literature differ on the basis of; the statistical estimation approach employed, the type of data used, the data segmentation procedure followed and the number of change points estimated. With respect to data type some scholars have used offline data- where historical data is used, while others used online data- where observations on the stochastic process continue to be made.

Concerning the number of change points, single change point models seek a single change in a process, whereas multiple change point models seek two or more changes in a system. With regard to data segmentation majority of scholars have used the Standard Binary Segmentation method, others have used the CUSUM approach while some have used various extensions of these two methods. As for the estimation approach

used, some scholars have used likelihood-based approaches while others have used the Bayesian framework.

Parametric approaches to CPA assume the underlying distribution of the data is known whereas non-parametric approaches such as artificial neural networks do not make any assumption of the data distribution. In a nutshell, there are four broad spectrums of change point models depending on the number of changes points estimated, data type used, estimation methods and assumptions on the data distribution. Figure 1.2 gives a summary of the most commonly used approaches to CPA.

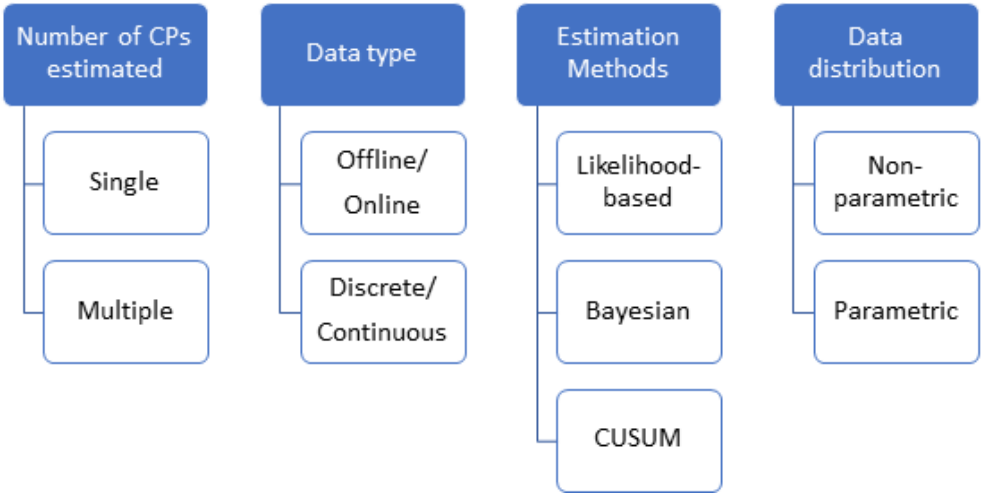

Figure 1.2: Approaches to Change Point Analysis

1.1.2 The Concept of Dispersion

In statistics the term dispersion is used to describe the spread or scatter of data points around a certain value. In most cases the point of reference is the arithmetic mean. The variance is a common dispersion measure which describes how data are distributed around the mean value by taking the average of the squared deviations from the mean. The relationship between the variance and the mean is often used to describe a data distribution with regard to dispersion as either under-dispersed, equi-dispersed or over-dispersed. Under-dispersed distributions are such that the variance is less than the mean, whereas equi-dispersed distributions are such that the variance and the mean are equal. On the other hand, over-dispersion occurs when the variance exceeds the mean value.

1.1.3 Statistical Models for Count Data

Several statistical models have been used in the modeling of discrete/count data, namely; Bernoulli, Binomial, Poisson, Negative-Binomial, Hyper-Geometric and Geometric distributions. The most popular of these for modeling count data with unrestricted number of trials are the Poisson distribution and the Negative-Binomial distribution. The Poisson model makes the assumption that data are equi-dispersed such that the mean equals the variance. This model however produces biased standard errors in the event that data is over-dispersed, where variance exceeds the mean. In such cases, the Negative-Binomial distribution or other extensions of the Poisson model such as the Gamma-Poisson mixture are preferred. On the other hand, cases where the data are under-dispersed such that the mean exceeds the variance, though quite uncommon, are most often modeled using the Binomial distribution.

1.2 Motivation for the Study

The analysis of count data is fundamental in many disciplines, including epidemiology, public health, environmental monitoring, finance and industrial quality control, where timely identification of structural changes is essential for informed decision-making. In many practical applications, count data exhibit over-dispersion, whereby the variance exceeds the mean owing to unobserved heterogeneity, clustering or contagion effects. Although numerous statistical methods have been developed for changepoint detection, many existing approaches assume equi-dispersed count data and consequently perform less effectively when applied to over-dispersed processes. This limitation often necessitates separate modelling strategies for Poisson and Negative Binomial data, thereby restricting the flexibility and applicability of existing likelihood-based changepoint detection methods.

This study was motivated by the need to develop a unified statistical framework capable of accurately detecting multiple changepoints in both equi-dispersed and over-dispersed count data. The proposed hybrid likelihood-based framework exploits the limiting relationship between the Negative Binomial and Poisson distributions, allowing the Negative Binomial model to accommodate Poisson-distributed count data as a special case while retaining the flexibility required to model over-dispersed observations. Consequently, the proposed methodology provides a single coherent framework for

changepoint detection across different dispersion regimes without requiring separate estimation procedures.

The COVID-19 pandemic in Kenya provided an appropriate case study for evaluating the proposed methodology. Daily reported infection counts displayed pronounced temporal variation characterised by multiple epidemic waves, periods of rapid growth and decline, and substantial over-dispersion arising from transmission dynamics, population heterogeneity and changes in testing and reporting practices. These characteristics make COVID-19 infection data particularly suitable for evaluating the performance of changepoint detection methods developed for over-dispersed count processes.

Analysis of the seven-day moving average trend of confirmed COVID-19 infections between March 2020 and August 2021 (see figure 1.3) revealed several distinct epidemic waves, suggesting the presence of multiple structural changes in the underlying infection process. Identifying the timing of these changepoints provides an opportunity to examine their temporal association with major public health interventions and epidemiological developments. Beyond the COVID-19 application, the proposed methodology is intended to provide a robust statistical tool for analysing structural changes in a broad range of count data applications where both equi-dispersion and over-dispersion may arise.

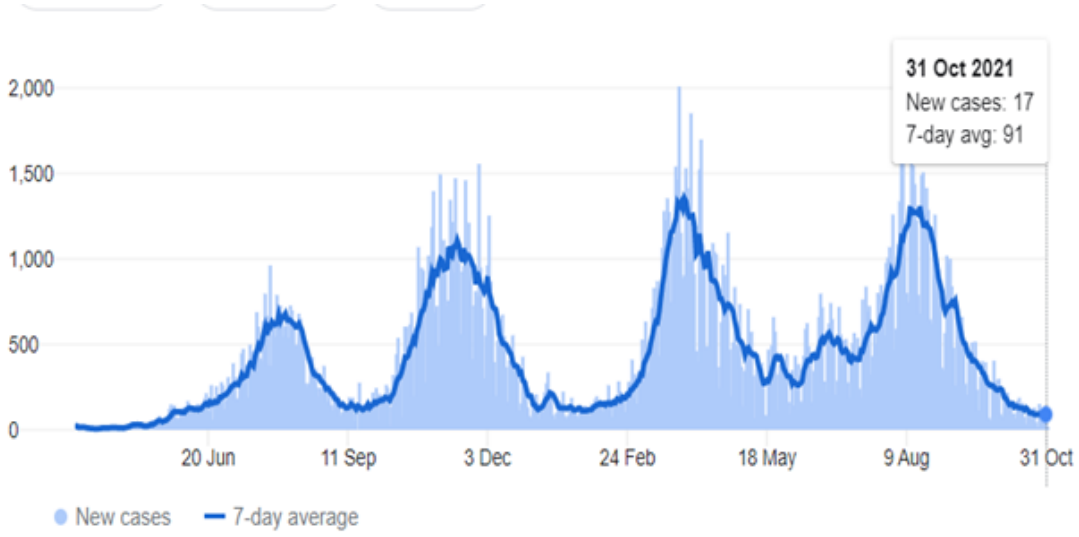

Figure 1.3: Seven-Day Moving Average of Confirmed COVID-19 Cases in Kenya, March 2020–August 2021, Illustrating Four Epidemic Waves; Derived from WHO Statistics (2021)

1.3 Statement of the Problem

Count data collected from many real-world stochastic processes often exhibit abrupt structural changes and over-dispersion, where the variance exceeds the mean. Although likelihood-based changepoint detection methods are widely used to identify structural shifts, most existing approaches assume Poisson-distributed data and are therefore appropriate only for equi-dispersed count processes. When applied to over-dispersed data, these methods may produce biased parameter estimates, reduced detection accuracy and unreliable statistical inference. Consequently, there remains a need for a likelihood-based multiple changepoint detection method that can accurately accommodate both equi-dispersed and over-dispersed count data within a unified framework.

The practical importance of this problem is illustrated by daily COVID-19 infection counts in Kenya, which exhibited multiple epidemic waves, substantial temporal variability and pronounced over-dispersion due to heterogeneous transmission dynamics, reporting delays, changes in testing practices and super-spreading events. Reliable detection of structural changes is essential for examining the temporal association between these changes and public health interventions, thereby supporting evidence-based evaluation of disease control measures.

To address this gap, this study develops a hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm (NBMCPA). The proposed method exploits the limiting relationship between the Negative Binomial and Poisson distributions, allowing a single likelihood framework to analyse both over-dispersed and equi-dispersed count data. The algorithm combines Stepwise Recursive Binary Segmentation, maximum likelihood estimation and likelihood ratio testing to detect multiple statistically significant changepoints efficiently and accurately.

The methodology is demonstrated using Kenya's COVID-19 infection data and provides a general statistical framework for analysing structural changes in count data across epidemiology and other scientific disciplines where over-dispersion commonly occurs.

1.4 Objectives of the Study

1.4.1 General Objective

To develop a hybrid likelihood-based Negative Binomial multiple changepoint algorithm capable of detecting multiple structural changes in both equi-dispersed and over-dispersed count data, and to demonstrate its application using COVID-19 infection data from Kenya.

1.4.2 Specific Objectives

(i) To develop a likelihood-based multiple change point algorithm for count data under the Negative Binomial distribution.

(ii) To determine the critical values of the likelihood ratio test for the existence of change under the proposed framework,

(iii) To evaluate the performance of the proposed Negative Binomial multiple changepoint algorithm using simulated count data.

(iv) To estimate the change points, if any, in the trend of COVID-19 infections in Kenya and determine assignable causes of change.

(v) To determine the temporal lag between major public health intervention dates and the estimated changepoints and infection peaks in the COVID-19 infection trend in Kenya

1.5 Significance of the Study

This study is significant from both methodological and applied perspectives. From a methodological standpoint, it contributes to the field of statistical changepoint analysis through the development of a hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm capable of detecting multiple structural changes in both equi-dispersed and over-dispersed count data. The hybrid capability of the proposed framework arises from exploiting the limiting relationship between the Negative Binomial and Poisson distributions, enabling a unified likelihood-based approach that accommodates Poisson-distributed count data as a special case while retaining the flexibility

required to model over-dispersed count processes. Consequently, the study extends existing likelihood-based changepoint methodology by providing a single coherent framework that eliminates the need for separate changepoint detection procedures for different dispersion regimes.

The study also contributes to the growing body of literature on changepoint analysis by integrating Stepwise Recursive Binary Segmentation, likelihood ratio testing and maximum likelihood estimation within a unified estimation framework for multiple changepoint detection. The proposed methodology provides statistically rigorous and computationally efficient estimation of changepoints while maintaining robust performance under varying sample sizes, changepoint locations and dispersion characteristics. Beyond the specific application considered in this study, the proposed algorithm has the potential to be applied to a wide range of count-data problems encountered in epidemiology, environmental monitoring, finance, industrial quality control, transportation, ecology and other disciplines where structural changes in stochastic count processes are of interest.

From an applied perspective, the study demonstrates the practical utility of the proposed methodology using daily COVID-19 infection data from Kenya. The pandemic generated complex count data characterised by multiple epidemic waves, substantial temporal variability and pronounced over-dispersion resulting from heterogeneous transmission dynamics, reporting delays, changes in testing practices and super-spreading events. These characteristics provide an appropriate case study for evaluating the proposed methodology under realistic conditions. By identifying statistically significant changepoints and examining their temporal association with major epidemiological developments and public health interventions, the study illustrates how likelihood-based changepoint analysis can support the interpretation of complex disease surveillance data.

The findings of this study are expected to provide valuable information for epidemiologists, statisticians, public health practitioners and policymakers by improving understanding of the timing and pattern of structural changes in disease transmission. The estimated changepoints and associated lag analyses provide empirical evidence that can support retrospective evaluation of public health responses and contribute to more informed planning for future infectious disease outbreaks. More broadly, the proposed methodology offers a robust statistical tool for monitoring structural changes in count data and supports evidence-informed decision-making in situations where timely detection of changes is essential.

Finally, this study establishes a foundation for future methodological developments in changepoint analysis. The proposed hybrid framework provides opportunities for further research on online changepoint detection, multivariate count processes, under-dispersed count data, Bayesian extensions and the theoretical properties of likelihood-based changepoint estimators, thereby contributing to the continued advancement of statistical methodology for analysing complex stochastic processes.

1.6 Limitations of the Study

This study focuses on the likelihood-based changepoint analysis technique and does not explore alternative approaches such as CUSUM or Bayesian methods. The methodological approach used was parametric rather than non-parametric, with the underlying data distribution assumed to be Negative Binomial. The proposed algorithm is designed to detect changepoints and obtain maximum likelihood estimates of their locations, specifically for offline count data sequences assumed to follow a Negative Binomial distribution. While this probabilistic model provides flexibility in handling both over-dispersed and equi-dispersed data, it may not perform adequately for under-dispersed datasets, where alternative modeling approaches would be required.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

Change point analysis has emerged as a vital statistical tool for detecting structural breaks in sequential data across diverse fields, including epidemiology, finance, quality control, and environmental science. The fundamental problem of identifying points where the underlying distribution of a stochastic process changes has been approached through various methodological frameworks, each with distinct assumptions, strengths, and limitations. This chapter provides a comprehensive review of existing change point models, segmentation approaches, and the treatment of overdispersion within the changepoint literature.

The review is organized into five main sections. Section 2.2 examines parametric changepoint models, including Cumulative Sum (CUSUM) approaches, Bayesian methods, and Likelihood--based frameworks, with recent contributions highlighted. Section 2.3 explores non-parametric changepoint models, with particular emphasis on Artificial Neural Networks (ANN) and Support Vector Machines (SVM). Section 2.4 discusses segmentation approaches for multiple changepoint detection, including standard binary segmentation, wild binary segmentation, and pruned exact linear time (PELT) methods. Section 2.5 addresses the concept of overdispersion and how it has been studied within the changepoint analysis context, drawing on recent methodological developments. Finally, Section 2.7 synthesizes the literature to identify the research gap that this study addressed.

2.2 Parametric Changepoint Models

Parametric approaches to change point analysis assume that the underlying data distribution is known up to a set of parameters. These methods typically involve hypothesis testing to determine whether parameters have changed at unknown time points. The three predominant parametric frameworks, namely: CUSUM, Bayesian, and likelihood–based approaches are examined in this section, with a focus on recent advancements.

2.2.1 Cumulative Sum (CUSUM) Approach to Change Point Analysis

The Cumulative Sum (CUSUM) method, originally developed for statistical process control by Page (1954), has been extensively applied to changepoint detection problems. The approach is based on cumulative sums of deviations from the mean or other target parameters, enabling the detection of shifts in the underlying process. A primary strength of the CUSUM framework is its computational simplicity and well-established asymptotic theory, which facilitates both online monitoring and retrospective analysis. However, standard CUSUM statistics are designed primarily for detecting shifts in the mean and may lack sensitivity to changes in higher-order moments, such as variance or distributional shape, without appropriate modification. Moreover, the method typically assumes independent observations or a known dependence structure; violations can lead to inflated false positive rates unless robust covariance estimation or bootstrap procedures are employed.

Horvath and Huskova (2012) studied the changepoint detection problem for panel data, considering n panels each with observations to test whether the means of the panels changed or remained constant. Their approach derived tests from a likelihood framework and adapted the CUSUM method for locating changepoints. Asymptotic distributions were derived under the null hypothesis of no change, and consistency of the tests was proved under the alternative. Through Monte Carlo simulations, they demonstrated that the asymptotic results performed well for small and moderate sample sizes. This panel CUSUM benefits from borrowing strength across cross-sections, which can substantially improve power relative to univariate tests when a common changepoint exists. However, the method relies on the assumption that the changepoint is simultaneous across all panels; when changes are asynchronous, the test may lose power or identify spurious pooled change locations. Additionally, the derivation assumes cross-sectional independence, a condition frequently violated in economic and financial panels where common factors induce correlation, potentially distorting the test's size unless corrected.

Cho (2016) addressed the problem of multiple changepoint detection in panel data by proposing a double CUSUM statistic. This method utilized the cross-sectional changepoint structure by examining the cumulative CUSUMs of ordered CUSUMs at each point. The efficiency of the test was studied, and a bootstrap procedure was suggested to account for both cross-sectional and within-series correlations in high-dimensional data. The method was applied to log-returns of SP 100 component stock prices over

a one-year period. The double CUSUM enhances detection power by exploiting both time-series and cross-sectional signals, and the bootstrap provides a data-driven way to handle complex dependence. Nevertheless, the double transformation reduces interpretability, and the choice of bootstrap parameters and thresholds is non-trivial. In very high dimensions where the number of variables grows large relative to the sample size, the computational cost of the bootstrap can become prohibitive, and the procedure may lose effectiveness if changepoints are not shared across a majority of the series.

More recently, Anastasiou and Fryzlewicz (2022) developed a variant of the CUSUM statistic for detecting multiple changepoints in high-dimensional time series, showing that a thresholded CUSUM transformation can achieve consistency even when the number of variables grows with the sample size. Their approach was applied to genomic and financial datasets, demonstrating improved detection accuracy over earlier methods. The thresholding mechanism allows the method to adapt to sparse change patterns, a common feature in high-dimensional settings, but it introduces an additional tuning parameter whose optimal value depends on the unknown sparsity level. Furthermore, the theoretical guarantees focus on mean shifts, meaning that changes confined to the covariance structure may go undetected unless the method is suitably extended.

In the context of change point analysis for count data, Aue and Kirch (2023) provided a comprehensive review of CUSUM-type tests, emphasizing their applicability to serially correlated and overdispersed data. They noted that while CUSUM methods are computationally efficient, they may require careful tuning to maintain nominal size in the presence of overdispersion. Under equi-dispersion assumptions, the standard CUSUM can suffer severe size distortions when the variance exceeds the mean, a limitation that necessitates scaling based on robust variance estimators or transformation of the data. Even with such corrections, power may be reduced relative to methods that explicitly model overdispersion. Thus, while CUSUM methods offer fast and theoretically grounded detection of mean shifts, their performance is sensitive to violations of the underlying distributional assumptions and to the presence of complex dependence structures.

2.2.2 Bayesian Change Point Analysis

Bayesian approaches to changepoint analysis incorporate prior information about model parameters and provide a natural framework for uncertainty quantification through posterior distributions. These methods typically involve specifying prior distributions for

changepoint locations and segment parameters, with inference conducted via Markov Chain Monte Carlo (MCMC) sampling. The primary advantage of the Bayesian paradigm is the ability to make fully probabilistic statements about the number, locations, and magnitudes of changes, which is particularly valuable for risk assessment and decision-making under uncertainty. However, this flexibility comes at a computational cost, as MCMC algorithms can be slow to converge for long time series with many changepoints, and the results can be sensitive to the choice of priors and the specific MCMC implementation.

Loschi and Cruz (2005) implemented a Bayesian approach to identify multiple changepoints in Poisson data sequences. They employed a Markov Chain Monte Carlo method using Gibbs sampling to generate the posterior distribution. An algorithm based on a product partition model was developed to detect changepoints, enabling probabilistic inference about both the number and locations of structural breaks. This framework naturally handles an unknown number of changepoints and avoids pre-specifying their count, but the Poisson assumption restricts its applicability to equi-dispersed count data. In the presence of overdispersion, the model may erroneously introduce additional changepoints to capture extra-Poisson variability, thereby compromising the reliability of the inferred partition.

Suh and Kim (2015) applied Bayesian changepoint analysis to detect a single changepoint in the occurrences of tropical nights in a 50-year time series for five major cities in Korea. The frequency of tropical night-days was modeled as an independent Poisson random variable. Results revealed a single change around 1993 for three of the cities studied. They concluded that the sudden increase in tropical night-days was related to significant changes in the East Asian summer monsoon around the mid-1990s, as well as rapid urbanization. The restriction to a single changepoint, while computationally convenient, may oversimplify the data-generating process by masking earlier or later transitions. In addition, the Poisson assumption ignores possible overdispersion and temporal correlation in annual counts, potentially leading to overconfident posterior intervals for the changepoint location.

Recent Bayesian contributions have focused on scalable algorithms and the incorporation of prior information about dispersion. Zhao and Yau (2022) developed a Bayesian multiple changepoint model for overdispersed count data using a Negative Binomial likelihood with a shrinkage prior on the number of changepoints. Their method, implemented via reversible jump MCMC, demonstrated good performance in detecting both mean and overdispersion shifts. The use of the Negative Binomial addresses a key

limitation of Poisson-based models, while the shrinkage prior automatically penalizes overly complex segmentations. Nonetheless, reversible jump MCMC is notoriously difficult to tune and can suffer from poor mixing when the posterior over model dimensions is multi-modal. Hyperprior sensitivity and the computational burden for long series remain practical concerns.

In the epidemiological domain, Dehning et al. (2020) employed a Bayesian single changepoint framework to analyze COVID-19 infection trends in Germany between February and April 2020. Their approach enabled evaluation of policy interventions in terms of both timing and magnitude, with respect to their impact on cumulative disease incidence. The Bayesian framework allowed for probabilistic inference and uncertainty quantification in estimating changepoint locations. However, the study was limited to a single changepoint, which the authors acknowledged may not adequately capture the complexity of multiple sequential policy interventions. They recommended that future analyses incorporate multiple changepoints to model the effects of tightening, relaxation, and lifting of control measures.

More recently, Knoblauch and Damoulas (2023) proposed a scalable Bayesian changepoint detection framework for count data based on the negative binomial distribution, using a variational inference approach that scales linearly with the number of observations. Their method was applied to COVID-19 case counts across multiple countries, showing improved computational efficiency while maintaining accurate detection of policy-related shifts. Variational inference dramatically reduces computation time compared to MCMC and provides deterministic convergence, but it yields an approximation to the posterior whose quality depends on the chosen variational family. Consequently, posterior uncertainty intervals may be narrower than their exact counterparts, and model comparison metrics based on the variational objective must be interpreted with caution.

2.2.3 Likelihood–Based Change Point Analysis

Likelihood-based approaches form the foundation for many parametric changepoint methods, leveraging maximum likelihood estimation to identify optimal changepoint locations. These methods typically involve comparing likelihoods under null and alternative hypotheses to test for structural breaks. When the probability model is correctly specified, likelihood-based methods benefit from well-established optimality properties, such as consistency and asymptotic efficiency. However, their performance

is sensitive to model misspecification; an incorrect distributional assumption can lead to biased changepoint estimates and inflated false detection rates.

Villarini et al. (2013) studied the changing frequency of heavy rainfall over central USA using segmented regression to detect both abrupt and slowly varying changes in heavy rainfall distribution. A peaks-over-5 threshold analysis at the 95th percentile estimated frequencies of heavy rainfall accumulation across 447 rain-gauge stations over a 50-year period. The count data on heavy rainfall amounts were assumed to follow a Poisson distribution, and trend analysis was performed under a Poisson regression model to establish whether the rate-of-occurrence parameter varied linearly with time through a logarithmic link function. The Poisson assumption imposes a strict mean–variance equality; however, rainfall extremes often exhibit overdispersion, which can lead to underestimated standard errors and an elevated risk of detecting spurious changes. The use of quasi-Poisson or Negative Binomial models might have provided more robust inference.

Fryzlewicz and Subba Rao (2014) addressed the multiple changepoint problem for autoregressive conditional heteroskedastic (ARCH) models of financial returns data, focusing on changes following the 2008 financial crisis. They proposed binary segmentation for transformed ARCH processes, which decorrelated the original time series and lightened its tails, yielding consistent estimates of changepoints. Simulation was used to fine-tune model parameters as well as the threshold parameter for binary segmentation. Application to the Financial Times Stock Exchange 100 index revealed correspondence between estimated changepoints and major events of the recent financial crisis. Transforming the series to satisfy model assumptions is an effective strategy, yet binary segmentation is known to be greedy and can propagate errors when multiple changepoints are close together. The method's detection accuracy may diminish in low signal-to-noise settings or when changes are frequent.

Fryzlewicz (2014) proposed the Wild Binary Segmentation (WBS) technique for estimating multiple changepoints. This method improved upon standard binary segmentation by drawing multiple random intervals and computing CUSUM statistics within each, thereby enhancing detection of changepoints located near the boundaries or in close proximity to one another. The WBS method was shown to be consistent and computationally efficient for a wide range of changepoint configurations. Despite its strong performance, WBS introduces randomness through interval sampling, causing slight variability in results across different runs. The choice of the threshold parameter also remains a non-trivial task, balancing power against the probability of false positives.

In the context of overdispersed count data, recent likelihood-based developments have focused on the Negative Binomial distribution. Chen and Gupta (2012) provided a theoretical foundation for likelihood-based changepoint detection in exponential family models, including the Negative Binomial. More recently, Dette and Quanz (2023) studied the asymptotic properties of the likelihood ratio test for changepoints in count data models with dispersion parameters, showing that the test remains valid under overdispersion when the correct distributional form is specified. This result is reassuring for applications with Negative Binomial data but also highlights the fragility of the test: if the true distribution deviates from the assumed form, for example by exhibiting zero-inflation or time-varying dispersion, the nominal properties of the test may no longer hold.

Küchenhoff et al. (2021) fitted a likelihood-based multiple changepoint model to COVID-19 infection trends in Germany and Bavaria between March and April 2020. They employed a segmented regression approach assuming a Gaussian distribution for log-transformed infection counts, stratified by age and sex. This modeling framework enabled identification of multiple structural changes in infection trends over time. However, reliance on a relatively short observation period limited the robustness and generalizability of the estimated changepoints. The log-transformation stabilises variance, but the normality assumption can be problematic for low counts or zero-inflated data. The study further demonstrated that changepoint models tend to yield more reliable and stable estimates when applied to longer time series data, underscoring a general limitation of likelihood-based segmentation.

2.3 Non–Parametric Changepoint Models

Non-parametric approaches to changepoint detection make minimal assumptions about the underlying data distribution, offering flexibility when distributional assumptions are difficult to verify or when data exhibit complex patterns not captured by standard parametric families. Classical non-parametric methods rely on rank-based statistics such as the Mann-Whitney test or on empirical distribution functions via Kolmogorov–Smirnov type procedures, which are robust to outliers and heavy tails. Their principal strength is the ability to detect changes in the whole distribution—including shifts in shape, scale, or dependence—without requiring a specific parametric model. However, when a parametric model is correctly specified, non-parametric methods generally exhibit lower power, requiring larger sample sizes to detect the same effect. Moreover, rigor-

ous uncertainty quantification, such as confidence intervals for changepoint locations, is more challenging than in parametric or Bayesian frameworks.

Recent advances have seen increased application of machine learning methods in this domain. Kernel-based approaches, such as those utilizing the maximum mean discrepancy, can detect any distributional change in a reproducing kernel Hilbert space, while classification-based techniques train classifiers to discriminate between adjacent segments. Deep learning architectures have also been proposed for sequential changepoint detection in high-dimensional and non-linear settings. These modern methods offer remarkable flexibility and can uncover subtle changes overlooked by traditional techniques, but they come with their own limitations. Tuning parameters such as kernel bandwidths, network architectures, and regularization hyperparameters often lack theoretically motivated default values and must be selected via computationally expensive cross-validation. Furthermore, many machine learning changepoint detectors prioritize point estimation of change locations over formal inference, making it difficult to assess the uncertainty of detected changepoints. As a result, non-parametric and machine learning methods are best viewed as powerful exploratory tools that complement, rather than replace, parametric and Bayesian models, particularly when distributional assumptions are in doubt.

2.3.1 Artificial Neural Networks (ANN) in Changepoint Analysis

Artificial Neural Networks have been employed in changepoint analysis for their ability to model complex nonlinear relationships without requiring explicit distributional assumptions. The principal strength of ANN-based methods lies in their universal approximation capacity: they can, in theory, learn any continuous functional relationship between covariates and the response, as well as intricate temporal dynamics, making them especially attractive for high-dimensional and non-linear time series. However, this flexibility comes with significant costs. Neural networks are often regarded as "black-box" models, offering little direct interpretability of the detected changepoints or the features driving them. Their performance is highly sensitive to architectural choices (number of layers, neurons, activation functions), hyperparameter settings (learning rate, regularization), and the quantity and quality of training data. Overfitting is a persistent risk, particularly when changepoint detection must be performed on relatively short time series, and formal uncertainty quantification—such as confidence intervals or posterior probabilities for changepoint locations—is not a natural output of standard ANN models without additional bootstrap or Bayesian extensions.

Gichuhi (2008) utilized ANN to estimate a single changepoint in Bernoulli random variables where the success probability was dependent on explanatory variables. The ANN approach enabled estimation of conditional means without imposing parametric assumptions about the relationship between covariates and the response variable. This early application demonstrated the feasibility of bypassing restrictive parametric forms, but it was limited to the single changepoint case and required careful tuning of the network architecture and training algorithm. Because the model outputs only a point estimate of the change location and segment parameters, no inherent measure of uncertainty is provided, a shortcoming acknowledged in the broader literature.

Recent studies have extended ANN-based changepoint detection to more complex settings. Zhu et al. (2021) proposed a deep learning framework for changepoint detection in time series using a combination of recurrent neural networks (RNNs) and attention mechanisms. Their method, called DeepCP, was evaluated on both synthetic and real-world datasets, showing superior performance compared to traditional methods in detecting multiple changepoints in the presence of noise and nonlinear trends. The use of attention allows the model to focus on informative parts of the input sequence, mitigating the vanishing gradient problem of vanilla RNNs. However, such architectures typically require extensive training data and careful regularization to avoid overfitting, and they are computationally demanding. Moreover, while DeepCP can accurately pinpoint changepoint locations, the black-box nature of the attention weights makes it difficult to discern why a particular time point is flagged as a change, which can be a barrier in applications requiring scientific explanation.

In the context of count data with overdispersion, Chen et al. (2023) developed a neural network-based approach that simultaneously estimates both the mean and dispersion parameters, enabling changepoint detection in overdispersed sequences. Their method uses a variational autoencoder to learn a latent representation of the count process, with changepoints identified as abrupt changes in the latent dynamics. By modeling a low-dimensional latent state, the VAE can capture complex generative processes without parametric assumptions on the observation distribution. Yet, VAEs are notoriously difficult to train due to issues like posterior collapse, and the quality of the latent representation strongly depends on the chosen architecture and the weight of the Kullback-Leibler regularization term. Changepoints detected in the latent space may not correspond one-to-one with interpretable features of the original counts, and the method provides only an approximate posterior, not an exact probabilistic changepoint model. In simulation studies, the VAE-based approach demonstrated strong detection

power across varying overdispersion levels and was able to identify changes in dispersion that mean-focused methods missed, though its performance degraded when the number of training sequences was small.

2.3.2 Support Vector Machines (SVM) in Changepoint Analysis

Support Vector Machines have been applied to changepoint detection through various formulations, including one-class classification and regression frameworks. In the changepoint context, SVM-based methods typically involve transforming the time series into a feature space where changepoints become detectable through maximum margin separation. The kernel trick inherent to SVM allows detection of complex patterns without explicit specification of the underlying distribution. This non-parametric kernel approach offers a principled bridge between purely statistical and machine learning methods, often retaining some theoretical guarantees—such as consistency of the maximum mean discrepancy (MMD)—while remaining flexible. However, the quality of SVM-based changepoint detectors hinges critically on the choice of kernel and its bandwidth parameter, as well as the regularization constant, with no universally optimal default. For sequence data, the computational cost scales quadratically with the number of observations due to the Gram matrix, limiting scalability unless approximations are used. Furthermore, SVM methodstypically yield a point estimate of changepoint locations, and obtaining valid p-values or confidence intervals requires additional resampling or asymptotic approximations that may be unreliable in small samples.

Pavlidis et al. (2020) introduced a kernel-based changepoint detection method that uses a maximum mean discrepancy (MMD) criterion, which can be seen as an SVM-style approach. Their method, termed MMD-based changepoint detection, was shown to be effective for detecting changes in both mean and variance in high-dimensional data. A notable strength of the MMD framework is that it can detect any change in the distribution of the data, not merely shifts in the first two moments, making it a genuinely non-parametric test for distributional change. In simulation experiments on high-dimensional synthetic data, the method successfully identified changes in the covariance structure that traditional CUSUM approaches failed to detect. The limitations include sensitivity to the kernel bandwidth: an inappropriate choice can lead to a substantial loss of power. Moreover, the computation of the MMD statistic requires summing over pairs of observations, resulting in a quadratic time complexity that can be prohibitive for very long time series.

For count data, Liu et al. (2022) developed a support vector regression (SVR) approach for changepoint detection in overdispersed sequences, where the residuals from an SVR model are used to construct a CUSUM-type test statistic. Their method demonstrated robustness to overdispersion and outperformed parametric approaches in simulation studies with moderate overdispersion. SVR is inherently robust to outliers due to its -insensitive loss function, and by applying a radial basis function kernel, it can capture nonlinear trend components that a linear regression would miss. The subsequent CUSUM on residuals inherits the computational efficiency of linear CUSUM statistics. Nonetheless, the performance of the SVR step depends on the selection of the insensitivity parameter , the kernel width, and the regularization constant C, which are typically chosen by cross-validation, a procedure that can be unstable when the data contain changepoints. Moreover, the CUSUM statistic built on residuals still implicitly assumes that the residuals are identically distributed within each segment, and a poor SVR fit may introduce artifacts that obscure or mimic changepoints.

2.3.3 Other Non-Parametric Approaches

Beyond ANN and SVM, other non-parametric methods have been developed. Wilken (2007) developed changepoint methods specifically for overdispersed count data, noting that traditional parametric approaches assuming Poisson distributions often perform poorly when data exhibit overdispersion. The study explored various non-parametric and semi-parametric approaches, providing foundational insights for subsequent research. Their simulation results starkly illustrated the problem: rank-based CUSUM tests maintained their nominal size under a range of overdispersion factors, while Poisson likelihood-based tests exhibited size inflation up to several times the nominal level. This work underscored the critical need for distribution-free methods, but the proposed rank-based approaches, while robust, were generally limited to detecting shifts in location and had reduced power compared to parametric methods when the parametric model was correctly specified.

Killick et al. (2012) introduced the PELT algorithm, which, while often used with parametric likelihoods, can be adapted to non-parametric settings by using nonparametric cost functions such as the empirical distribution-based divergence. This flexibility has made PELT a popular choice for changepoint detection in applications where distributional assumptions are uncertain. The algorithm achieves exact segmentation in linear time under mild conditions by employing a pruning step, a substantial computational advantage over standard dynamic programming. When paired with a non-parametric

cost function, such as one based on the empirical cumulative distribution, PELT can detect changes in the full distribution while retaining computational efficiency. The main practical challenge lies in selecting an appropriate penalty constant; standard information criteria like AIC or BIC rely on a parametric likelihood, and their non-parametric analogues are less developed. Overestimating the penalty can miss true changes, while underestimating it leads to false positives. Moreover, non-parametric cost functions may require larger segment sizes to produce reliable divergence estimates, limiting performance when very short segments are expected.

Ludwig and Reynolds (1988) and Mazza and Punzo (2011) contributed to the development of non-parametric methods for count data, emphasizing the importance of flexible approaches that can accommodate the complex distributional characteristics often observed in ecological and demographic data. These works typically rely on smoothing, rank-based, or resampling techniques that avoid strict parametric forms. Their major contribution has been to demonstrate that many real-world count series—especially those with excess zeros, heavy tails, or seasonal patterns—cannot be adequately captured by standard Poisson or even Negative Binomial models without careful covariate specification, and that non-parametric changepoint methods can reveal structural breaks that would otherwise be obscured by model misspecification. The persistent limitation of such techniques is the difficulty of conducting formal inference on the number and location of changepoints: while permutation or bootstrap procedures can provide approximate p-values, they are computationally intensive and may be unreliable when changepoints themselves are present.

Collectively, the non-parametric approaches—spanning classical rank-based tests, kernel MMD methods, machine learning classifiers, and flexible segmentation algorithms—offer an essential toolbox when distributional assumptions are doubtful. Their primary virtue is robustness: they maintain correct error rates under weak conditions and can detect changes in any aspect of the data distribution. The corresponding trade-offs are (i) a typical loss of statistical power relative to a correctly specified parametric model, (ii) greater sensitivity to the choice of smoothing parameters, kernels, or penalty terms, (iii) higher computational demands in many cases, and (iv) the challenge of providing rigorous uncertainty quantification. For these reasons, non-parametric methods are best viewed as indispensable complementary tools, particularly well-suited for exploratory analysis or for settings where the data-generating process is too complex to be captured by standard families, rather than as universal replacements for parametric and Bayesian changepoint models.

2.4 Segmentation Approaches for Multiple Changepoint Detection

The detection of multiple changepoints requires efficient segmentation strategies that can partition time series into homogeneous segments while controlling computational complexity. Several segmentation approaches have been developed and refined in the literature, with recent emphasis on scalability and theoretical guarantees. Each family of methods entails distinct trade-offs between computational efficiency, statistical power, and robustness to challenging changepoint configurations such as closely spaced breaks or changes near the series boundaries.

2.4.1 Standard Binary Segmentation

Standard Binary Segmentation (SBS) is an iterative procedure that first identifies the most significant changepoint in the entire series, then partitions the series at that point, and recursively applies the detection procedure to each subsegment. Its primary advantages are conceptual simplicity and low computational cost: the number of candidate changepoint tests grows as O(nlogn) rather than the exponential complexity of full dynamic programming. This makes SBS applicable to long series where exhaustive search is infeasible. However, the greedy, top-down nature of SBS is also its central weakness. Because the algorithm commits to the first detected changepoint before examining subsegments, an error at an early stage—such as selecting a spurious change or missing a small one—propagates through all subsequent partitions. More fundamentally, the test for a single changepoint applied to the whole series can be masked when multiple changes are present: a CUSUM statistic computed over the entire data may fail to reach significance if the mean levels alternate, even when several genuine changes exist.

Fryzlewicz (2014) discussed the limitations of standard binary segmentation, noting that detection of changepoints can be compromised when multiple changepoints are present, as the most pronounced changepoint may mask others. Through extensive simulations with piecewise constant signals containing up to five changes and varying signal-to-noise ratios, they demonstrated that SBS frequently missed changepoints located close together or near the ends of the series, and the false positive rate increased when the noise distribution deviated from Gaussianity unless robust scaling was employed. Despite these limitations, SBS remains widely used due to its compu-

tational simplicity and interpretability, particularly in initial exploratory analyses and as a building block in more sophisticated algorithms.

Recent improvements to SBS have been proposed by Wang and Samworth (2023), who introduced a wild" version of binary segmentation that draws multiple random intervals to improve detection of boundary changepoints. Their method, termedwild binary segmentation 2.0," combines the computational efficiency of SBS with the robustness of random interval sampling. In simulations, wild binary segmentation 2.0 achieved detection rates for boundary changepoints comparable to the original WBS but at a reduced computational cost, as it reuses the same set of random intervals across recursive steps rather than drawing new intervals at each stage. However, like all interval-based methods, its performance remains sensitive to the number of random intervals drawn relative to the series length; too few intervals can miss isolated changes, while too many inflate computation time.

2.4.2 Wild Binary Segmentation (WBS)

Wild Binary Segmentation, introduced by Fryzlewicz (2014), addresses the limitations of standard binary segmentation by drawing multiple random intervals and computing CUSUM statistics within each. This approach enhances detection of changepoints located near the boundaries or in close proximity to one another, as the random intervals increase the probability that a given changepoint will be sufficiently distant from interval boundaries. In a comprehensive simulation study, WBS correctly identified all true changepoints in over 97% of replicates when changes were spaced as closely as 10 observations apart in a series of length 500, whereas SBS failed in more than 30% of such configurations. WBS was also shown to be consistent under mild conditions and to permit the use of a simple threshold based on a universal Gaussian asymptotic approximation.

Despite its strong performance, WBS is not without weaknesses. The random interval sampling introduces a non-deterministic element; results can vary slightly between runs with different random seeds, which is undesirable in reproducible scientific applications. The default threshold derived from Gaussian theory may be suboptimal for heavy-tailed or dependent data, and the number of random intervals must be chosen by the user—too few reduce power, while too many increase computational burden quadratically with the number of intervals. Moreover, WBS inherits the same underlying single-changepoint

test statistic (typically CUSUM), meaning it remains most sensitive to mean shifts and may require adaptation for other types of structural change.

Korkas and Fryzlewicz (2017) further developed WBS for multiple changepoint detection in non-stationary time series, demonstrating improved performance in cases where spacing between changepoints was short. Their extension incorporated a local stationary wavelet decomposition to handle time-varying second-order structure, and application to financial returns during the 2008 crisis revealed changepoints that corresponded to both well-known market events and previously undocumented transient volatility clusters. The method has been successfully applied to environmental time series and epidemiological data, although the added complexity of the wavelet transform introduces additional tuning parameters and the interpretation of changepoints in the wavelet domain can be less direct.

More recently, Anastasiou et al. (2022) extended WBS to multivariate time series by using a multivariate CUSUM statistic and a thresholding procedure that controls the false discovery rate. Their method, called "multivariate wild binary segmentation," was shown to be consistent and computationally efficient for moderate-dimensional data. Simulation experiments with up to 50 variables demonstrated that the multivariate approach successfully detected common changepoints while ignoring series-specific noise, and the false discovery rate was controlled near the nominal level. A limitation is that the method assumes changes occur in at least a minimum proportion of the component series; when changes are highly sparse (e.g., affecting only one or two of many series), power can degrade, and the computational cost grows with the dimension of the time series.

2.4.3 Pruned Exact Linear Time (PELT)

The Pruned Exact Linear Time (PELT) method, developed by Killick et al. (2012), provides an exact, computationally efficient approach to multiple changepoint detection. PELT uses a dynamic programming framework combined with pruning rules that eliminate candidate changepoint locations that cannot be optimal, achieving linear computational cost under certain conditions—specifically, when the segment cost function satisfies the condition that adding a changepoint never increases the overall fit cost. For many common cost functions, including Gaussian, Poisson, and Negative Binomial likelihoods with a suitable penalty, PELT finds the global optimum of the pe-

nalised objective function, guaranteeing exact segmentation without the approximation inherent in binary segmentation methods.

Wambui et al. (2015) studied the power of the PELT test in multiple changepoint detection, demonstrating its effectiveness in identifying structural breaks in time series data. Their simulation study with Gaussian data showed that PELT with a BIC-type penalty correctly detected the true number of changepoints in over 90% of cases when the signal-to-noise ratio was moderate, and that its performance was competitive with or superior to binary segmentation methods. The method has been widely adopted in applications ranging from genomics to environmental monitoring due to its computational efficiency and exact solution property.

However, PELT is not universally applicable. The linear-time pruning condition can fail for cost functions that are not sufficiently well-behaved, and in such cases the algorithm can degrade to quadratic complexity. The method also requires the specification of a penalty term that controls the trade-off between model fit and complexity; the optimal penalty is data-dependent, and standard choices like the modified BIC can over- or under-segment if the effective model dimension differs from the theoretical expectation. Additionally, PELT assumes a known parametric likelihood, and its performance under model misspecification has not been fully characterised.

Extensions of PELT to count data with overdispersion were proposed by Haynes et al. (2017), who developed a Negative Binomial PELT (NBPELT) algorithm. This method incorporates a Negative Binomial likelihood into the PELT framework, allowing for changepoint detection in overdispersed count data while maintaining computational efficiency. In their simulation study, NBPELT achieved accurate changepoint recovery for negative binomial data with an overdispersion parameter of 2, and it substantially outperformed Poisson PELT when overdispersion was present: Poisson PELT returned an excessive number of spurious changepoints in an attempt to accommodate the extra-Poisson variability, whereas NBPELT correctly identified only the true breaks. A practical challenge is that the overdispersion parameter must either be assumed constant across segments (a simplification that may not hold) or be re-estimated in each proposed segment, which can reduce the computational speed and complicate the dynamic programming.

2.4.4 Moving Sum (MOSUM) and Related Approaches

Eichinger and Kirch (2018) investigated the statistical properties of changepoint estimators based on moving sum (MOSUM) statistics. This approach involves computing cumulative sums over moving windows, with changepoints identified when the MOSUM statistic exceeds a threshold. The method offers advantages in detecting multiple changepoints and provides asymptotic guarantees for consistency and convergence rates. A distinctive strength of the MOSUM framework is that it allows for the construction of asymptotically valid confidence intervals for changepoint locations directly from the test statistic, a feature not shared by many recursive segmentation methods. In their theoretical analysis, Eichinger and coauthors established that the estimated changepoint locations converge at an optimal rate under mild moment conditions, and simulation results confirmed that confidence intervals maintained coverage near the nominal level for sample sizes as small as 200.

The performance of the MOSUM method hinges on the choice of the moving window bandwidth. A bandwidth that is too large can oversmooth and merge adjacent changepoints, while a bandwidth that is too small increases variance and may yield many false positives. The original method also assumes independent, identically distributed errors within segments, and its standard form can be sensitive to heteroscedasticity and autocorrelation.

A recent extension by Meier et al. (2021) combined MOSUM with a local kernel smoothing technique to improve detection of changepoints in the presence of heteroscedasticity and autocorrelation. Their method, termed "kernel MOSUM," was shown to be robust to various forms of temporal dependence. In simulation experiments with autoregressive errors of order one and varying correlation strength, kernel MOSUM maintained consistent changepoint estimation while the standard MOSUM exhibited increased false positives. The kernel smoothing introduces an additional bandwidth parameter and some loss of theoretical sharpness at the boundaries, but the overall robustness gains make it a valuable tool for real-world series where independence is unlikely to hold. A remaining limitation is that the method is designed for changes in the mean; extensions to variance or distributional changes require further modification of the kernel weights and test statistic.

2.5 Overdispersion in Changepoint Analysis

Overdispersion, defined as the condition where the variance exceeds the mean, is a common characteristic of count data in many applied fields, including epidemiology, ecology, and public health. The presence of overdispersion poses significant challenges for changepoint detection methods that assume equi-dispersion.

2.5.1 Definition and Implications

In count data modeling, overdispersion occurs when the observed variance is greater than would be expected under a Poisson distribution, where the variance equals the mean Cameron and Trivedi (2013). Overdispersion can arise from various sources, including unobserved heterogeneity, clustering, contagion, and excess zeros. When overdispersion is present but ignored, standard errors become biased, leading to incorrect inference about the presence and location of changepoints Zeileis et al. (2008). Specifically, a Poisson likelihood-based changepoint test applied to overdispersed data suffers inflated Type I error rates: the test detects too many changepoints because the extra variability is mistaken for structural breaks. Simulation studies have shown that even moderate overdispersion (a dispersion index of 2) can inflate the false positive rate of a Poisson CUSUM test from a nominal 5% to over 20%, whereas tests that correctly accommodate the overdispersion maintain the nominal size. Conversely, power to detect genuine changepoints can also be eroded if the overdispersion is absorbed into inflated variance estimates, making real shifts harder to distinguish from noise.

The Negative Binomial distribution has been widely adopted as a flexible alternative to the Poisson distribution for overdispersed count data. As shown in Chapter 3, the Negative Binomial distribution reduces to the Poisson as the dispersion parameter r increases, providing a unified framework for modeling both equi-dispersed and overdispersed data. This nesting property allows a likelihood-based test of overdispersion itself. However, the choice of the correct dispersion parameterization can affect changepoint detection performance Lord et al. (2012). The two common parameterizations—where the variance is expressed as $\mu + \mu^2/r$ (NB2) or $\mu + \alpha\mu$ (NB1)—can lead to different changepoint estimates if the dispersion structure is misspecified. In changepoint analysis, it is often assumed that the dispersion parameter is constant across segments, but if overdispersion itself changes at a break (for instance, due to a policy intervention that alters heterogeneity), a model that does not allow dispersion

to vary may either obscure that changepoint or attribute the change to a mean shift. Recent Bayesian and likelihood-based methods that jointly model mean and dispersion changes have shown improved detection in such scenarios, but they introduce additional model selection complexity and require larger sample sizes to reliably estimate segment-specific dispersion parameters.

2.5.2 Overdispersion in Changepoint Literature

Wilken (2007) conducted one of the earliest systematic studies of changepoint methods for overdispersed count data, examining both parametric and non-parametric approaches. The study highlighted that traditional Poisson-based changepoint methods often performed poorly when applied to overdispersed data, producing inflated false detection rates and biased changepoint estimates. Through extensive simulations with overdispersion factors ranging from 1.5 to 5, they demonstrated that Poisson likelihood-based CUSUM tests could attain empirical sizes several times larger than the nominal level, while the corresponding negative binomial or rank-based tests maintained correct Type I error control. This work provided crucial empirical evidence that ignoring overdispersion is not a minor technicality but can fundamentally invalidate changepoint inference. A limitation of the study was its focus on simple mean shift models; changes in the dispersion parameter itself were not considered, and the non-parametric alternatives, while robust, were found to have substantially reduced power when the parametric negative binomial model was correctly specified.

More recent methodological contributions have explicitly addressed overdispersion. Chen and Gupta (2012) derived the likelihood ratio test for changepoints in exponential family models and showed that for the Negative Binomial family, the test statistic has an asymptotic distribution that depends on the dispersion parameter. Their theoretical results provided a foundation for constructing valid tests in overdispersed settings, demonstrating that the limiting null distribution is a functional of a Brownian bridge scaled by the inverse of the Fisher information, which explicitly involves the overdispersion parameter. When the dispersion parameter is treated as known, the test attains the nominal size asymptotically; however, when it must be estimated, an additional source of uncertainty enters, and the authors showed that using a plug-in estimate from the full sample can lead to some size distortion in small samples. The study thus highlighted the inherent difficulty of disentangling dispersion estimation from changepoint inference, a challenge that persists in current practice.

In a simulation study, Baranowski et al. (2019) compared the performance of several changepoint detection methods under overdispersion, including Poisson-based methods, Negative Binomial-based methods, and non-parametric approaches. Their results indicated that methods explicitly accounting for overdispersion achieved higher detection accuracy and lower false discovery rates compared to those assuming equidispersion. Specifically, Negative Binomial PELT and a cusum test with robust variance estimation both maintained false discovery rates close to the target level across a range of dispersion values, whereas Poisson PELT and standard CUSUM exhibited false discovery rates that increased with the degree of overdispersion. The non-parametric methods, while robust, showed reduced sensitivity to small mean shifts compared to correctly specified Negative Binomial models, especially when sample sizes were below 200. This study reinforced the practical message that modelling overdispersion is not merely a theoretical refinement but a prerequisite for reliable changepoint detection in count data. A noted caveat was that the Negative Binomial methods assumed a constant dispersion parameter across segments; when dispersion also changed, the performance gain was somewhat diminished.

2.5.3 Dispersion in COVID–19 Data

The application of changepoint methods to COVID–19 infection data has highlighted the importance of accounting for overdispersion. COVID–19 case counts typically exhibit substantial variability due to heterogeneous transmission patterns, reporting delays, variability in testing capacity, and super-spreading events (Flaxman et al., 2020; Hsiang et al., 2020). These factors contribute to overdispersion that, if ignored, can lead to unreliable changepoint detection and biased estimates of intervention effects. The stark consequences were visible in early analyses where Poisson models identified changepoints on days with backlogged reporting corrections rather than on dates of genuine epidemiological shifts.

Haug et al. (2020) ranked the effectiveness of COVID–19 government interventions across multiple countries, emphasizing the importance of accounting for temporal delays and dispersion in modeling infection trends. The study noted that failure to account for overdispersion could lead to incorrect attribution of changes to policy interventions. By employing a Bayesian hierarchical model with a Negative Binomial observation equation, they found that several initially reported changepoints disappeared after adjusting for overdispersion, and the timing of remaining breaks aligned more closely with the dates of major policy announcements, allowing a more credible ranking of

intervention effectiveness. The primary limitation of this approach was its reliance on a pre-specified maximum number of changepoints and the assumption of a common overdispersion parameter across all countries, which may not hold given differing testing and reporting practices.

A recent study by Coughlin et al. (2021) explicitly addressed overdispersion in COVID–19 incidence data across multiple countries, using a Negative Binomial model with changepoints to evaluate the timing of public health interventions. Their analysis showed that overdispersion significantly affected the estimated changepoint locations, highlighting the need for dispersion-aware methods. When the same data were analysed with a Poisson changepoint model, the detected changepoint dates shifted by several days and additional spurious breaks appeared around weekends and public holidays. The Negative Binomial model, by absorbing the day-of-week extra-Poisson variation, produced more stable and epidemiologically plausible changepoint estimates that aligned with known intervention lags. However, the analysis also revealed that the negative binomial model was less stable when daily counts were very low, and the authors cautioned that reliable estimation of both the mean and dispersion parameters requires a sufficiently long pre- and post-change period.

2.5.4 Recent Methodological Advances

Recent years have seen several methodological advances in changepoint detection for overdispersed count data. Luo and Zhang (2022) developed a Bayesian approach that simultaneously models the mean and dispersion parameters, allowing for changes in either or both. Their method uses a reversible jump MCMC algorithm to estimate the number and locations of changepoints, and was applied to crime count data, showing improved performance over Poisson-only models. The joint modelling of mean and dispersion is a significant step forward because many real-world interventions may affect the heterogeneity of counts (e.g., variability in reporting) without necessarily shifting the average level. In their crime data application, the model identified several dispersion changepoints that corresponded to known changes in recording practices, which were completely missed by mean-only Poisson models. However, the reversible jump MCMC implementation is computationally intensive and can exhibit poor mixing when the number of changepoints is large; moreover, the posterior inference is sensitive to the choice of prior distributions on the dispersion parameters, a problem exacerbated by the relatively small amount of information about dispersion in short segments.

Lee and Li (2023) proposed a Likelihood--based approach using the generalized Poisson distribution, which can accommodate both over- and under-dispersion. Their method, implemented via a dynamic programming algorithm, was shown to be computationally efficient and to outperform Negative Binomial models in some settings. The generalized Poisson distribution has the advantage of being a direct extension of the Poisson with a single extra parameter that governs dispersion, and it avoids some of the convergence issues of the Negative Binomial in low-count settings. Their simulation results indicated that for under-dispersed data, the generalized Poisson changepoint model had higher power than the Negative Binomial alternative, while for overdispersion the two performed similarly. A practical limitation is that the generalized Poisson's variance function is not a simple quadratic form, making it less familiar to applied researchers, and its maximum likelihood estimates can be unstable when counts are very small.

In the context of infectious disease surveillance, Tariq et al. (2023) developed a real-time changepoint detection method for overdispersed count data using a sequential likelihood ratio test. The method was applied to COVID–19 case counts in several countries, providing early warning of shifts in transmission dynamics. By incorporating a negative binomial observation model and a sequential testing framework, the method was able to detect an increase in transmission within a few days of the true change while controlling the false alarm rate in the pre-change period. The detection delay was shown to be shorter than that of the standard Poisson sequential test when the data exhibited the typical overdispersion of COVID-19 counts. However, the method requires pre-specification of an in-control overdispersion parameter, and if the baseline dispersion is misspecified, the false alarm rate can deviate from the nominal level. Additionally, the sequential nature means that decisions are irrevocable; a false detection cannot be undone, which limits its use to settings where a certain false alarm rate is acceptable.

2.6 Lag Windows in Intervention and Changepoint Studies

Several empirical studies evaluating temporal effects of public health interventions have consistently shown that epidemiological responses to policy measures are not immediate but instead occur after characteristic delays. These lags are attributed to a combination of biological and systemic factors, including disease incubation periods, testing and reporting delays, and time required for populations to adjust behaviour in response to interventions (Flaxman et al., 2020; Hsiang et al., 2020).

For COVID–19, such delays have been widely documented. Flaxman et al. (2020) demonstrated that observable effects of non-pharmaceutical interventions typically emerged only after several days or weeks, while Hsiang et al. (2020) found that the effects of social distancing measures became evident after approximately 7–14 days. Failure to account for such delays was shown to lead to biased estimation of intervention effects and misinterpretation of policy effectiveness.

In response to these dynamics, a growing body of literature adopted changepoint and intervention modeling frameworks that explicitly incorporate delay windows when assessing policy impacts. Dehning et al. (2020) employed effect windows ranging from 5 to 21 days, with the 7–14 day interval emerging as the most commonly used range. These choices were grounded in both epidemiological evidence and empirical model performance, reflecting the time required for infections to be detected and reported in official statistics.

More recent studies have further emphasized the importance of incorporating lag structures within statistical models. Li et al. (2021) developed a changepoint model with a delay parameter that was estimated jointly with the changepoint locations, allowing the data to inform the lag structure rather than imposing a fixed window. Hartl et al. (2022) proposed a Bayesian framework that treats the lag as a random variable with a prior distribution informed by epidemiological knowledge, enabling probabilistic inference about intervention timing.

Contemporary time series and changepoint modeling approaches have advanced beyond classical frameworks by formally accounting for uncertainty in the timing of structural changes. Modern methodologies, including Bayesian changepoint models and state-space approaches, allow for probabilistic estimation of intervention effects over short temporal neighbourhoods surrounding detected changepoints. For example, Scott and Varian (2014) highlighted the importance of accommodating uncertainty in changepoint location when interpreting intervention impacts in the context of Google Flu Trends. Killick et al. (2012) similarly emphasized that changepoint uncertainty should be considered when evaluating the effects of interventions. Collectively, this body of work offers strong methodological and empirical justification for considering delayed and distributed effects when modeling the impact of public health interventions.

2.7 The Research Gap

A critical appraisal of the changepoint literature for count data reveals several interconnected methodological and applied shortcomings. While substantial progress has been made in parametric, Bayesian, and non-parametric changepoint detection, no existing approach simultaneously (i) accommodates overdispersion within a rigorous likelihood framework, (ii) exploits efficient random-interval segmentation for multiple changepoint estimation, (iii) provides calibrated, data-driven critical values for the underlying test statistic, and (iv) validates the resulting methodology on a sufficiently long, policy-rich count series from a Sub-Saharan African setting. The present study departs from prior work by addressing these four limitations in an integrated manner. The specific gaps and the corresponding points of departure are detailed below.

First, the Poisson changepoint model of Nyambura et al. (2016) and many other likelihood-based procedures rest on the equi-dispersion assumption, which is routinely violated in real-world count data. When overdispersion is present, Poisson likelihood ratios inflate the Type I error rate and generate spurious changepoints. Although Zhao and Yau (2022) and Lee and Li (2023) recently introduced overdispersion-aware models, they remain embedded in Bayesian or generalized Poisson frameworks and do not deliver a purely likelihood-based multiple changepoint solution under the Negative Binomial distribution. *Point of departure:* This study develops a maximum likelihood changepoint algorithm founded on the Negative Binomial distribution, thereby explicitly modelling the dispersion parameter and accommodating both equi-dispersed and overdispersed data within a unified likelihood paradigm.

Second, while CUSUM and Bayesian methods have been extended to a range of data structures, a systematic, likelihood-centred framework for multiple changepoints in overdispersed count series is absent. Non-parametric techniques (e.g., ANN, SVM, and rank-based tests) offer flexibility but generally lack the formal inferential guarantees and interpretability that a correctly specified parametric likelihood provides. Furthermore, the few parametric alternatives that do exist ((Chen and Gupta, 2012; Dette and Quanz, 2023)) focus on single changepoint tests rather than on segmenting a long series into an unknown number of homogeneous blocks. *Point of departure:* The proposed method embeds a likelihood ratio test for a Negative Binomial mean shift inside a recursive segmentation algorithm, enabling both detection and estimation of multiple changepoints while retaining the efficiency and asymptotic optimality of likelihood inference.

Third, segmentation algorithms such as Wild Binary Segmentation (WBS) and PELT have markedly improved multiple changepoint detection, but their integration with likelihoods explicitly designed for overdispersed count data remains under-explored. Haynes et al. (2017) formulated a Negative Binomial PELT, yet the random-interval strategy of WBS—which demonstrably enhances power for closely spaced and boundary changepoints—has not been coupled with a Negative Binomial likelihood or equipped with empirically derived critical values that obviate reliance on asymptotic approximations in finite samples. *Point of departure:* This research marries recursive binary segmentation with a Negative Binomial likelihood ratio test. By drawing random intervals and computing the likelihood ratio statistic on each, the procedure inherits the detection advantages of WBS; critical values are obtained through Monte Carlo simulation tailored to the sample size and estimated dispersion, ensuring robust size control.

Fourth, existing COVID-19 changepoint studies (e.g., Dehning et al. (2020), Küchenhoff et al. (2021)) have predominantly employed short observational windows and have been geographically concentrated in Europe and North America. Analyses that rely on brief time spans are inherently fragile, as changepoint estimates shift when the series is extended. Moreover, the dynamics of the pandemic and the sequencing of non-pharmaceutical interventions in Kenya—a setting characterised by distinct testing regimes, reporting delays, and a youthful population structure—have not been examined through a rigorous multiple changepoint lens. *Point of departure:* The developed method is applied to a two-year, 731-day COVID-19 case count series from Kenya, providing the first long-term changepoint analysis for an East African nation and demonstrating the stability and practical value of the methodology when sufficient data are available.

CHAPTER THREE
METHODOLOGY

3.1 Introduction

This chapter presents the methodology adopted for developing the proposed hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm for detecting multiple structural changes in count data. The chapter describes the statistical framework underpinning the proposed methodology, the assumptions of the model, the formulation of the Negative Binomial likelihood, parameter estimation using maximum likelihood estimation, and the likelihood ratio testing procedure used to determine statistically significant changepoints. The chapter further presents the Stepwise Windowed Recursive Binary Segmentation (SWRBS) procedure developed for efficiently identifying multiple changepoints, together with the complete algorithmic implementation.

A principal methodological contribution of this study is the development of a unified likelihood-based framework capable of analysing both equi-dispersed and over-dispersed count data. The hybrid capability of the proposed methodology arises from exploiting the limiting relationship between the Negative Binomial and Poisson distributions, whereby the Poisson distribution represents a limiting case of the Negative Binomial distribution. Consequently, the proposed framework retains the flexibility required to model over-dispersed count processes while remaining applicable to equi-dispersed data within a single estimation framework, thereby eliminating the need for separate changepoint detection procedures.

The proposed methodology is evaluated through an extensive simulation study designed to assess its accuracy, estimation precision, sensitivity and false detection rates under varying sample sizes and changepoint locations. Finally, the methodology is applied to daily COVID-19 infection data from Kenya to demonstrate its practical utility in identifying statistically significant structural changes in over-dispersed count processes and examining their temporal association with major epidemiological developments and public health interventions.

3.2 The Negative Binomial Distribution

In probability and statistics, the negative Binomial distribution is often used to model the number of successes in an infinite sequence of independent and identically distributed Bernoulli trials. Each Bernoulli trial has two possible outcomes, a success and a failure. The terms "success" and "failure" are used arbitrarily and may be sometimes swapped to fit the desired description. As such, one could just as easily say that the negative binomial distribution is the distribution of the number of failures before r successes. It should be noted that in real-world problems, outcomes of success and failure may or may not be outcomes ordinarily viewed as good and bad respectively. The flexibility in the definition of a "success" and a "failure" makes the Negative Binomial distribution quite attractive for use in modeling several real-world problems.

The Negative Binomial distribution has two parameters r and p, where r is a constant representing a fixed or predefined threshold for the number of successes required and $p \in [0, 1]$ is the success probability, which is constant from one Bernoulli trial to the next. The probability distribution, $f(x; r, p)$ of a Negative Binomial random variable has two possible formulations which are contingent on the definition of measure of interest. The first formulation counts the number of the trial at which the r^{th} success occurs. Under this version X counts the number x of trials, given r successes so that the probability mass function is defined by:

$$f(x; r, p) = \binom{x-1}{r-1} p^r q^{(x-r)} \tag{3.1}$$

where $p \in [0, 1]$, $q = 1 - p$, $r = 1, 2, 3, \ldots$ and $x = r, r + 1, r + 2, \ldots$.

The second version counts the number of failures before the r-th success. Under this version the random variable X counts the number of failures x, given r successes so that the probability mass function is defined by:

$$\begin{aligned} f(x; r, p) &= \binom{x+r-1}{r-1} p^r q^x \\ &= \frac{(x+r-1)!}{x!(r-1)!} p^r q^x \\ &= \frac{\Gamma(x+r)}{x!\,\Gamma(r)} p^r q^x \end{aligned} \tag{3.2}$$

where $p \in [0, 1]$, $q = 1 - p$, $r = 1, 2, 3, \ldots$ and $x = 0, 1, 2, 3, \ldots$.

The standard formulation of the Negative Binomial distribution is such that the mean and variance of random variable X are both derived quantities whose values are obtained from the primary parameters r and p as:

$$E[X] = \mu = \frac{rq}{p} \quad \text{and} \quad \text{Var}(X) = \sigma^2 = \frac{rq}{p^2} = \mu + \frac{1}{r}\mu^2 \tag{3.3}$$

The parameter r of the Negative Binomial distribution was regarded as the dispersion parameter, and in particular the overdispersion parameter since r was a positive integer. The reason for this regard was that the variance of X, σ^2 in equation (3.3) exceeded the mean, μ by a function of the parameter r.
The current study assumed the second version of the Negative Binomial distribution in equation (3.2). Further, this study utilized an alternative parameterization of the Negative Binomial distribution to allow the use of the model for cases of both overdispersion and equidispersion as described in chapter 1.

3.2.1 Parameterization of the Negative Binomial Distribution for the Proposed Change Point Algorithm

Under the second formulation of the Negative Binomial distribution with the probability mass function defined in equation (3.2), the mean, μ, and variance, σ^2, were considered as the primary quantities, while the parameters r and p were treated as derived quantities. The parameters r and p were expressed in terms of the mean and variance of the data distribution as:

$$r = \frac{\mu^2}{\sigma^2 - \mu}, \quad \text{and} \quad p = \frac{r}{r + \mu}. \tag{3.4}$$

Using this parameterization, the values of r and p were computed directly from the mean, μ, and variance, σ^2, of the data distribution. The factor $1/r$ in the variance expression defined in equation (3.3) was interpreted as a dispersion or "clumping" parameter. As the difference between the variance and the mean decreased, that is, as $\sigma^2 - \mu \to 0$, it followed that $r \to \infty$ and consequently $1/r \to 0$. In this case, the variance $\sigma^2 = \mu + \frac{\mu^2}{r}$ converged to the mean μ, and the Negative Binomial distribution approached the Poisson distribution.

$$\lim_{r\to\infty} NB\left(r, p = \frac{r}{r+\mu}\right) = Pois(\mu). \tag{3.5}$$

An alternative scenario was considered when the dispersion factor $1/r$ was large. This situation arose when the variance exceeded the mean, indicating that the data were overdispersed. As the difference $\sigma^2 - \mu$ increased, the parameter r in equation 3.4 decreased, approaching zero, and consequently $1/r$ increased without bound. The parameter $1/r$ therefore served as a measure of dispersion, with larger values corresponding to a higher degree of overdispersion in the data.

The Poisson distribution is commonly used in literature to model count data under the assumption of equidispersion, where the mean and variance are equal. However, equation (3.5) demonstrated that the Negative Binomial distribution reduced to the Poisson distribution as the dispersion parameter r increased. This property made the Negative Binomial distribution suitable for modeling count data exhibiting overdispersion and provided a flexible alternative to the standard Poisson model.

3.2.2 The Likelihood Function of NB(r, p) Distribution

Given a sample $x_1, x_2, x_3, \ldots x_n$ from a NB(r, p) distribution with probability distribution $f(x; r, p)$ described as in equation (3.2). The likelihood function was obtained as:

$$\begin{aligned} L(r,p) &= \prod_{i=1}^{n} \left[f(x_i; r, p)\right] \\ &= \prod_{i=1}^{n} \left[\frac{\Gamma(x_i + r)}{x_i!\,\Gamma(r)}\, p^r\, q^{x_i}\right] \\ &= \left(\prod_{i=1}^{n} \frac{\Gamma(x_i + r)}{x_i!}\right) \left(\frac{1}{\Gamma(r)}\right)^n p^{nr}\, q^{\sum_{i=1}^{n} x_i} \end{aligned} \tag{3.6}$$

The log-likelihood function was given by:

$$l(r,p) = \sum_{i=1}^{n} \ln\left\{\frac{\Gamma(x_i+r)}{x_i!\,\Gamma(r)} p^r q^{x_i}\right\} \tag{3.7}$$

$$= \sum_{i=1}^{n} \{\ln(\Gamma(x_i+r)) - \ln(\Gamma(r)) - \ln(x_i!) + \ln(p^r) + \ln(q^{x_i})\}$$

$$= \sum_{i=1}^{n} \ln(\Gamma(x_i+r)) - n\ln(\Gamma(r)) - \sum_{i=1}^{n} \ln(x_i!) + nr\ \ln(p) + \sum_{i=1}^{n} x_i\ \ln(q) \tag{3.8}$$

3.2.3 Regularity Conditions for Maximum Likelihood Estimation

The asymptotic properties of the maximum likelihood estimators and the likelihood ratio test employed in this study rely on a number of standard regularity conditions. These conditions ensure the consistency, asymptotic normality and asymptotic efficiency of the maximum likelihood estimators and provide the theoretical basis for the asymptotic distribution of the likelihood ratio statistic.

Comprehensive discussions of these regularity conditions are provided by Wasserman (2013) and Pawitan (2013) while the asymptotic framework adopted in this study follows the likelihood ratio theory developed by Gombay and Horvath (1996) and the general asymptotic results of van der Vaart (2000). These regularity conditions are standard in likelihood-based statistical inference and provide the theoretical justification for the asymptotic properties of maximum likelihood estimation and likelihood ratio testing (Wasserman, 2013; Pawitan, 2013). Under these conditions, the maximum likelihood estimators are consistent, asymptotically normally distributed and asymptotically efficient, while the likelihood ratio statistic converges to its limiting distribution, thereby providing a rigorous theoretical basis for changepoint estimation and hypothesis testing within the proposed hybrid likelihood framework (Gombay and Horvath, 1996; van der Vaart, 2000).

3.2.4 Theoretical Properties of the Proposed Estimator

The proposed hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm is founded on maximum likelihood estimation and likelihood ratio inference. Consequently, its theoretical properties are inherited from classical likelihood theory

(Severini, 2021; van der Vaart, 2000) under the regularity conditions presented in Section 3.2.3. Although the primary objective of this study was the development, implementation and empirical evaluation of the proposed methodology, the estimator is expected to possess several desirable large-sample statistical properties.

Consistency

Under the standard regularity conditions for maximum likelihood estimation, the maximum likelihood estimators of the model parameters are consistent. That is,

$$\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}, \qquad n \to \infty,$$

where $\boldsymbol{\theta}$ denotes the true parameter vector and $\hat{\boldsymbol{\theta}}$ its maximum likelihood estimator. Consequently, as the sample size increases, the estimated changepoint locations are expected to converge to their true values, provided that the changepoints are identifiable and sufficiently separated.

Asymptotic Normality

Maximum likelihood estimators are asymptotically normally distributed. Specifically,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} N\left(\mathbf{0}, I(\boldsymbol{\theta})^{-1}\right),$$

where $I(\boldsymbol{\theta})$ denotes the Fisher Information Matrix. This property provides the theoretical basis for statistical inference and for constructing approximate confidence intervals for the estimated model parameters.

Statistical Efficiency

Under the stated regularity conditions, the maximum likelihood estimators are asymptotically efficient in the sense that they attain the Cramér–Rao lower bound asymptotically. Consequently, no regular unbiased estimator possesses a smaller asymptotic variance than the maximum likelihood estimator.

Asymptotic Behaviour of the Likelihood Ratio Statistic

The likelihood ratio statistic employed for changepoint detection follows the asymptotic distribution established by Gombay and Horváth (1990). This asymptotic result provides the theoretical justification for the critical values used throughout the proposed hypothesis testing procedure and ensures that the significance levels of the test are asymptotically correct.

Algorithmic Convergence

The recursive segmentation procedure terminates after a finite number of iterations because each accepted changepoint partitions the sequence into smaller sub-sequences, while recursion ceases whenever the likelihood ratio statistic fails to exceed the prescribed asymptotic critical value. Consequently, the algorithm is guaranteed to terminate after a finite number of recursive steps.

The present study establishes these theoretical properties through their reliance on well-established likelihood theory and validates the practical performance of the proposed methodology using extensive simulation studies and application to real-world COVID-19 infection data. Formal mathematical proofs of estimator consistency, convergence rates, asymptotic optimality and finite-sample properties for the proposed hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm remain important avenues for future methodological research.

3.3 Dispersion Test

A count data distribution may exhibit one of three forms of dispersion: under-dispersion, equi-dispersion, or over-dispersion. The nature of dispersion is determined by the relationship between the mean and variance of the data. A commonly used descriptive measure is the variance-to-mean ratio (VMR), also referred to as the index of dispersion, defined as:

$$D = \frac{S^2}{\bar{X}} \tag{3.9}$$

where S^2 denotes the sample variance and $\bar{X}$ represents the sample mean. The value of D provides an initial indication of the dispersion pattern in the data. Specifically, $D \approx 1$ suggests equi-dispersion, $D > 1$ indicates over-dispersion, and $0 < D < 1$

implies under-dispersion. These relationships may guide the selection of candidate probability models for count data, as summarized in Table **??**.

Table 3.2: Indicative Mean–Variance Relationships and Candidate Distributions

Dispersion Index	Dispersion Type	Candidate Distribution(s)
$0 < D < 1$	Under-dispersion	Binomial; Conway–Maxwell–Poisson (subject to context)
$D \approx 1$	Equi-dispersion	Poisson
$D > 1$	Over-dispersion	Negative Binomial; Generalized Poisson

Note. D denotes the dispersion index, computed as the ratio of the sample variance to the sample mean.

While the VMR provides a useful descriptive measure, formal statistical inference is required to assess whether the observed dispersion significantly departs from equi-dispersion. In this study, the index of dispersion was computed for the full sample as well as for each data partition prior to the application of the change point detection algorithm. The dispersion test was conducted under the null hypothesis that the data were equi-dispersed, consistent with a Poisson process:

$$H_0 : D = 1 \quad \text{(Equi-dispersion)}$$
$$H_1 : D \neq 1 \quad \text{(Departure from equi-dispersion)}$$

For small sample sizes ($n < 30$), the test statistic is given by:

$$d = (n-1)D \tag{3.10}$$

Under the null hypothesis, d is approximately distributed as a chi-square random variable with $(n-1)$ degrees of freedom. The null hypothesis is rejected at significance level α if:

$$d < \chi^2_{\alpha/2,\, n-1} \quad \text{or} \quad d > \chi^2_{1-\alpha/2,\, n-1}$$

where $\chi^2_{\alpha/2,\, n-1}$ and $\chi^2_{1-\alpha/2,\, n-1}$ denote the lower and upper critical values of the chi-square distribution, respectively. Rejection in the lower tail indicates under-dispersion, while rejection in the upper tail indicates over-dispersion.

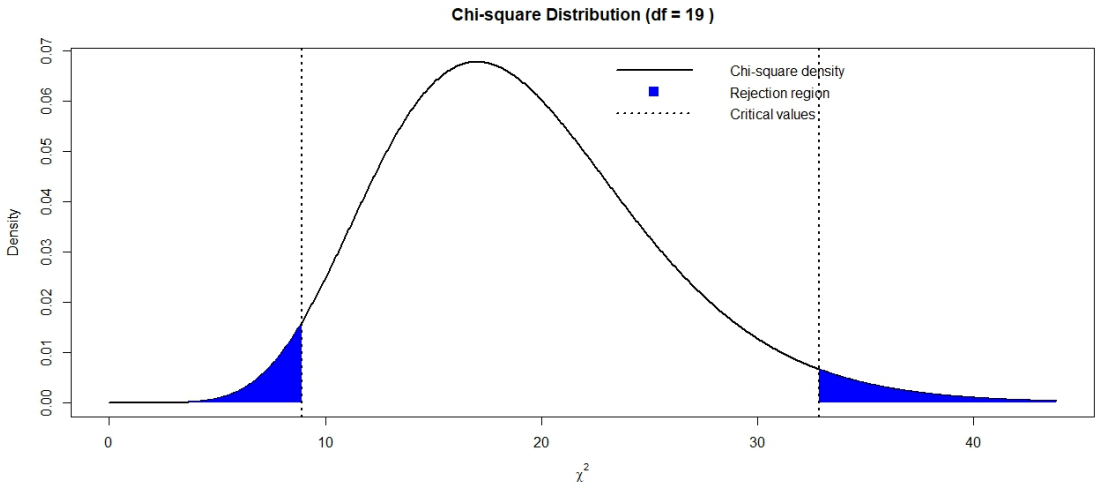

Figure 3.4: Two-Tailed Chi-Square Chart with Rejection Regions Shown

For larger samples ($n \geq 30$), a normal approximation may be employed using the transformation:

$$d = \sqrt{2(n-1)D} - \sqrt{2(n-1)-1} \tag{3.11}$$

Under the null hypothesis, d is approximately standard normally distributed. The null hypothesis is rejected at significance level α if:

$$|d| > z_{\alpha/2}$$

A positive value of d suggests over-dispersion, whereas a negative value indicates under-dispersion. Important to note, the index of dispersion test was based on sample moments and did not account for covariate effects or model structure. Consequently, the VMR served as an initial diagnostic tool. Where appropriate, further assessment of dispersion may be conducted within a regression framework to account for potential heterogeneity in the data.

3.4 The Negative Binomial Multiple Change Point Algorithm

The Negative Binomial Multiple Change Point Algorithm (NBMCPA) algorithm was developed as an iterative 3-step process (see figure 3.5) involving recursive binary segmentation, hypothesis testing for existence of statistically significant change, and estimation of the change points, if any, using maximum likelihood approach.

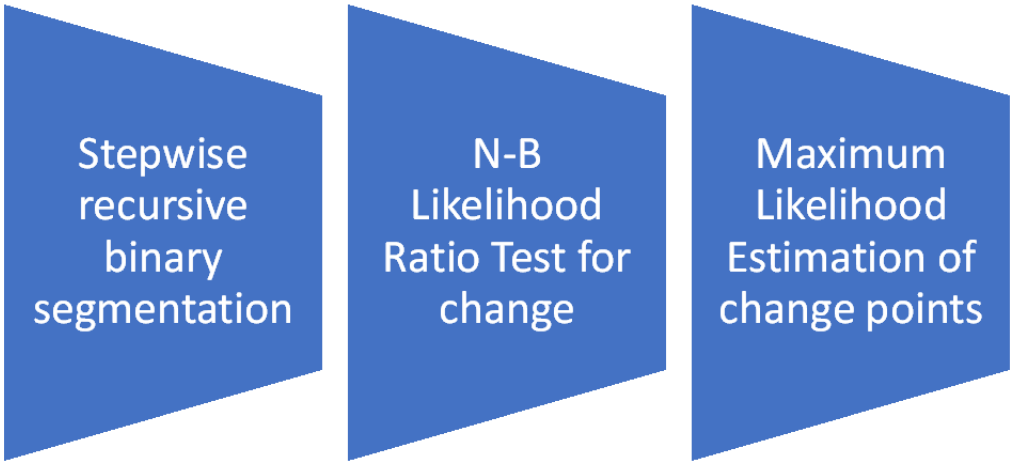

Figure 3.5: The Negative Binomial Multiple Change Point Algorithm

3.4.1 Assumptions for the NBMCP Algorithm

1. **Step changes occur in distribution parameters.** The model assumed that the data may exhibit step changes, which can affect either the mean, the variance, or specific parameters of the distribution. In this context, the proposed algorithm focuses on detecting simultaneous step changes in both parameters r and p of the Negative Binomial distribution at the same change points.

2. **Simultaneous change in parameters r and p.** It is assumed that any change in the distribution occurs as a sudden shift (step change), and that both parameters r (dispersion parameter) and p (probability parameter) change at the same time at each detected change point.

3. **No temporal dependencies in the Data.** The sample data are assumed to be independent over time, implying that observations do not depend on previous values. This means that there is no autocorrelation or temporal dependence in the data. It is further assumed that any existing temporal dependence must be removed before applying the algorithm. If present, such dependencies should be addressed using techniques like differencing or averaging to satisfy model assumptions

4. **Parameterization dependency between r and p.** Under the adopted parameterization, the parameter r depends only on the mean μ and variance σ^2, while

p depends on both r and μ. Therefore, a change in parameter r automatically implies a change in parameter p.

5. **Focus on changes in the overdispersion parameter r.** Due to the dependency between parameters, the algorithm explicitly tests only for a step change in the overdispersion parameter r. The corresponding change in p is inferred implicitly, thereby simplifying the hypothesis testing procedure.

6. **Equivalence to mean-based changepoint detection.** The proposed algorithm is expected to yield the same number and location of change points as an algorithm that seeks to detect step changes in the mean of the Negative Binomial distribution under the standard parameterization.

7. **Random count data generating Process.** The data are assumed to be generated from a random count process and may exhibit either over-dispersion or equi-dispersion. This assumption is consistent with the properties of the Negative Binomial distribution.

8. **Verification of dispersion structure.** It is assumed that the dispersion characteristics of the data can be assessed using a dispersion test, as described in Section 3.3, to determine whether the data are over-dispersed or equi-dispersed.

3.4.2 The Step-Wise Recursive Binary Segmentation (SWRBS) Procedure

The proposed Step-Wise Recursive Binary Segmentation (SWRBS) procedure is a recursive likelihood-based algorithm developed to identify multiple changepoints in count data modelled under the Negative Binomial distribution. The procedure combines recursive binary segmentation with likelihood ratio hypothesis testing and maximum likelihood estimation to iteratively partition the data into increasingly homogeneous segments. At each stage, the algorithm evaluates whether sufficient statistical evidence exists to justify further partitioning of the sequence. The overall SWRBS procedure is illustrated in Figure 3.6.

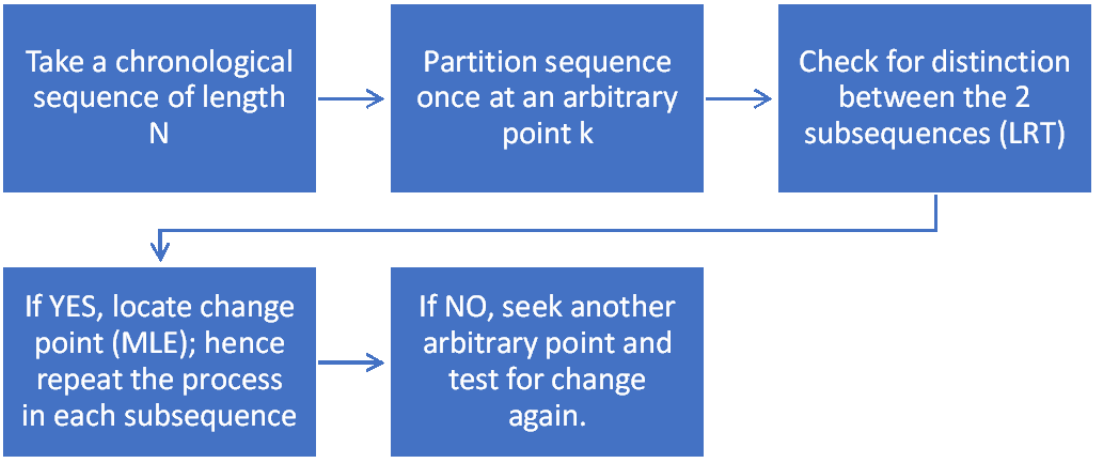

Figure 3.6: The Stepwise Recursive Binary Segmentation Procedure

Starting with a chronological sequence of length N assumed to follow a Negative Binomial(r, p) distribution, the algorithm first considers all admissible candidate changepoint locations satisfying the trimming conditions described in subsection 3.4.5. For each candidate location k, the sequence is partitioned into two sub-sequences and the likelihood ratio statistic is computed by comparing the null hypothesis of a common parameter against the alternative hypothesis of distinct parameters before and after the candidate changepoint.

The candidate location producing the maximum likelihood ratio statistic $Z_n = \max_{2\leq k\leq n-2} \Lambda_k$, is selected as the estimated changepoint. The statistical significance of this candidate changepoint is then evaluated using the likelihood ratio test described in Section 3.3. A changepoint is accepted only if $Z_n > C_\alpha$, where C_α denotes the asymptotic critical value corresponding to the chosen significance level α. Thus, the likelihood ratio test provides the *threshold criterion* governing recursive partitioning. Only statistically significant changepoints are retained, thereby reducing the likelihood of spurious detections arising from random sampling variability.

Once a statistically significant changepoint has been identified, the original sequence is partitioned into two sub-sequences at the estimated changepoint. The same estimation and hypothesis testing procedure is then applied independently to each resulting segment. This recursive process continues until no further statistically significant changepoints are detected within any partition. Consequently, the stopping rule adopted in this

study is entirely determined by statistical significance: recursion terminates whenever the maximum likelihood ratio statistic within a segment fails to exceed the corresponding asymptotic critical value. This stopping criterion ensures that additional partitions are introduced only when supported by sufficient statistical evidence, thereby avoiding unnecessary over-segmentation of the data.

The SWRBS procedure was selected in preference to alternative multiple changepoint algorithms such as Pruned Exact Linear Time (PELT), Moving Sum (MOSUM) and Wild Binary Segmentation (WBS) for several methodological reasons. First, recursive binary segmentation integrates naturally with the likelihood ratio testing framework developed in this study, allowing changepoint estimation and hypothesis testing to be performed simultaneously within a unified statistical framework. Second, the recursive nature of the algorithm provides a transparent and easily interpretable estimation procedure, making it straightforward to identify the sequence of detected changepoints and the corresponding data partitions.

Furthermore, SWRBS offers favourable computational efficiency for datasets of moderate size, such as the COVID-19 infection data analysed in this study, while remaining considerably simpler to implement than optimisation-based procedures such as PELT. Although PELT guarantees globally optimal segmentation under additive cost functions, its implementation requires specification of an appropriate penalty function and optimisation criterion, neither of which aligns directly with the likelihood ratio hypothesis-testing framework adopted in this study. Similarly, MOSUM is primarily designed for detecting local changes through moving-window statistics and is less naturally integrated with likelihood-based estimation of Negative Binomial parameters. Wild Binary Segmentation improves detection of closely spaced changepoints through random interval selection; however, its stochastic nature introduces additional algorithmic complexity and was not considered necessary for the relatively well-separated changepoints investigated in this study.

The proposed SWRBS procedure therefore provides an appropriate compromise between statistical rigour, computational efficiency, interpretability and compatibility with the proposed hybrid likelihood-based framework. Moreover, the recursive segmentation strategy enables direct integration of maximum likelihood estimation, likelihood ratio testing and asymptotic hypothesis testing within a single coherent methodology for detecting multiple changepoints in both equi-dispersed and over-dispersed count data.

3.4.2.1 Partitioning the Sequence

A sequence of random calendar time points was considered: $t_0, t_1, t_2, \cdots, t_N$. Let $x_1, x_2, \ldots, x_n$ be a sequence of observed values or realizations of a stochastic process. The interval $[t_0, t_1]$ represented the length of the partition between consecutive time points 0 and 1, while $[t_1, t_2]$ represented the length of the partition between consecutive time points 1 and 2, and so on. It was assumed that each partition $[t_j, t_k]$ for $j \in [1, N]$ and $k \in [0, N]$, with $j \neq k$, of the original sequence $\{x_i\}$ consisted of a random sub-sequence $\{x_h, x_{h+1}, x_{h+2}, \ldots\}$. It was further assumed that the observations in each partition followed a Negative Binomial distribution with parameter r_j for $j \in [1, n]$, as shown in Figure 3.7.

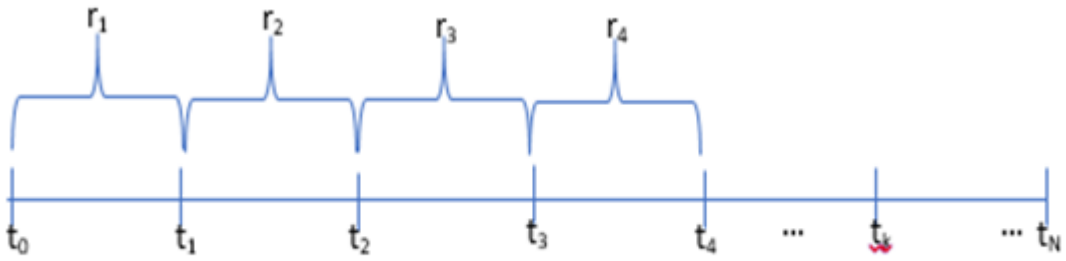

Figure 3.7: Timeline Diagram Showing Sequence Partitions

3.4.3 Computational Complexity and Scalability of the Proposed Algorithm

An important consideration in the development of any recursive changepoint detection algorithm is its computational efficiency, particularly when analysing long sequences of observations. The computational complexity of the proposed hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm is determined primarily by the number of likelihood ratio evaluations performed during recursive partitioning and the recursive nature of the segmentation procedure.

For a sequence of length n, the likelihood ratio statistic is evaluated at each admissible candidate changepoint. Consequently, the computational cost of a single segmentation stage is proportional to the number of candidate locations, giving a computational complexity of $O(n)$.

Whenever a statistically significant changepoint is detected, the sequence is partitioned into two sub-sequences and the estimation procedure is applied recursively to each

partition. Let $T(n)$ denote the computational cost for analysing a sequence of length n. The recursive nature of the algorithm may therefore be represented by the recurrence relation

$$T(n) = T(n_1) + T(n_2) + O(n),$$

where $n_1 + n_2 = n$.

For reasonably balanced partitions, where $n_1 \approx n_2 \approx \frac{n}{2}$, the recurrence becomes $T(n) = 2T\left(\frac{n}{2}\right) + O(n)$ which, by the Master Theorem, yields an expected computational complexity of

$$T(n) = O(n \log n).$$

This complexity is comparable to that of conventional recursive binary segmentation procedures and is computationally efficient for moderate and large datasets.

In the worst-case scenario, where successive changepoints occur very close to one end of the sequence, the recursive partitions become highly unbalanced. In this case, the recurrence relation approaches

$$T(n) = T(n-1) + O(n),$$

which gives a worst-case computational complexity of $T(n) = O(n^2)$.

Although this represents the theoretical upper bound, such highly unbalanced partitions are uncommon in practice and were not encountered during either the simulation study or the application to the Kenyan COVID-19 infection data.

From a scalability perspective, the proposed algorithm is well suited to datasets containing several hundred to several thousand observations because recursive partitioning substantially reduces the size of the search space at each stage of the analysis. Moreover, the independent treatment of each partition provides opportunities for parallel implementation, whereby different segments may be analysed simultaneously using multi-core computing environments. Consequently, the proposed methodology can readily be extended to larger datasets without fundamental modification of the estimation procedure.

While optimisation-based algorithms such as the Pruned Exact Linear Time (PELT) algorithm may achieve linear computational complexity under appropriate penalty

functions, the proposed hybrid NBMCPA was designed to prioritise statistical interpretability, direct integration with likelihood ratio hypothesis testing, and flexibility for modelling both equi-dispersed and over-dispersed count data within a unified likelihood framework. The resulting computational performance was found to be satisfactory for both the simulation study and the real-world application presented in this research.

3.4.4 The Likelihood Ratio Test for Existence of the First Change

An investigation into whether a change existed at some random point k in a sequence of observations was conducted using a likelihood ratio test. The null hypothesis stated that there was no change in the over-dispersion parameter r across the two data segments formed with the boundary at point k. The alternative hypothesis stated that a difference existed in the over-dispersion parameter r between the two data segments at point k. The first partition of the sequence in the interval $[t_0, t_k]$ had parameter r_α, while the second partition in the interval $[t_{k+1}, t_N]$ had parameter r_β, where $r_\beta \neq r_\alpha$. The mathematical hypotheses were expressed as in equation (3.12):

$$\begin{aligned} &H_0 : r_1 = r_2 = \ldots = r_N = r \\ &\text{versus} \\ &H_1 : r_1 = r_2 = \ldots = r_k = r_\alpha \neq r_{k+1} = r_{k+2} = \ldots = r_N = r_\beta \end{aligned} \tag{3.12}$$

The log-likelihood function under H_0 was:

$$l_0(\hat{r}, \hat{p}|x_i) = \sum_{i=1}^{n} \ln \left\{ \frac{\Gamma(x_i + \hat{r})}{x_i! \Gamma(\hat{r})} \hat{p}^{\hat{r}} \hat{q}^{x_i} \right\} \tag{3.13}$$

where;

$$\hat{r} = \frac{\bar{X}^2}{S^2 - \bar{X}}, \quad \hat{p} = \frac{\hat{r}}{\hat{r} + \bar{X}}, \quad \hat{q} = 1 - \hat{p}$$

The statistics S^2 and $\bar{X}$ were the unbiased estimates of the population variance and mean, respectively, obtained as:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{X})^2}{n - 1}$$

The log-likelihood function under H_1 was given by:

$$l_1\left(\hat{r}_\alpha, \hat{r}_\beta, \hat{p}_\alpha, \hat{p}_\beta \middle| x_i\right) = \sum_{i=1}^{k} \ln\left\{\frac{\Gamma\left(x_i + \hat{r}_\alpha\right)}{x_i!\,\Gamma(\hat{r}_\alpha)}\left(\hat{p}_\alpha\right)^{\hat{r}_\alpha}\left(\hat{q}_\alpha\right)^{\hat{x}_i}\right\} + \sum_{i=k+1}^{n} \ln\left\{\frac{\Gamma\left(x_i + \hat{r}_\beta\right)}{x_i!\,\Gamma(\hat{r}_\beta)}\left(\hat{p}_\beta\right)^{\hat{r}_\beta}\left(\hat{q}_\beta\right)^{\hat{x}_i}\right\} \tag{3.14}$$

The parameter estimates of r_α, r_β, p_α, p_β, q_α, and q_β were obtained as:

$$\hat{r}_\alpha = \frac{\bar{X}_k^2}{S_k^2 - \bar{X}_k}, \quad \hat{p}_\alpha = \frac{\hat{r}_\alpha}{\hat{r}_\alpha + \bar{X}_k}, \quad \hat{q}_\alpha = 1 - \hat{p}_\alpha$$

$$\hat{r}_\beta = \frac{\bar{X}_{n-k}^2}{S_{n-k}^2 - \bar{X}_{n-k}}, \quad \hat{p}_\beta = \frac{\hat{r}_\beta}{\hat{r}_\beta + \bar{X}_{n-k}}, \quad \hat{q}_\beta = 1 - \hat{p}_\beta$$

The statistics S_k^2 and $\bar{X}_k$ were the unbiased estimates of the sub-population variance and mean for the first k observations, obtained as:

$$\bar{X}_k = \frac{\sum_{i=1}^{k} x_i}{k}, \quad S_k^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{X}_k)^2}{k-1}$$

Similarly, the statistics S_{n-k}^2 and $\bar{X}_{n-k}$ were the unbiased estimates of the sub-population variance and mean for the next $n-k$ observations, obtained as:

$$\bar{X}_{n-k} = \frac{\sum_{i=k+1}^{n} x_i}{n-k}, \quad S_{n-k}^2 = \frac{\sum_{i=k+1}^{n}(x_i - \bar{X}_{n-k})^2}{n-k-1}$$

The likelihood ratio statistic at a candidate changepoint t_k was defined consistently with the classical likelihood ratio testing framework discussed in Casella & Berger (2002) and Chen & Gupta (2012) and based on the log-likelihood functions defined in equations 3.13 and 3.14 as:

$$\hat{\Lambda}_k = -2\left(l_0(\hat{r}, \hat{p}) - l_1(\hat{r}_\alpha, \hat{r}_\beta, \hat{p}_\alpha, \hat{p}_\beta)\right) \tag{3.15}$$

where l_0 denotes the maximized log-likelihood under the null hypothesis of no structural change, assuming a single Negative Binomial model for the entire series, and l_1 denotes

the maximized log-likelihood under the alternative hypothesis of a change at t_k, with separate parameter sets for the two segments.

Under standard regularity conditions, the likelihood ratio statistic asymptotically follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the competing models. However, in the present changepoint setting, the location of the changepoint is not identified under the null hypothesis, violating these regularity conditions. Consequently, the asymptotic distribution of $\hat{\Lambda}_k$ deviates from the classical χ^2 distribution, and inference was based on specialized asymptotic results for critical values as discussed in section 3.4. The change point t_k was estimated such that the likelihood ratio test statistic was maximized with the optimal value Z_n estimated as:

$$\hat{Z}_n = \max_{2 \leq k \leq n-2} \hat{\Lambda}_k \tag{3.16}$$

The decision rule was such that H_0 in equation (3.12) was rejected if the test statistic satisfied $\hat{Z}_n > C$. The constant C was a critical value determined by the size of the test α, the sample size n, and the null distribution of the likelihood ratio test statistic, as described in Gombay and Horvath (1990). Otherwise, small values of $\hat{Z}_n$ indicated that no statistically significant change existed at the time point t_k.

3.4.5 Testing for the Second and Subsequent Change Points

Once the first change point was obtained, the likelihood ratio test was repeated within each of the two data segments formed. For example, if there were $n = 100$ observations and the first change point was estimated at time t_{20}, the data were divided into two segments: $t_1, \ldots, t_{20}$ and $t_{21}, \ldots, t_{100}$. While a single change point algorithm would have stopped at this stage, the multiple change point algorithm continued by testing for additional change points in each segment separately. To identify the second change point, the hypotheses in equation 3.17 were tested:

$$\begin{aligned} &H_0 : r_1 = r_2 = \ldots = r_{20} = r \\ &\text{versus} \\ &H_1 : r_1 = r_2 = \ldots = r_k = r_\alpha \neq r_{k+1} = r_{k+2} = \ldots = r_{20} = r_\beta \end{aligned} \tag{3.17}$$

To identify the third change point, the hypotheses in equation 3.18 were tested:

$$H_0 : r_{21} = r_{22} = \ldots = r_{100} = r$$

versus

$$H_1 : r_{21} = r_{22} = \ldots = r_k = r_\alpha \neq r_{k+1} = r_{k+2} = \ldots = r_{100} = r_\beta \tag{3.18}$$

The existence and location of the second and third change points resulted in further sub-partitions of the data, producing four segments in total for the sequence of size $n = 100$. Each of these segments was then tested for change, and the process was repeated, hence the name step-wise recursive binary segmentation. The position of any valid change point t_k satisfied the inequality $t_2 \leq t_k \leq t_{n-2}$, ensuring that a change point could not occur at the first or last observation. Instead, the change point was required to lie between two observations. In the context of multiple change point analysis, given a sample of size n, the maximum possible number of change points was $n - 3$, excluding the first and last two observations.

3.4.6 Estimation of the Dispersion Parameter

The dispersion parameter, r, plays a fundamental role in the Negative Binomial model because it governs the degree of over-dispersion present in the count data. Unlike the Poisson distribution, where the variance equals the mean, the Negative Binomial distribution accommodates additional variability through

$$\text{Var}(X) = \mu + \frac{\mu^2}{r},$$

where μ denotes the mean of the distribution. Smaller values of r correspond to greater over-dispersion, while as $r \to \infty$, the Negative Binomial distribution converges to the Poisson distribution.

In the proposed methodology, the dispersion parameter was estimated simultaneously with the remaining model parameters using the method of maximum likelihood. Maximum likelihood estimation was selected because of its desirable large-sample properties, including consistency, asymptotic normality and asymptotic efficiency under the regularity conditions discussed in Section 3.2.3. (Pawitan, 2013; van der Vaart, 2000). The estimated value of r was subsequently used in constructing the likelihood ratio statistic for changepoint detection.

The precision of the maximum likelihood estimator of r may be assessed using the observed Fisher information matrix. Under standard regularity conditions, approximate $(1-\alpha)\times 100\%$ confidence intervals for the dispersion parameter can be constructed as

$$\hat{r} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{r})},$$

where $\widehat{\mathrm{Var}}(\hat{r})$ is obtained from the inverse of the observed Fisher information matrix and $z_{\alpha/2}$ denotes the corresponding standard normal quantile (Pawitan, 2013; Wasserman, 2013). Although confidence intervals were not the primary focus of the present study, they provide a useful measure of estimation uncertainty and may be incorporated in future applications of the proposed methodology.

The performance of likelihood-based changepoint detection depends, to some extent, on accurate estimation of the dispersion parameter. Underestimating the degree of over-dispersion may inflate the likelihood ratio statistic and increase the probability of false changepoint detection, whereas overestimating the dispersion parameter may reduce the sensitivity of the procedure to genuine structural changes. Consequently, accurate estimation of r is essential for reliable hypothesis testing and changepoint estimation.

A formal sensitivity analysis of the dispersion parameter was beyond the scope of the present study. Nevertheless, the simulation study presented in Chapter 4 evaluated the proposed algorithm under a range of sample sizes and changepoint locations, thereby providing indirect evidence of the robustness of the estimation procedure across different data configurations. Future research may extend the proposed methodology by undertaking a comprehensive sensitivity analysis to investigate the influence of alternative dispersion levels on changepoint estimation accuracy, statistical power and false detection rates.

3.5 Determining the Critical Values for the Likelihood Ratio Test for Existence of Change

The proposed likelihood ratio test requires critical values for deciding whether the observed test statistic provides sufficient evidence to reject the null hypothesis of no changepoint. In this study, the critical values were derived from asymptotic theory rather than by Monte Carlo simulation or empirical approximation. Specifically, the

asymptotic distribution theory developed by Gombay and Horvath (1990) was adopted and extended to the proposed Negative Binomial likelihood framework.

The study makes use of the methods described by Gombay and Horvath (1996) regarding the asymptotics of maximum-likelihood ratio-type statistics for testing a sequence of observations for no change in parameters against a possible change while some nuisance parameters may remain constant over time. Gombay and Horvath obtained extreme value approximations as well as Gaussian-type approximations for the square root of the likelihood ratio in equation (3.16). They also approximated the maximum likelihood ratio using Ornstein-Uhlenbeck processes and obtained the upper bounds for the rate of approximation.

In order to get the critical values of the likelihood ratio test statistic defined in section 3.3 the current study made use of the asymptotic distribution of $\sqrt{Z_n}$ where $Z_n = \max_{2\leq k\leq n-2} \Lambda_k$ as described in equation 3.16. Different critical values were obtained for various small and medium sample sizes ($n = 12, 20, 50, 100, 200,$ and, 500), various test sizes ($\alpha = 1\%, 5\%$ and 10%).

The asymptotic critical values of $\sqrt{Z_n}$ were obtained as follows:

Let $0 < \alpha < 1$ Define:

$$z_n = z_n(1-\alpha) = \sup(x : P\{Z_n^{1/2} \leq x\} \leq 1-\alpha)$$

and

$$\begin{aligned} r(h,l) &= r(h,l;1-\alpha) \\ &= \sup(x : P\{\sup_{h\leq t\leq 1-t} \{B^{(d)}(t)/(t(1-t))\}^{1/2} \leq x\} = 1-\alpha) \end{aligned}$$

Then, according to Gombay and Horvath (1990), if the null hypothesis of no change holds and $h(n)$ and $l(n)$ are chosen such that both exceed $1/n$ then

$$\lim_{n\to\infty} P\left\{Z_n^{1/2} > r(h(n), l(n))\right\} = \alpha$$

So that $r(h(n), l(n))$ is an asymptotically correct critical value of size α.

It can be shown that for all $0 < h < 1 - l < 1$.

$$\sup_{h\leq t\leq 1-t} \left(\frac{B^{(d)}(t)}{t(1-t)}\right)^{1/2} = \sup_{0\leq t\leq \log(1-h)(1-l)/hl} \Delta(t).$$

Such that for any $T > 0$

$$P\left\{\sup_{0\leq t\leq T}\Delta(t) > x\right\} = \frac{x^d \exp(-x^2/2)}{2^{d/2}\Gamma(d/2)}\left\{T - \frac{d}{x^2}T + \frac{4}{x^2} + O\left(\frac{1}{x^4}\right)\right\} \tag{3.19}$$

The approximations for the distributions of $r(h(n), l(n))$ and $\sqrt{Z_n}$ as described in Gombay and Horvath (1990) were applied to data assumed to follow the exponential, Poisson and Normal distributions. Further, the approximations were based on a convenient choice of the parameters h and l such that

$$h(n) = l(n) = \frac{(\log n)^{(3/2)}}{n}$$

The current study, extended the methods in Gombay and Horvath (1990) to data obtained from the Negative Binomial distribution. Further, alternative values of $h(n) = l(n)$ were tested in an attempt to get the best asymptotic critical values .

3.5.1 Applying the Negative-Binomial Multiple Change Point Algorithm to Simulated Data

The change point algorithm developed in section 3.4 were tested using a set of simulated data from the Negative Binomial (r, p) distribution. The simulation study was performed using the R statistical software for $10,000$ iterations. In order to check the performance of the algorithm at different changepoint locations, the study considered a set of three predetermined changepoints in the simulated data for a sample of size n. The change points were set at the locations: $n/4$, $n/2$, and $3n/4$ where parameters r and p were assumed to change simultaneously, while the distributional form remained the same throughout the sample.

The algorithm was fitted iteratively to simulated data of different sample sizes and the test statistic values stored for each value of possible change point k, where $2 \leq k \leq n-2$. Graphs of the likelihood ratio statistic were plotted for each sample size and their maximum values compared to the critical values of the likelihood ratio test obtained in section 3.2 to check for significance of change.

3.5.1.1 The Null Model: No Changepoint

The simplest model considered in this study assumes that the entire sequence of count observations arises from a single Negative Binomial distribution with constant parameters throughout the observation period. Under this model, no structural change is assumed to occur, and all observations are governed by a common mean and dispersion parameter.

Let $x_1, x_2, \ldots, x_n$ denote a sequence of independent count observations such that $x_i \sim NB(r, p), \qquad i = 1, 2, \ldots, n$ where r denotes the dispersion parameter and p represents the success probability of the Negative Binomial distribution.

The null hypothesis is therefore defined as

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_n = \theta,$$

indicating that the underlying distribution remains constant throughout the observation period.

Under the null model, a single likelihood function is fitted to the entire dataset, and the corresponding maximum likelihood estimates are obtained using all observations. This model serves as the reference model against which alternative changepoint models are compared using the likelihood ratio test developed.

The no-changepoint model provides the baseline for evaluating whether observed fluctuations in the data can be attributed solely to random sampling variability or whether they indicate statistically significant structural changes in the underlying count process.

3.5.1.2 The Single-Changepoint Model

The next level of complexity assumes that the sequence contains a single structural change occurring at an unknown location k. The objective is to estimate both the location of the changepoint and the corresponding model parameters before and after the change.

Suppose that $2 \leq k \leq n - 2$. The sequence is partitioned into two segments, $x_1, \ldots, x_k$ and $x_{k+1}, \ldots, x_n$,

which are assumed to follow separate Negative Binomial distributions,

$$x_i \sim NB(r_1, p_1), \qquad i \leq k,$$

and

$$x_i \sim NB(r_2, p_2), \qquad i > k.$$

The corresponding hypotheses were:

$$H_0 : (r_1, p_1) = (r_2, p_2),$$

against

$$H_1 : (r_1, p_1) \neq (r_2, p_2).$$

For each admissible candidate location k, the likelihood ratio statistic was computed by comparing the likelihood under the null model with the likelihood obtained after partitioning the sequence at k. The candidate location producing the maximum likelihood ratio statistic was selected as the estimated changepoint,

$$\hat{k} = \arg \max_{2 \leq k \leq n-2} \Lambda_k.$$

The statistical significance of the estimated changepoint is subsequently assessed using the asymptotic critical values derived earlier. If the likelihood ratio statistic exceeds the corresponding critical value, the null hypothesis of parameter homogeneity is rejected and a single changepoint is declared. Otherwise, the sequence is regarded as containing no statistically significant structural change.

The single-changepoint model provides the methodological foundation for the simulation experiments presented in Chapter 4 and forms the basic building block of the proposed recursive multiple changepoint algorithm. The multiple changepoint model considered in this study was a natural extension of the single-changepoint framework, in which recursive binary segmentation was employed to identify successive statistically significant changepoints until no further evidence of structural change remained.

3.6 Assessing the Performance of the NBMCP Algorithm

Evaluation of performance of the NBMCPA for detecting and locating change points considers the view of changepoint detection similar to a binary classification problem,

where each data point is expected to belong to either the change-point class or non-change point class (van den Burg and Williams, 2020). The step-wise recursive binary segmentation procedure is applied to simulated datasets, and the change detection test conducted on each segment formed with the boundary at some value k, the possible change point. If a variation in the dispersion parameter is found to exist between the two data segments, then k falls into the detected change-points class, otherwise it belongs in the non-change point class. The detected change points may or not coincide with the true changepoints that are fixed in the data simulations; hence the need to check how well the algorithm estimates change points.

In this study, the NBMCPA was evaluated through computation of various metrics based on the concepts of a confusion matrix for the binary classification problem in the change-point detection setting. The metrics, including sensitivity or recall, accuracy, precision, and empirical coverage, were used in the analysis of change detection error. In addition to numerical measures, the kernel density estimate (KDE) of the distribution of estimated change points was obtained and visualized. The density curve was used to analyze the distribution of estimated changepoints in contrast to the true change points. The algorithm performance was checked by visually inspecting the features of the KDE curve at varying locations of true changepoints, while controlling for the sample size and magnitude of the distinction between data segments.

3.6.1 NBMCPA Performance Metrics

3.6.1.1 True Positive Rate (Sensitivity)

In the context of change point detection, a True Positive (TP) refers to a correctly detected change point, while a False Negative (FN) is a missed changepoint or a true changepoint that was not detected by the model. Sensitivity, evaluates how well the model detects change points that are truly present.

In the current study, for each iteration, if the model identified a change point at position k, where $k = n/4, n/2$ or $3n/4$ or within a defined tolerance, the changepoint estimate counted as a true positive. in the other hand, if the algorithm failed to detect the preset change point, then the event counted as a false negative. A high sensitivity indicated

that the model was good at detecting change points. Mathematically, the true positive rate was calculated according to Aminikhanghahi and Cook (2017) as:

$$\begin{aligned}\text{True Positive Rate} &= \frac{\text{Number of correctly detected change points}}{\text{Total number of true change points}} \\ &= \frac{\text{True Positives}}{\text{True Positives + False Negatives}}\end{aligned} \tag{3.20}$$

3.6.1.2 False Positive Rate (FPR)

This metric determines the rate of false alarms in change point detection. It measures how often the model incorrectly detects a change point where there is none. In the simulation study, a false alarm was found to occur if the model detected a changepoint outside of the expected region, so that the estimated changepoint was far from the positions $k = n/4, n/2$ or $3n/4$ respectively. A low FPR was desirable, so that the model did not falsely locate a change point too frequently.

$$\text{False Positive Rate} = \frac{\text{Number of false positives}}{\text{Total number of iterations}} \tag{3.21}$$

3.6.1.3 Change Point Location Accuracy

This metric evaluates how well the algorithm locates a true change point, particularly when it correctly detects a change point. For each iteration, the absolute difference between the estimated change point and the true change point was computed, then the location accuracy was determined as the average of these differences over all iterations. Smaller values of mean absolute error (MAE) suggested higher accuracy in locating the changepoints, and if there was a consistently low error, then the model was consistently accurate. Mathematically, the MAE was calculated according to Aminikhanghahi and Cook (2017) and Chander et al. (2024) as:

$$\text{Mean Absolute Error (MAE)} = \frac{1}{T}\sum_{i=1}^{T}|\hat{c}_i - c| \tag{3.22}$$

where c was the true/known change point, $\hat{c}_i$ was the estimated changepoint for iteration i, and T was the total number of iterations where a change point is detected.
In addition to MAE, the median and interquartile range of the absolute errors were calculated. A narrow interquartile range and a median error close to 0 would indicate

that the model was precise and accurate. A wider range or a high median error would indicate issues with precision or bias. Unlike the MAE, the Median Absolute Error was less sensitive to outliers; hence it gave the typical magnitude of error in the model predictions, providing a better measure of change location accuracy in the presence of outliers.

3.6.1.4 Precision

In the context of change point analysis, precision measures how often the detected changepoints are correct or rather, how often a estimated change point matches the true changepoint. In the current study, precision was computed as the ratio of true positives to the total number of all detected change points (among true positives and false positives). High precision meant that when the model detected a change point, it was usually correct. Mathematically, precision of the algorithm was computed according to Aminikhanghahi and Cook (2017) as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}} \tag{3.23}$$

3.6.1.5 F1 Score

The F1 score gave a combined measure of precision and true positive rate, providing a single metric to evaluate the model performance. An F1 score closer to 1 indicated a good balance between precision and recall, suggesting that the model was both accurate and reliable in detecting change points. Mathematically, the F1 score was computed from the model precision and sensitivity according to Killick et al. (2012); van den Burg and Williams (2020) as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision + Sensitivity}} \tag{3.24}$$

3.6.1.6 Empirical Coverage

In the current study, the empirical coverage was defined as the proportion of times the true changepoint fell within a certain region, or a window around the true changepoint, where the model estimated a changepoint. This approach was consistent with Killick et al. (2012) and Truong et al. (2020) who stated that in the evaluation of change

detection models using classification metrics such as precision and empirical coverage, it was common to define a margin of error or threshold around the known change location to allow for minor discrepancies. The tolerance level was defined as the allowable error in the change detection so that estimated changepoints that were close, though not exact, to the real changepoint were not classified as missed changes. Given a tolerance range, say within $\pm k$ positions of $n/2$, the proportion of iterations in which the model estimated a change point within this range was computed. A higher empirical coverage indicated that the model was correctly detecting change points close to the true location most of the time. Empirical coverage was computed according to van den Burg and Williams (2020) as:

$$\text{Empirical coverage} = \frac{\text{Number of detected change points within tolerance range}}{\text{Number of iterations}} \tag{3.25}$$

Selection of the threshold k depended on the scale of the data and magnitude of changes. If data spanned a wide range or if changes were subtle, a larger threshold was set to account for small variations that should not be penalized as false detection. For smaller datasets or data sequences where changes were more pronounced, a smaller threshold was used to ensure that the detected change points under the model were very close to the true changepoint. In the current simulation study, a threshold of $k = 1$ was used for small samples of size $n <= 60$, for sample sizes between $n = 60$ and $n = 200$ a threshold of $k = 3$ was used, while a tolerance of $k = 5$ was applied for sample sizes $200 < n < 500$.

3.6.2 Visual Analysis of Estimation Results

3.6.2.1 Histograms of Estimated Change Points

Frequency histograms were plotted to show the general distribution of the estimated change point locations across all iterations. The charts gave an idea of the scatter or spread of the model estimates around the true location. The tallest bar on the chart showed the modal changepoint estimate, while the general shape of the histogram helped to infer the distribution of estimated changepoints around the true/known value. Bars located farther from the true value pointed to the existence of false positives.

3.6.2.2 Kernel Density Estimate of Change Points

A Kernel Density Estimate (KDE) is a non-parametric way to estimate the probability density function of a random variable. For the change point estimation problem, when comparing the estimated change points against the true change point, the KDE helps to visualize how the estimated change points are distributed relative to the true change point Chen (2017). In this study, the Kernel density estimate of the detected changepoints across all iterations were plotted to provide a smoothed estimate of the data distribution. Mathematically, the KDE was computed as:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{3.26}$$

where:
$\hat{f}_h(x)$ was the estimated probability density function at point x.
n was the number of estimated changepoints
X_i was the estimated changepoint from iteration i.
h was the bandwidth parameter, controlling for the smoothness of the density estimate.
$K(\cdot)$ was the kernel function.
The current study used the Gaussian (Normal) kernel defined by:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \tag{3.27}$$

Plunging the Gaussian kernel in equation 3.27, the KDE in equation 3.26 transformed to:

$$\hat{f}_h(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{x - X_i}{h}\right)^2} \tag{3.28}$$

This kernel ensured that each estimated change point contributed to the density estimate according to the normal distribution centered at X_i, with a spread controlled by the bandwidth h. If the model was performing well, the KDE would show a peak near the true change point. The width of the peak depended on the chosen bandwidth parameter h, with a small bandwidth resulting in a more jagged estimate (with sharper peaks and valleys), while a large bandwidth gave a smoother, broader peak as discussed in Kang and Kim (2019).
Visualizing the KDE enabled assessment of whether the estimated change points were concentrated around the true change point or if there was significant bias or variability

in the model estimates. This study utilised the Silverman's Rule of Thumb to identify the optimal bandwidth parameter h as:

$$h = 0.9 \cdot \min\left(\text{std}(X), \frac{\text{IQR}(X)}{1.34}\right) \cdot n^{-1/5} \tag{3.29}$$

Where $\text{std}(X)$ is the standard deviation , $\text{IQR}(X)$ is the interquartile range, and n is the number of change points estimated.

3.7 Estimating Change Points in COVID-19 Infections Data for Kenya and Identifying Assignable Causes of Change

3.7.1 Overview

The fourth specific objective of this study was to estimate the change points, if any, in the trend of COVID-19 infections in Kenya and to determine the assignable causes of such changes. This section outlines the procedures followed, including data selection, transformation and smoothing the raw COVID-19 data sequence. Further, the methodology entails subsequent identification of potential causal events associated with structural shifts in the epidemic curve, by interpreting each changepoint with reference to specific epidemiological or policy events in order to determine plausible assignable causes

3.7.2 Data Collection, Cleaning, and Study Period

Daily confirmed COVID-19 infection counts for Kenya were obtained from the John Hopkins University Website. The study period was restricted to approximately 500 days, similar to the largest considered threshold for the critical values determined in section3.4. More specifically, the study period spanned the time interval between March 2020, when the first positive case of COVID-19 in the country was confirmed, and August 2021 in order to capture the major epidemic waves and key public health interventions implemented in Kenya during this time. The data sequence was truncated so that all observations falling outside the study period were dropped. Outliers believed to arise from reporting delays or data backlogs were assessed for epidemiological relevance and retained or adjusted based on contextual justification.

3.7.3 Exploratory Data Transformation

Let $\{x_t\}_{t=1}^{T}$ denote the raw daily COVID-19 infection counts, where $T \approx 500$ observations. and $\{y_t\}_{t=1}^{T}$ be the transformed infection counts. Daily case counts typically display right skewness, heteroscedasticity, and non-linear trends. To stabilize variance and approximate normality, three transformation functions were explored for the raw COVID-19 data sequence as follows:

1. Natural logarithm:
$$y_t = \log(x_t)$$

2. Log-plus-one transformation:
$$y_t = \log(x_t + 1)$$

3. Square-root transformation:
$$y_t = \sqrt{x_t}$$

Diagnostic checks, including distributional plots and variance assessments, were used to evaluate the suitability of each transformation. The square-root transformation exhibited the most stable variance pattern; thus the working series for the current study was defined as:
$$y_t = \sqrt{x_t}.$$

3.7.4 Smoothing Using a Moving Average

To reduce short-term variability and day-of-week reporting effects, a 7-point symmetric moving average was applied to the transformed data. The smoothed series was defined by:
$$\tilde{Y}_t = \frac{1}{7}\sum_{i=-3}^{3} y_{t+i}, \qquad t = 4, 5, \ldots, T-3.$$
This smoothing approach preserved medium-term pandemic trends, while mitigating noise from short-term variations that could otherwise influence changepoint detection. The moving average also helped to reduce the impact of random fluctuations or outliers, making patterns more visible. The Negative Binomial Multiple Changepoint model described in section 3.3 was fitted to the smoothed data series and change points in the COVID-19 infections trend estimated.

3.7.5 Identification of Assignable Causes

To address the objective of identifying assignable causes of structural changes in COVID–19 incidence, each statistically significant changepoint estimated by the proposed Hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm (HL-NBMCPA) was first mapped to its corresponding calendar date. The estimated changepoint locations were then subjected to a systematic post-estimation interpretation framework to evaluate whether the observed structural changes could plausibly be associated with external epidemiological, behavioural or administrative events occurring during the study period.

It is important to emphasize that the identification of assignable causes is performed only after the changepoint estimation procedure has been completed. Consequently, the determination of assignable causes does not form part of the proposed estimation algorithm and has no influence on the likelihood function, parameter estimation, likelihood ratio calculations, recursive segmentation procedure or hypothesis testing. The algorithm relies exclusively on the observed count data and the underlying Negative Binomial likelihood framework to estimate the locations of statistically significant changepoints. External information is introduced solely during the interpretation stage and therefore cannot influence, modify or bias the estimated changepoint locations.

Specifically, the estimated changepoint dates were cross-referenced against a curated timeline of nationally documented events and interventions obtained from official government publications, Ministry of Health situation reports, public health surveillance reports and other authoritative policy documents. The following categories of potential assignable causes were considered:

- Non-pharmaceutical interventions, including lockdowns, curfews, school closures, travel restrictions and phased reopening of economic activities;
- Changes in testing capacity, testing eligibility criteria or case reporting protocols;
- Major public holidays, mass gatherings and other events associated with increased population mobility;
- Emergence or widespread detection of new SARS-CoV-2 variants of concern;
- Milestones in the national COVID–19 vaccination programme, including programme initiation and substantial increases in vaccine coverage;

- Observable changes in population mobility, compliance with public health measures or other documented behavioural changes.

A temporal alignment criterion was adopted whereby an external event was considered plausibly associated with a detected changepoint if it occurred within a ±7-day window of the estimated changepoint. This window was selected to accommodate delays inherent in the infection-reporting process, including the incubation period, delays in seeking testing, laboratory processing times and reporting delays. Where temporal alignment existed, the direction of the estimated structural change was qualitatively compared with the expected epidemiological effect of the corresponding event. For example, restrictive public health interventions would generally be expected to precede downward shifts in reported incidence, whereas increased population mobility, relaxation of restrictions or the emergence of more transmissible viral variants would be expected to precede upward shifts in infection counts.

The attribution of assignable causes was undertaken solely to provide epidemiological context for the statistically detected changepoints and did not constitute a formal causal analysis. Owing to the observational nature of the data, multiple interventions may have been implemented concurrently, behavioural responses may have occurred gradually, and unmeasured factors may also have influenced transmission dynamics. Consequently, the identified assignable causes should be interpreted as plausible temporal associations rather than definitive evidence of causal relationships. Where multiple candidate events occurred within the specified temporal window, priority was assigned based on epidemiological plausibility, supporting evidence from official surveillance reports and consistency between the expected and observed changes in the infection trend.

3.8 Determining the Average Lag Between Policy Dates and Calendar Change Points in the COVID-19 Infections Trend in Kenya

3.8.1 Overview

The Novel COVID-19 pandemic struck the world in 2019 with the first case tested in Wuhan, China following which the disease quickly spread to many countries in the world. Since the first COVID-19 case was confirmed in Kenya in March 2020, the government of Kenya (GoK) implemented a range of public health and socio-

economic measures, such as restriction of movement by imposing international and inter-county travel bans as well as nation-wide and county-specific curfews. Other measures included the closure of schools and government workplaces, expansion of health insurance, cash and food aids, and tax relief (McDade et al., 2020). However, Kenya's cases and fatalities continued to grow, and the economy remained strained dues to movement restrictions and closure of businesses.

Easing lock down and travel restrictions to help boost economic growth posed a risk of further viral transmission. To ensure the outbreak did not worsen, the government needed to implement strategies for identifying and protecting the most vulnerable, facilitated the establishment of Fangcang shelter hospitals, improved procurement and distribution of testing kits and PPE, and increased financial accountability over the response (McDade et al., 2020).

An interesting question is what guided the National Emergency Response Committee on Coronavirus on the timing of policy implementation including both enforcement and relaxation of various policies? How long did it take after policy implementation before a reversal in infection trend was noted? This study seeks to determine the average lag between policy implementation dates and the calendar changepoints in the COVID-19 infections trend in Kenya. The estimated average lag is then used to obtain a short-term prediction of the next peak in the COVID-19 incidence in response to government policy measures.

3.8.2 Classification of Policies

In epidemiological studies such as the current study, not all government measures have equal impact on transmission patterns. Policies are often classified as major, if they are expected to directly alter population movement, contact rates, or transmission risk, or minor, if they are largely administrative, advisory, or marginal adjustments unlikely to cause immediate shifts in infection trends. This classification helps to reduce noise in the lag analysis, focus changepoint matching on impactful events, and prevent over-counting small or overlapping actions.

In the current study, policy actions were categorized as major or minor based on their scale, degree of behavioral or economic impact, and epidemiological significance (see table 3.3), following the classification logic of the World Health Organization (WHO, 2020), the Oxford COVID-19 Government Response Tracker (Hale et al., 2021), and official policy timelines released by the Ministry of Health (Republic of Kenya, 2020,

2021).

Minor actions under the above classification included: routine public advisories (e.g., mask and hand-washing reminders), small curfew-hour adjustments (e.g., 10 pm to 9 pm), updates to gathering size limits, and regional (county-specific) directives without national impact. For the purposes of lag analysis in this study, both major and some minor COVID-19 policy actions were considered. These included government interventions with nationwide or multi-county impact likely to alter transmission dynamics, such as curfews, lock-downs, or large-scale re-openings. On the other hand, minor actions were excluded because they were unlikely to produce detectable shifts in national infection trends.

Table 3.3: Classification Criteria for COVID-19 Policy Measures in Kenya

Criterion	Explanation	Examples (Kenya)
1. Scale of restriction or relaxation	Nationwide or multi-county measures affecting mobility, gatherings, or operations.	National curfews, lockdown of Nairobi/Mombasa corridor, border closures, lifting of curfews
2. Degree of behavioral or economic impact	Measures expected to change how people work, travel, or socialize.	School closure, bar and restaurant business shutdown, school reopening phases
3. Epidemiological linkage	Actions temporally associated with clear infection trend shifts (change in slope or mean).	Lockdowns, major easing phases, reopening borders

3.8.3 Mathematical Formulation of Lag Analysis and Peak Prediction

In the context of this study, *lag* refers to the time delay between the date of implementation or relaxation of a government policy and the corresponding observable change in the COVID-19 infection trend. The concept of lag allows for an assessment of the temporal relationship between policy actions and subsequent shifts in epidemic dynamics.

3.8.3.1 Definition of Lag

Let $t_k^{(P)}$ denote the calendar date of the kth policy intervention, for $k = 1, 2, \ldots, K$, and let $t_j^{(C)}$ represent the calendar date of the jth estimated changepoint, for $j = 1, 2, \ldots, J$. The lag associated with policy k is defined as

$$L_k = t_{j(k)}^{(C)} - t_k^{(P)}, \tag{3.30}$$

where $t_{j(k)}^{(C)}$ is the changepoint temporally associated with policy k. Positive values of L_k indicate that the changepoint occurred after the policy action, whereas negative values indicate that the changepoint preceded the policy, suggesting a reactive intervention.

3.8.3.2 Matching Policies to Changepoints

To associate each policy intervention with the most relevant changepoint, a forward matching rule was adopted. Specifically, the changepoint corresponding to policy k is selected as

$$j(k) = \arg\min_j \left\{ t_j^{(C)} - t_k^{(P)} \;\middle|\; t_j^{(C)} \geq t_k^{(P)},\ t_j^{(C)} - t_k^{(P)} \leq W \right\}, \tag{3.31}$$

where W is a pre-specified maximum effect window (e.g., 7, 14, or 21 days). Policies with no changepoints occurring within the selected window were recorded as having no detectable short-term effect.

Choice of the Matching Window W

Because the changepoint $t_j^{(C)}$ is estimated with uncertainty, exact one-to-one matching between policy dates and changepoints may be unreliable. To account for this, a temporal matching window of width W days is introduced. A policy date $t_k^{(P)}$ and changepoint $t_j^{(C)}$ are considered associated if

$$\left| t_k^{(P)} - t_j^{(C)} \right| \leq W.$$

The selection of W is guided by both epidemiological and statistical considerations. Empirical studies of COVID-19 interventions frequently report effect windows between 7 and 21 days, corresponding to typical incubation periods, behavioural compliance ad-

justments, and reporting delays (Dehning et al., 2020, Flaxman et al., 2020; and Hsiang et al., 2020). From a statistical perspective, changepoint estimation procedures—particularly likelihood-based approaches—exhibit boundary uncertainty, whereby the estimated location of a true changepoint may deviate by several days (Killick & Eckley, 2014). Classical intervention theory similarly recommends evaluating intervention impacts within a short neighbourhood around the candidate point.
Considering these factors, a window size of $W = 14$ days is adopted in this study. This choice aligns with widely accepted public health delay estimates and accommodates typical statistical uncertainty in changepoint estimation. Sensitivity checks may also be performed by comparing results under alternative window sizes (e.g., $W = 7$ or $W = 21$) to assess robustness.

3.8.3.3 Estimation and Summary of Average Lag

For all policies with matched changepoints, the sample mean lag is computed as

$$\bar{L} = \frac{1}{n} \sum_{k=1}^{n} L_k, \tag{3.32}$$

where n is the number of valid policy–changepoint pairs. The sample variance and sample standard deviation are given by

$$s_L^2 = \frac{1}{n-1} \sum_{k=1}^{n} (L_k - \bar{L})^2, \qquad \text{and} \quad s_L = \sqrt{s_L^2}. \tag{3.33}$$

A 95% confidence interval for the mean lag is expressed as

$$\bar{L} \pm z_{0.975} \left(\frac{s_L}{\sqrt{n}} \right), \tag{3.34}$$

where $z_{0.975}$ denotes the 97.5th percentile of the standard normal distribution.

3.8.3.4 Hypothesis Testing for Lag

To determine whether policy actions are temporally associated with shifts in COVID-19 infection trends, the following hypothesis test was conducted:

$$H_0 : \mu_L = 0 \qquad \text{versus} \qquad H_1 : \mu_L \neq 0,$$

where μ_L is the true mean lag. The test statistic for the one-sample t-test is

$$T = \frac{\bar{L}}{s_L/\sqrt{n}}, \tag{3.35}$$

which is compared against the t-distribution with $n - 1$ degrees of freedom. A statistically significant result suggests that changepoints tend to occur after (or before) policy actions with a non-zero average lag.

3.8.3.5 Lag Between Policy Dates and Epidemic Peaks

Beyond changepoints, the study also examines the temporal distance between policy actions and subsequent infection peaks. Let $t_r^{(Q)}$ denote the calendar date of the rth peak in the smoothed infection curve. The lag between policy k and its associated peak is defined as

$$L_k^{(\text{peak})} = t_{r(k)}^{(Q)} - t_k^{(P)}, \tag{3.36}$$

where $t_{r(k)}^{(Q)}$ is the first peak following policy k within a specified window. The empirical distribution of $\{L_k^{(\text{peak})}\}$ is used to compute the mean and interquartile range of peak lags, which form the basis for short-term predictions.

3.8.3.6 Short-Term Lag Prediction

If a new policy is implemented on date $t_{\text{new}}^{(P)}$, the expected occurrence of the next peak can be predicted as

$$\hat{t}_{\text{next peak}} \approx t_{\text{new}}^{(P)} + \hat{\mu}_L, \tag{3.37}$$

where $\hat{\mu}_L$ is the estimated mean historical lag. A prediction interval may be constructed using the empirical interquartile range of historical lags, thereby providing a plausible window for the timing of the next peak.

This approach provides an interpretable, data-driven method for quantifying and predicting temporal delays between policy interventions and observable changes in epidemic behavior.

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the results of the study and discusses their implications in relation to the research objectives outlined in Chapter One. The chapter is organised into two principal components. The first presents the findings from an extensive simulation study conducted to evaluate the statistical performance of the proposed Hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm (NBMCPA) under different sample sizes and changepoint configurations. The second demonstrates the practical application of the proposed methodology to daily COVID–19 infection data from Kenya, illustrating its ability to identify statistically significant structural changes in a real-world epidemiological setting.

The simulation study evaluates the performance of the proposed algorithm using several criteria, including changepoint estimation accuracy, estimation precision, false detection rates and the influence of sample size and changepoint location on estimation performance. Where appropriate, the observed results are interpreted in light of likelihood theory, properties of maximum likelihood estimation and established principles of changepoint analysis. Comparisons are also made with findings reported in previous studies to demonstrate the consistency of the proposed methodology with existing theoretical and empirical evidence.

The empirical application focuses on detecting multiple changepoints in Kenya's COVID–19 infection time series and examining their temporal correspondence with documented public health interventions and epidemiological developments. In addition, a lag analysis is undertaken to investigate the temporal relationship between policy implementation dates and the estimated changepoints. Throughout the discussion, the findings are interpreted cautiously, recognising that the detected associations provide evidence of temporal correspondence rather than direct causal relationships.

Collectively, the results presented in this chapter provide empirical evidence regarding the statistical performance, practical applicability and robustness of the proposed NBMCPA for analysing structural changes in over-dispersed count data.

4.2 Developing a Multiple Change Point Algorithm Using the Negative Binomial Distribution

This section describes the step-by-step process that was followed in the detection and estimation of multiple change points in a data sequence assumed to come from a Negative Binomial distribution. The NBMCP algorithm consists of 9 steps outlined here-below

Step 1

Input or load the sample data into a statistical software such as R. These data could be either real observed data or simulated data from a Negative Binomial Distribution.

Step 2

Compute the sample mean $\bar{X}$ and variance S^2 for these data. Conduct a dispersion test on the data. If data are equi-dispersed or over-dispersed, proceed to find sample mean and variance under the parameterization discussed in subsection 3.3.2, otherwise discard the sample. Based on the sample statistics, find the method of moments estimates of the two parameters $\hat{r}$ and $\hat{p}$ of the negative binomial model according to equation (3.13)

Step 3

Using the sample data and the parameter estimates in step 2, compute the log-likelihood functions under the null and alternative hypotheses for various values of the arbitrary change point k as described in equations (3.13) and (3.14) respectively.
Next, compute the likelihood ratio statistic $\hat{\Lambda_k}$ as described in equation (3.15) for each possible value of the change point k. Store all computed values of the likelihood ratio statistic in a single vector.

Step 4

Plot a graph of the likelihood ratio statistic $\hat{\Lambda}_k$ or equivalently, its square root, against the possible values of k. A line plot shows at a glance how the likelihood ratio behaves over sequential values of k.

Step 5

Investigate whether a change exists at some point k by inspecting the graph in step 4 visually. As a guide, when there is no change the likelihood ratio statistic does not have a unique maximum point, rather it appears as a regular time series graph with several alternating rises and dips. On the other hand, when a change exists at some point k,

the graph of the likelihood ratio test statistic is such that it reaches a maximum value at exactly or in the neighborhood of the point k.
Figures(4.8) and (4.9) represent illustrative graphs of a likelihood ratio test statistic in the absence of change and in the presence of a single change respectively.

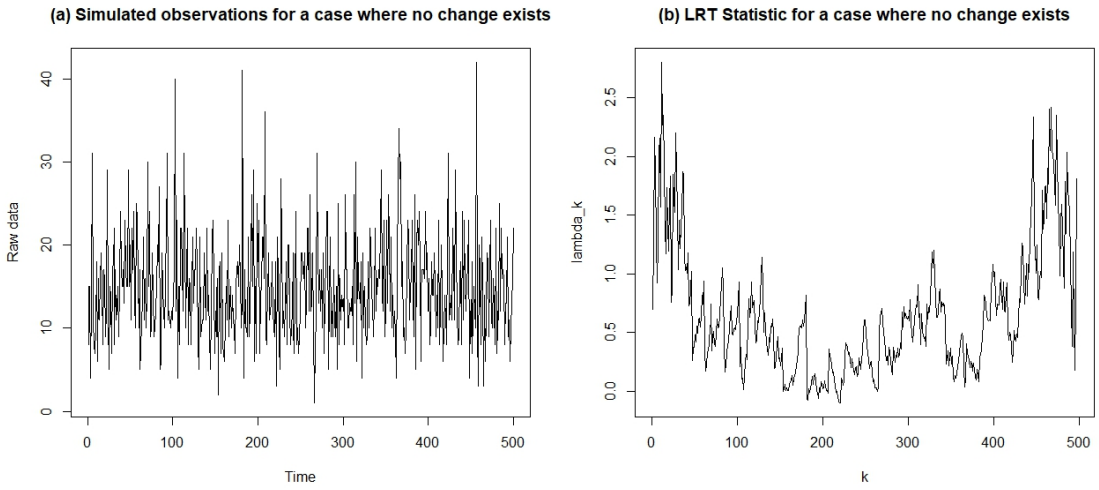

Figure 4.8: Sample Graph of the Likelihood Ration Test Statistic Illustrating a Case of No Change

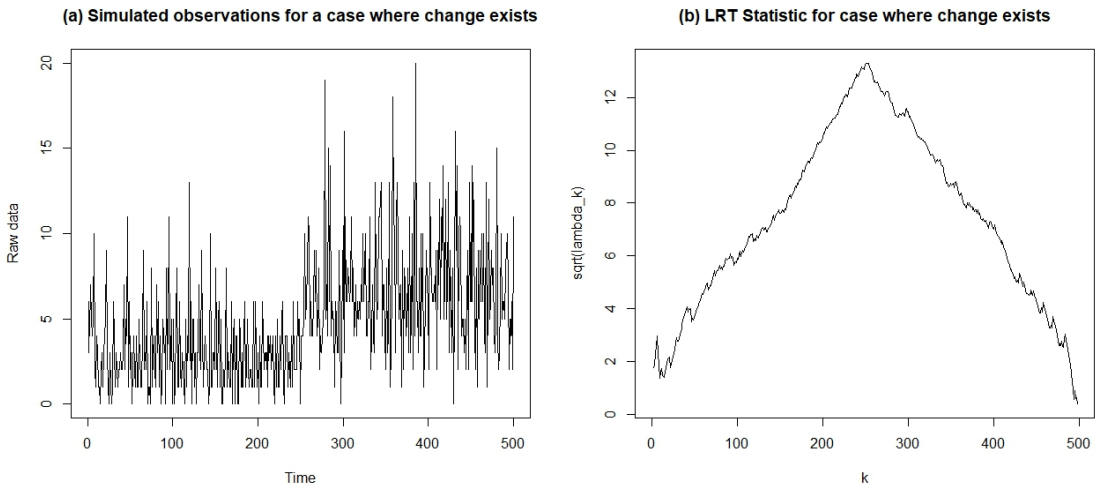

Figure 4.9: Sample Graph of the Likelihood Ratio Test Statistic Showing a Case of a Single Change at Position $n/2$

Step 6

In case there is no change in model parameters for the given dateset, the change point

algorithm comes to an end and no further change points are sought. However, in case a change exists, the next step is to approximate the location of the change point. The change point is estimated using the maximum likelihood approach as the point k at which the likelihood ratio statistic $\hat{\Lambda_k}$ attains a maximum value (see equation 3.16).

Step 7

Once a change has been detected and its location estimated, the next step entails determining whether or not the said change point is statistically significant. This is done by conducting a likelihood ratio test. The computed value of the likelihood ratio test statistic at the said change point is compared against the critical values of the likelihood ratio test statistic which is determined for a given sample size and level of significance as described in section 3.4.

The decision made under the test is such that if the estimated change point lies beyond or exceeds the critical value, the null hypothesis that change is not significant is rejected, otherwise the null hypothesis is not rejected.

Step 8

If a change is found to be statistically insignificant, the algorithm comes to an end and no further change points are sought. However, if a change is found to be significant at some point k, the current value of k is stored as a change point.

The multiple change point algorithm is developed such that only a single change can be identified at a time. In cases where only one change exists in an entire data set, then the algorithm stops when the first and only change point is located. However, in cases where there are two or more change points, the most pronounced change is detected and located first, then the next most pronounced change is identified and so on.

Step 9

Once a significant change point k is located and stored, the next step is to partition the input data set into two segments with the boundary corresponding to the change point estimate k. Each of the two data partitions is then taken, one at a time, and treated as the input data set in step 1 then checked for the existence of change and the process repeats until no further change points are found.

Given this is an iterative process, the estimated values of k at each iteration must be recorded and stored. This process constitutes the step-wise recursive binary segmentation procedure discussed in sub-subsection 3.3.2. Figure 4.10 represents a model flow chart for the Negative Binomial Multiple Change Point Algorithm.

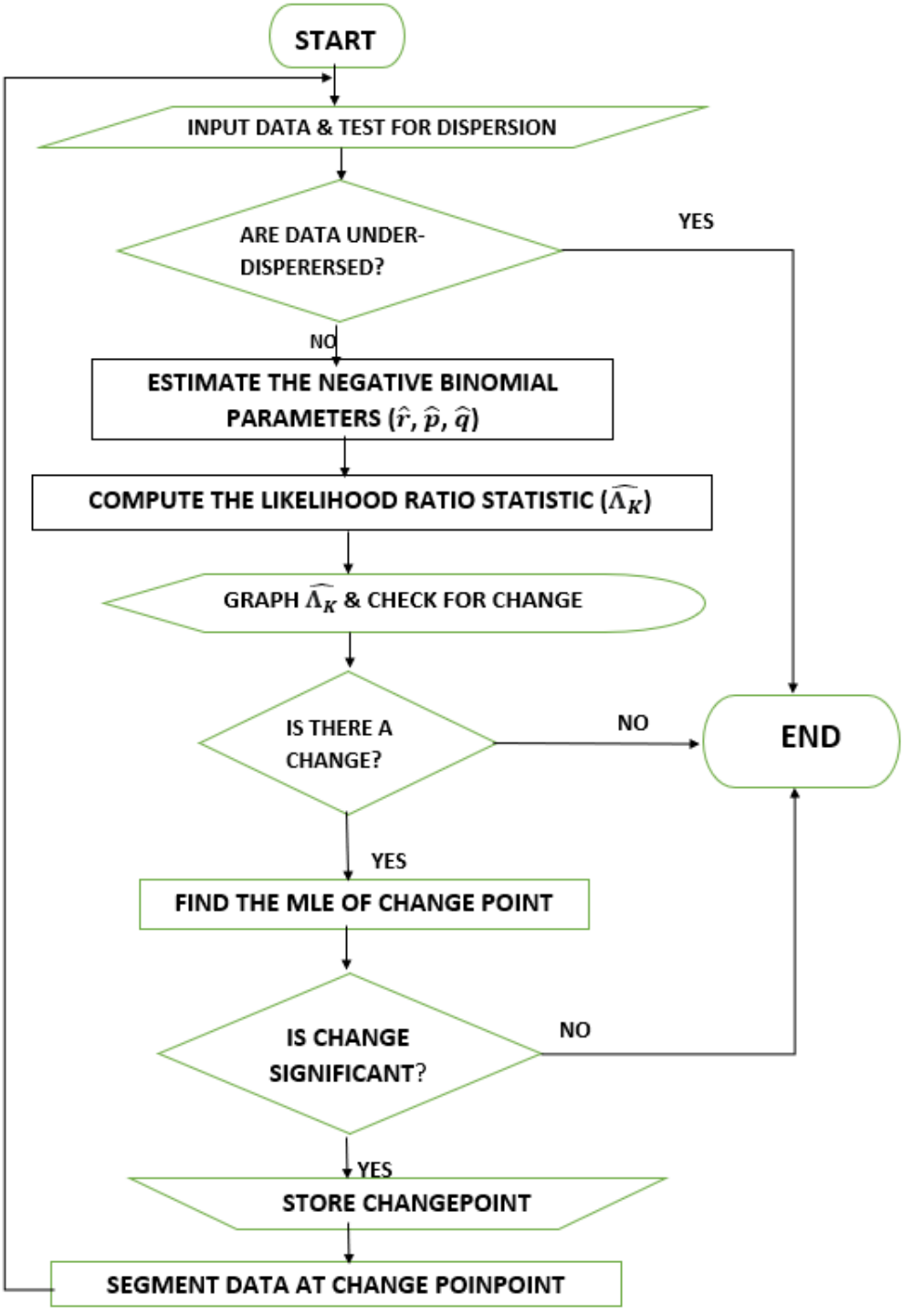

Figure 4.10: Schematic Diagram of the Negative Binomial Multiple Change Point Algorithm

4.3 Determining the Critical Values of the Likelihood Ratio Test for Change

This section refers back to the methodology described in chapter three section 3.4. The method proposed by Gombay and Horvath (1996) on the asymptotic distribution of the likelihood ratio test is applied in the determination of the critical values presented according to equation 3.19 which states that for all $T > 0$

$$P\left\{\sup_{0\leq t\leq T}\Delta(t) > x\right\} = \frac{x^d \exp(-x^2/2)}{2^{d/2}\Gamma(d/2)}\left\{T - \frac{d}{x^2}T + \frac{4}{x^2} + O\left(\frac{1}{x^4}\right)\right\} \tag{4.38}$$

In the method, the choice of values of $h(n)$ and $l(n)$ should be such that $h(n) \leq \frac{1}{n}$ and $l(n) \geq \frac{1}{n}$ and $0 < h(n) < 1 - l(n) < 1$. Gombay and Horvath (1996) used the formulations:

$$h(n) = l(n) = \frac{(\log n)^{(3/2)}}{n}$$

Various combinations of $h(n)$ and $l(n)$ were tested for the current study, ensuring the condition was met. The choices of h and l settled upon in this study were given by:

$$h(n) = l(n) = \frac{(\log n)^{(2)}}{n}$$

The parameter d in equation 4.38 was taken as $d = 2$ to equal the number of unknown parameters in the Negative Binomial model. The value of parameter T was dependent on the choice of h and l as:

$$T = \log(\frac{(1-h)(1-l)}{hl})$$

The asymptotic critical values for $\sqrt{Z_n}$ were determined in the R environment for various sample sizes n and different levels of significance α as summarized in table 4.4. Figure 4.11 shows the behaviour of the critical values for the LRT as the sample size increases and the level of significance of the test changes between $\alpha = 1\%, 5\%$ and 10%.

Table 4.4: Critical Values for the Likelihood Ratio Test for Existence of Change

Sample size (n)	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
12	2.900	3.184	3.735
20	3.019	3.294	3.830
50	3.183	3.443	3.958
100	3.275	3.527	4.029
200	3.349	3.594	4.086
500	3.428	3.666	4.128

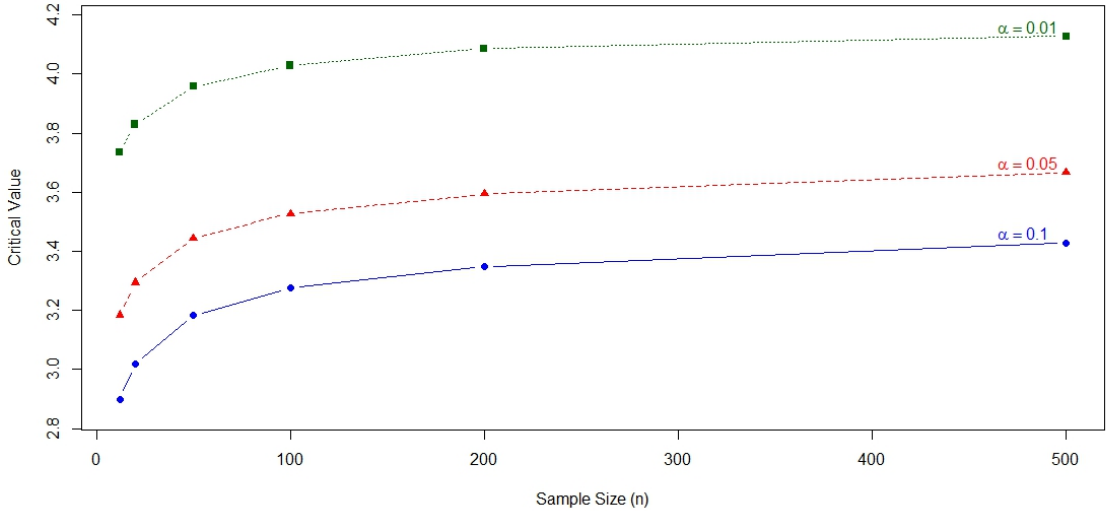

Figure 4.11: Behavior of the Critical Values for the LRT as n Varies for Various Test Sizes α

4.4 Testing the Negative Binomial Multiple Change Point Algorithm using Simulated Data

The change point algorithm developed in section 4.1 was validated using simulated data from a Negative Binomial distribution. This section gives a summary of the results of the simulation study performed in R using different sample sizes and varying model parameters where applicable. The first simulation (see subsection 4.2.1) involved change detection in a scenario where the data set had no change in the parameters r and p of the Negative Binomial distribution. As such the model parameters were both kept constant.
The second, third and fourth simulations (see subsection 4.2.2) were performed in a setting where there was a single change in the both model parameters for varying sample sizes.
The fifth and subsequent simulations (see subsection 4.2.3) were done to showcase a scenario where there were multiple simultaneous changes in both model parameters at various predefined points. For simplicity in illustration only two changes were considered for the multiple change-point study.

4.4.1 A case where No Change Exists

Data simulation was done using the Negative Binomial distribution with no predefined change point for a sample size, n. The entire data set was such that both parameters of the distribution remained constant throughout the time series. Figure 4.12 shows the graph of $\sqrt{\Lambda_k}$ for a sample of size $n = 500$. The graph of the likelihood ratio test statistic (see figure 4.12 panel (b)) does not exhibit a unique or single maxima. Instead, the graph appears rugged, similar to the graph of raw data as displayed in panel (a). This indicates that no change is detected.
Further, in addition to the lack of a single maxima on the graph, the absence of variation in the distribution parameters is emphasized by the fact that the highest point on the graph of the LRT statistic lies below the critical value, $C = 3.666$ $\alpha = 0.05$ (see the dashed horizontal line). It is concluded that at the 5% level of significance, there is no statistically significant change in the parameters of the distribution.

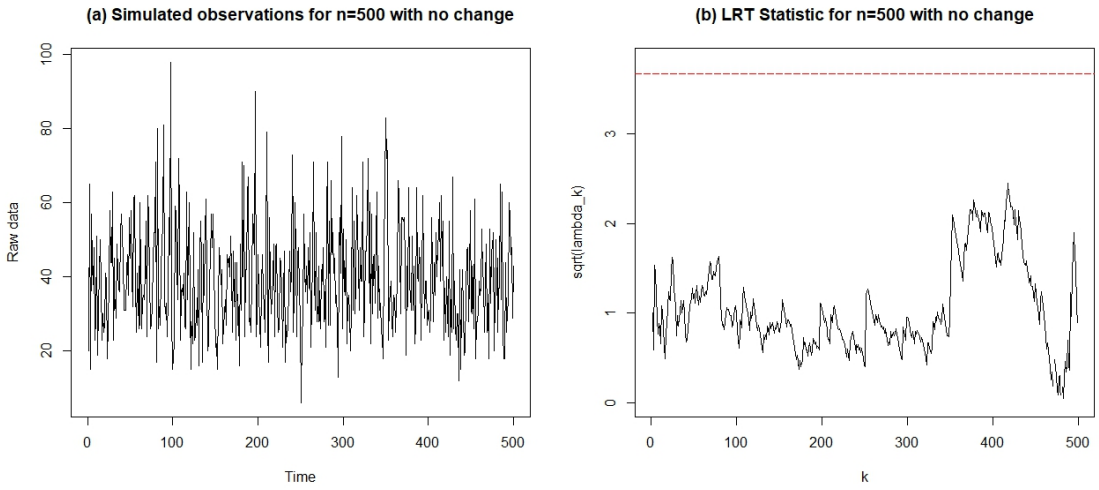

Figure 4.12: A Case of No Change for $n = 500$, at $\alpha = 0.05$

4.4.2 A Case where a Single Predefined Change Exists

Data simulations were done in R using the Negative Binomial distribution with 3 preset change points for different small and medium sample sizes, n. The change points were set such that both the parameters of the Negative Binomial distribution changed once in the entire series at either of the locations $n/4$, $n/2$ or $3n/4$. The change point estimates were obtained using the NBMCP algorithm and the results summarized in table 1. The graphs of the square root of the LRT statistics, are shown in figures 4.13 to 4.24.

4.4.2.1 When the Sample Size is $n = 50$

Figures (4.13), (4.14) and (4.15) represent the raw simulated data (panel a)) and the results of the likelihood ratio test (panel (b)) when the change point was set at $n/4$, $n/2$ and $3n/4$ respectively. The estimated change point in each case was indicated on the graph of the likelihood ratios statistic $\sqrt{\Lambda_k}$ using a vertical line. The horizontal dashed lines showed the critical values of the test which was used to determine whether or not a change, if present, was significant. The results displayed in Figure 4.13 showed that when the change point was set at position $n/4 = 12.5$, the MLE of the change point was $\hat{k} = 13$. On the other hand, when the change was set at the middle of the series, $n/2 = 25$, (see figure 4.14) the change point estimate was $\hat{k} = 25$. When the change point was set further away from the initial data point, at position $3n/4 = 37.5$ as shown in figure 4.15, the change point estimate was $\hat{k} = 38$.

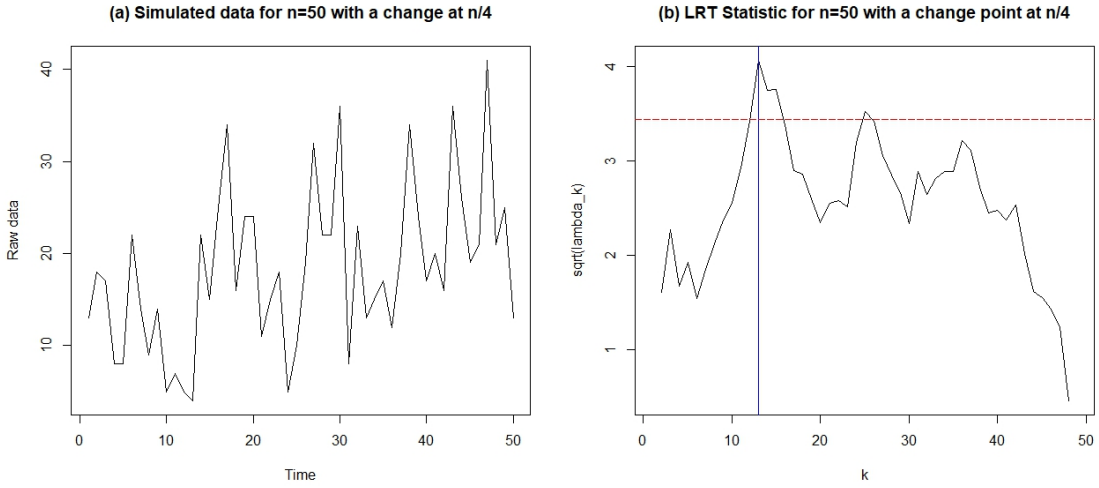

Figure 4.13: A Case of a Single Change at $n/4$ for $n = 50$, at $\alpha = 0.05$

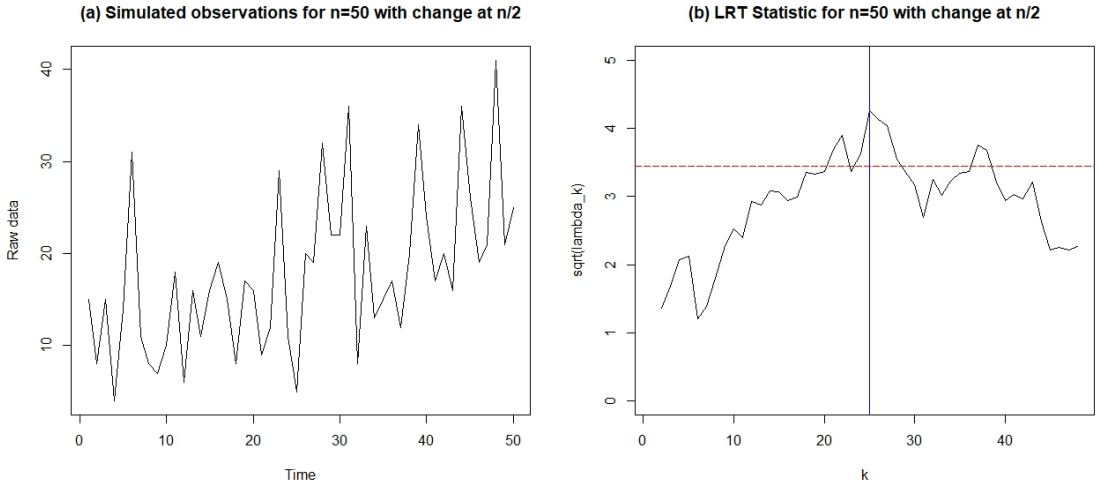

Figure 4.14: A Case of a Single Change at $n/2$ for $n = 50$, at $\alpha = 0.05$

4.4.2.2 When the Sample Size is $n = 100$

For the case where $n = 100$, the changepoint locations at $n/4 = 25$, $n/2 = 50$, and $3n/4 = 75$ were considered. The results indicated that the estimator performed very well across all positions. Specifically, when the changepoint was set at $k = 25$, the estimated changepoint was $\hat{k} = 24$, resulting in a small estimation error of $\Delta = 1$. At the midpoint, $k = 50$, the changepoint was estimated exactly, with $\hat{k} = 50$ and no error. Similarly, for $k = 75$, the estimator again performed perfectly, yielding $\hat{k} = 75$. These results suggest that the likelihood ratio test is highly accurate for moderate sample sizes,

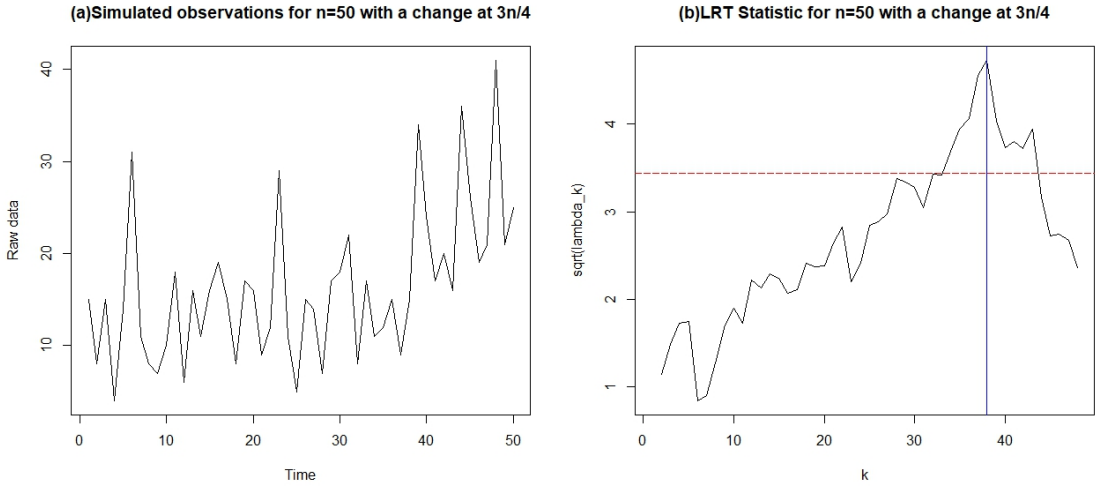

Figure 4.15: A Case of a Single Change at $3n/4$ **for** $n = 50$**, at** $\alpha = 0.05$

with negligible bias and excellent localization of the changepoint, particularly when the changepoint lies in the interior of the series. Figures (4.16), (4.17) and (4.18) represent the raw simulated data and the results of the likelihood ratio test when the change point was set at $n/4$, $n/2$ and $3n/4$ respectively. The estimated change point in each case was indicated on the graph of the likelihood ratios (panel (b)) using vertical lines.

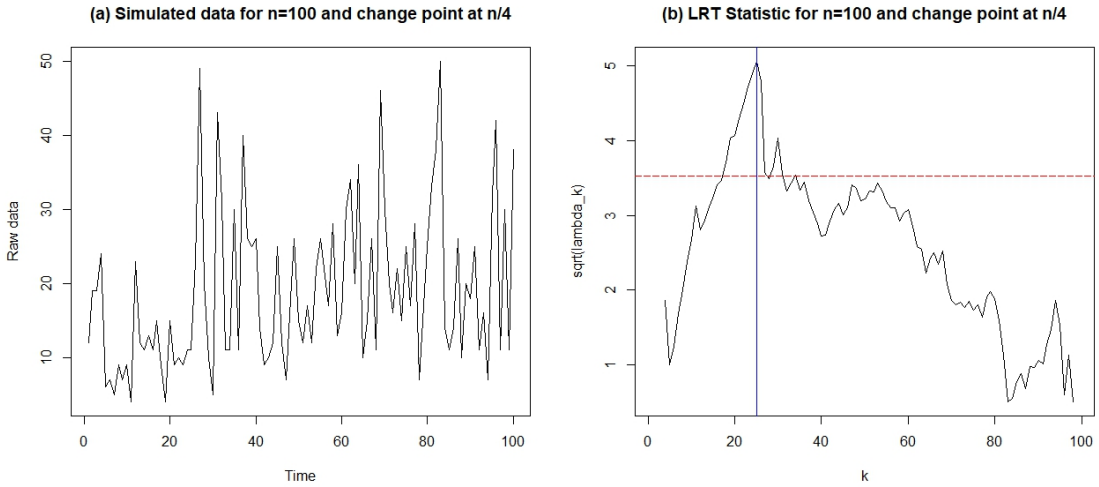

Figure 4.16: A Case of a Single Change at $n/4$ **for** $n = 100$

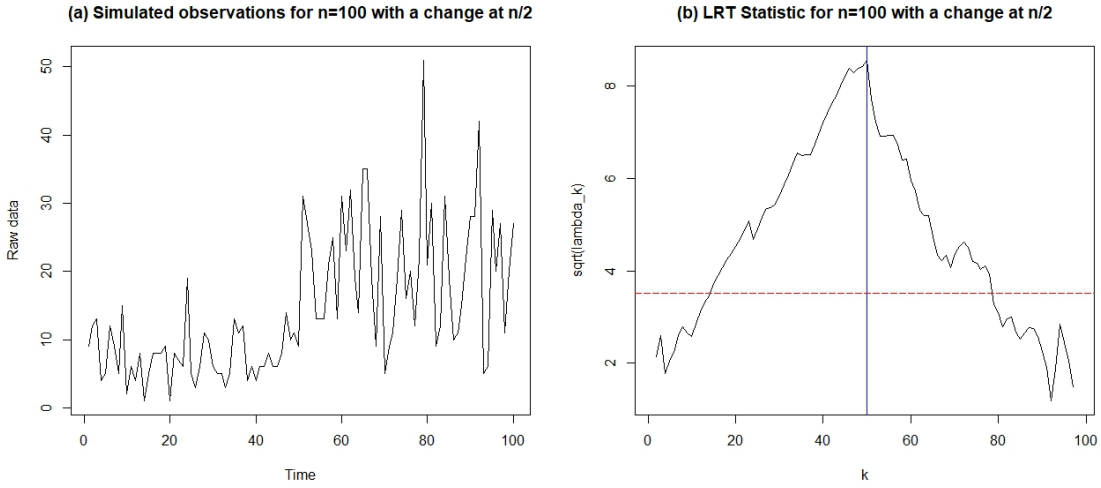

Figure 4.17: A Case of a Single Change at $n/2$ for $n = 100$

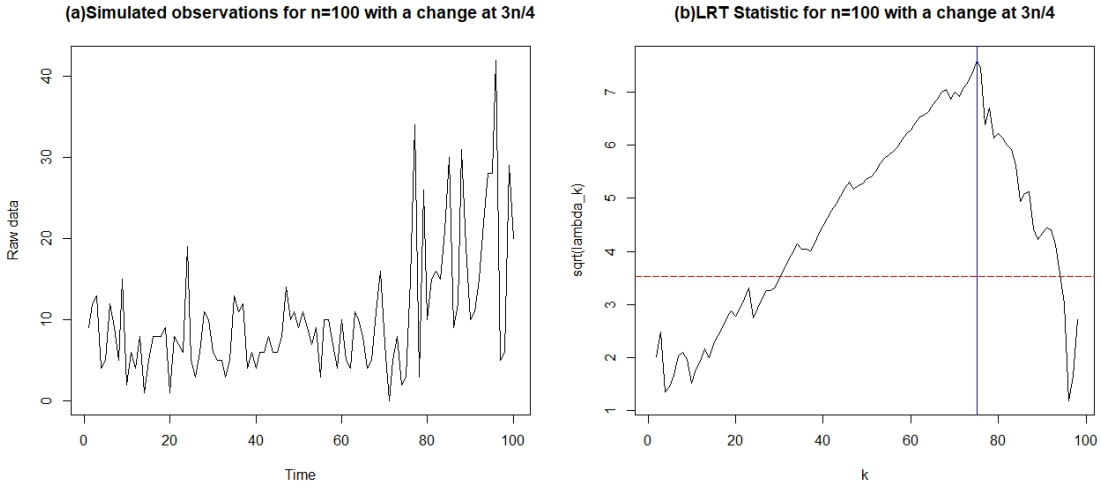

Figure 4.18: A Case of a Single Change at $3n/4$ for $n = 100$

4.4.2.3 When the Sample Size is $n = 200$

For $n = 200$, the actual changepoint locations were $k = 50$, $k = 100$, and $k = 150$. The estimator demonstrated even stronger performance, with exact recovery of the changepoint in all cases. That is, $\hat{k} = 50$, $\hat{k} = 100$, and $\hat{k} = 150$, yielding zero estimation error in all three scenarios. This finding indicated that as the sample size increases, the likelihood ratio test becomes increasingly precise, and the estimator achieves perfect localization of the changepoint. The results highlighted the asymptotic consistency of the estimator, as the accuracy improved with increasing sample size.

Figures 4.19, 4.20 and 4.21 represent the raw simulated data and the results of the likelihood ratio test when the change point was set at $n/4$, $n/2$ and $3n/4$ respectively for $n = 200$. The estimated change point in each case was indicated on the graph of the likelihood ratios (panel (b)) using vertical lines.

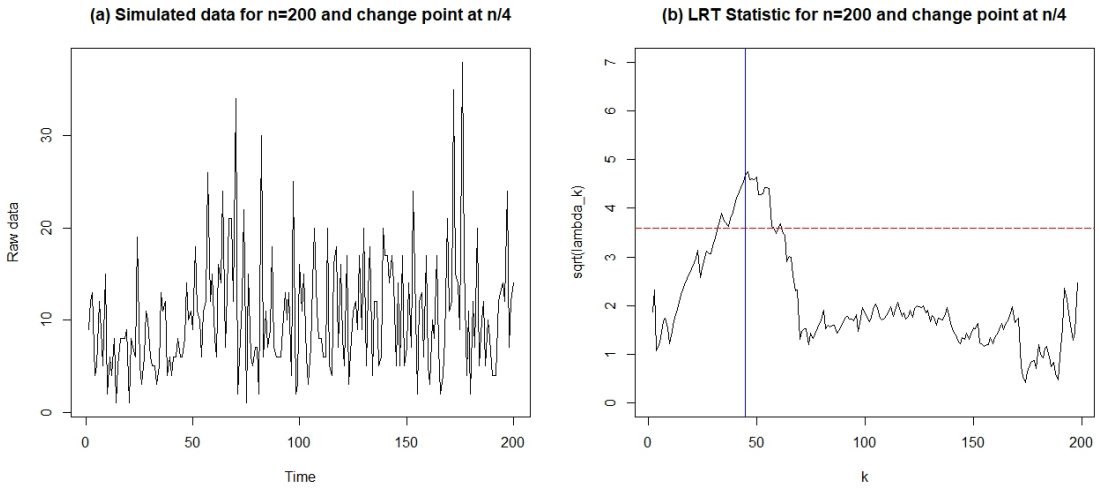

Figure 4.19: A Case of a Single Change at $n/4$ for $n = 200$

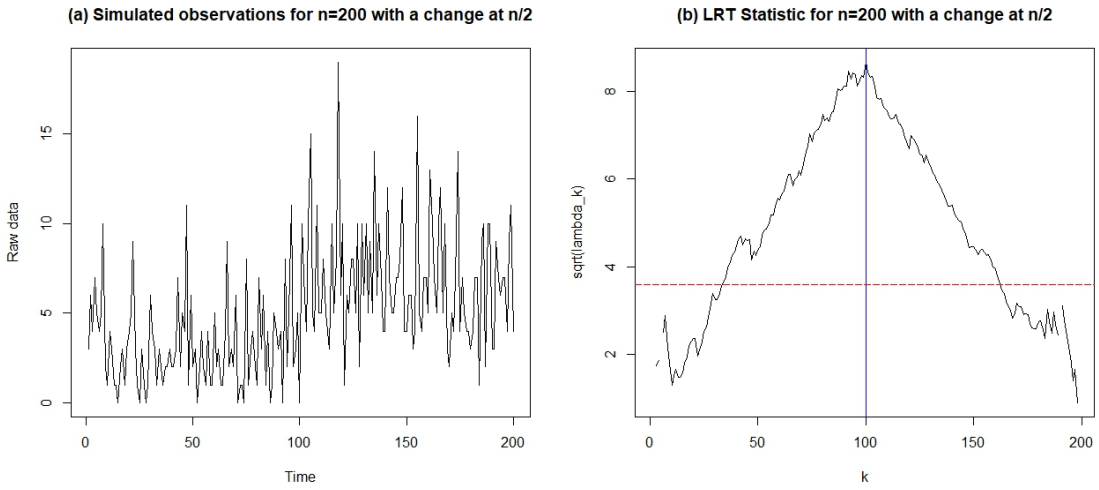

Figure 4.20: Graph showing a case of a single change at $n/2$ for a sample size $n = 200$

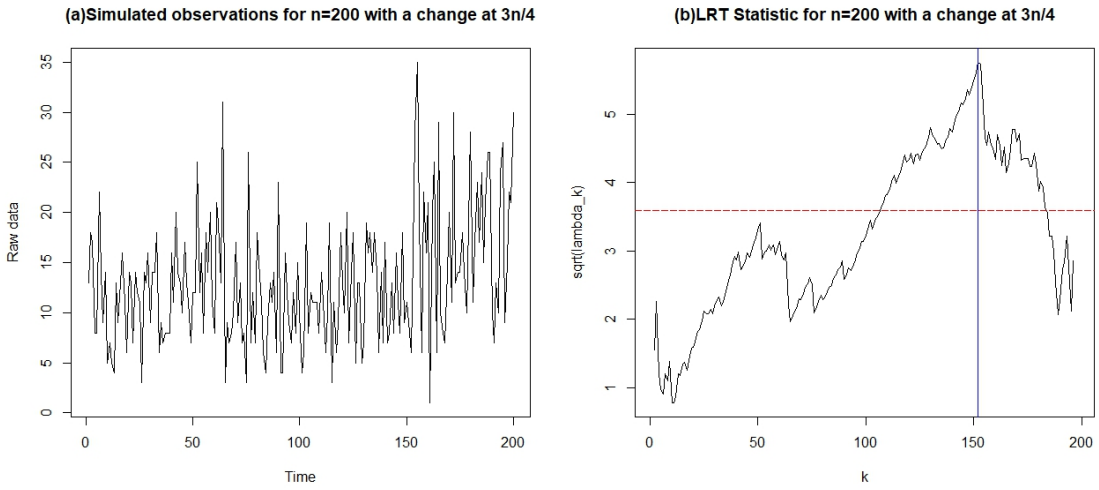

Figure 4.21: A Case of a Single Change at $3n/4$ **for** $n = 200$

4.4.2.4 When the Sample Size is $n = 500$

When the sample size was further increased to $n = 500$, the results remained consistent with the findings at $n = 200$. The changepoint locations $k = 125$ and $k = 250$ (corresponding to $n/4$ and $n/2$, respectively) were both estimated exactly, with $\hat{k} = 125$ and $\hat{k} = 250$, resulting in zero estimation error. These findings reinforced the robustness of the method for large samples, demonstrating that the likelihood ratio test not only detected the presence of a changepoint effectively, but also accurately identified its location with no observable bias in estimation. Figures (4.22), (4.23) and (4.24) represent the raw simulated data and the results of the likelihood ratio test when the change point is set at $n/4$, $n/2$ and $3n/4$ respectively for $n = 500$. The estimated change point in each case was indicated on the graph of the likelihood ratio statistic (panel(b)) using vertical lines. The horizontal dashed lines indicated the critical values of the test statistic at the 5% level of significance.

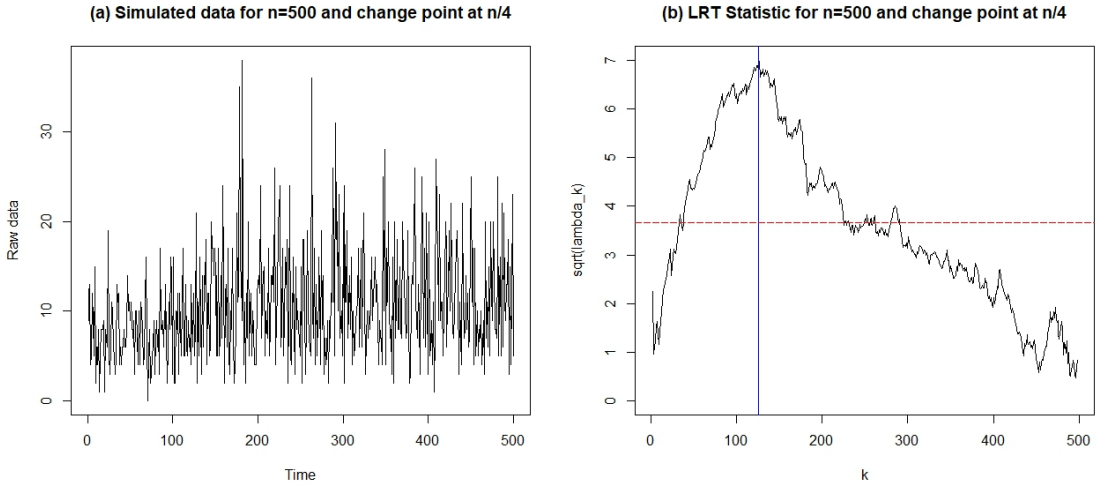

Figure 4.22: A Case of a Single Change at $n/4$ for $n = 500$, at $\alpha = 5\%$

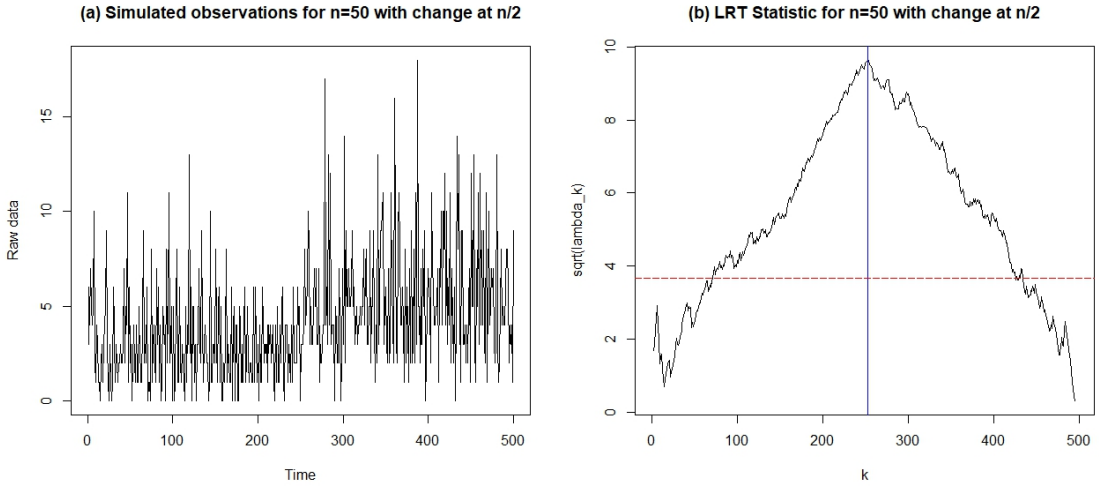

Figure 4.23: A Case of a Single Change at $n/2$ for $n = 500$, at $\alpha = 5\%$

4.4.3 A Case where Multiple Predefined Changes Exist

The NBMCP algorithm was found to work well for the single change point case as indicated by the results in subsection 4.3.3.1. The same algorithm was tested through a simulation study for a multiple change point case with a sample of size $n = 500$. For simplicity, two change points were specified quarter-way (at $n/4 = 125$) and three-quarter-way (at $3n/4 = 375$) respectively in the simulated data.

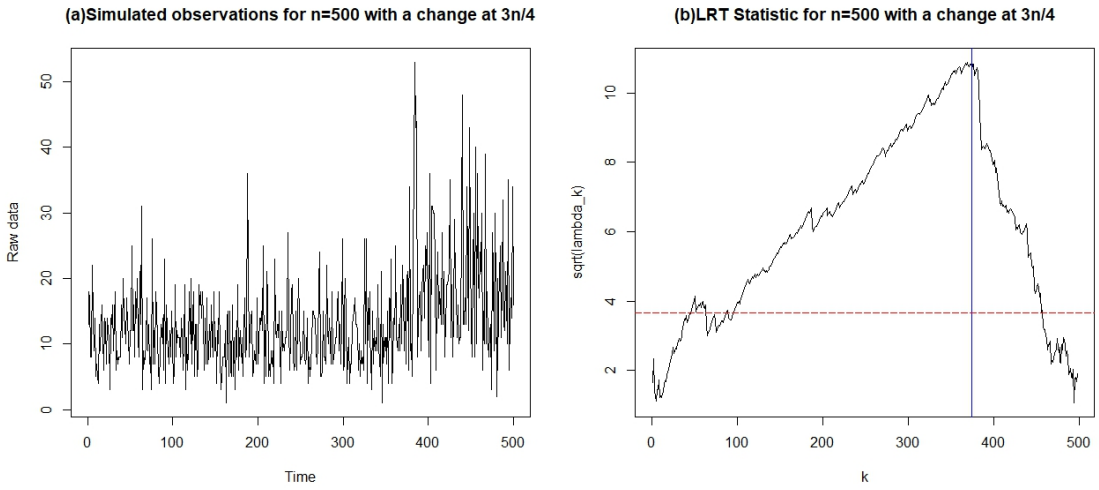

Figure 4.24: A Case of a Single Change at $3n/4$ **for** $n = 500$**, at** $\alpha = 5\%$

4.4.3.1 Location of First Change Point

The NBMCP algorithm detects changes by the order of their magnitude such that the first change point detected and estimated is the most pronounced among all changes present. The first change point lies at the point where the likelihood ratio test statistic first attains a maximum value.

The vertical blocked line in figure (4.25) shows the estimate of the first change point, which corresponds to the $3n/4$ value. Other potential change points only appear as peaks on the graph of the likelihood ratio statistic but are not marked as change points at this stage.

4.4.3.2 Location of Second Change Point

Once the first change point is identified, the algorithm divides the time series into two parts separated by the first estimated change point. The NBMCP algorithm proceeds to detect and locate a change point in the lower partition and then in the upper partition of the data set. The lower partition ranges from the first data point in the series to the first change point whereas the upper partition ranges from the first change point to the last data point in the series.

Figure(4.26) shows estimates of the first change point (vertical blocked line) and the second change point (vertical dashed line).

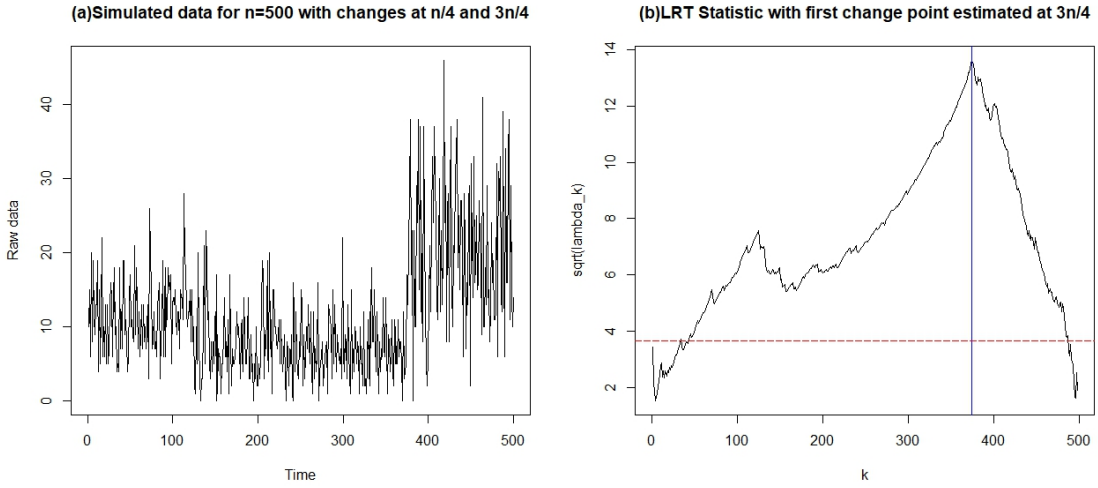

Figure 4.25: Simulated Observations (a) and Likelihood Ratio Curve (b) Showing the Location of First Change Point at $3n/4$ for $n = 500$, at $\alpha = 5\%$

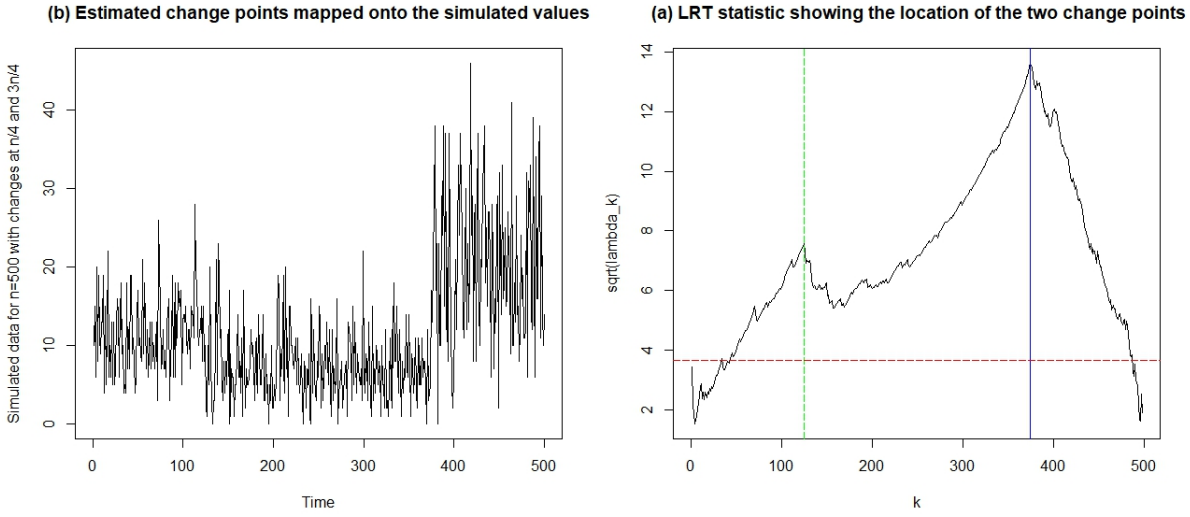

Figure 4.26: Location of the First Changepoint at $3n/4$ and the Second Changepoint at $n/4$ for $n = 500$, at $\alpha = 5\%$

4.4.3.3 Location of the Third and Subsequent Change Points

The detection and estimation of the first two change points at the positions $n/4 = 250$ and $3n/4 = 750$ results in three different partitions of the original simulated data set of size $n = 1000$. The boundaries of these partitions are located at the two estimated change points. The graph of simulated observations and likelihood ratio (see figure 4.27) indicates that there are no changes in the lower (consisting of the first set of observations 0-250), middle (consisting of the data points in positions 251-750) and

upper (consisting of the observations 751-1000) sub-partitions of the data set. Therefore no third, fourth, or higher change points exist in this simulated data set. The foregoing statement can be confirmed by making an assumption that there may be a third change point existing in the lower partition of the data set. The NBMCP algorithm is applied to the first 250 observations. As expected, the graph of the likelihood ratio for this sub-partition (see figure 4.27) does not show a unique or single maxima. This indicates that no change point is detected between the 1^{st} and 250^{th} observations.

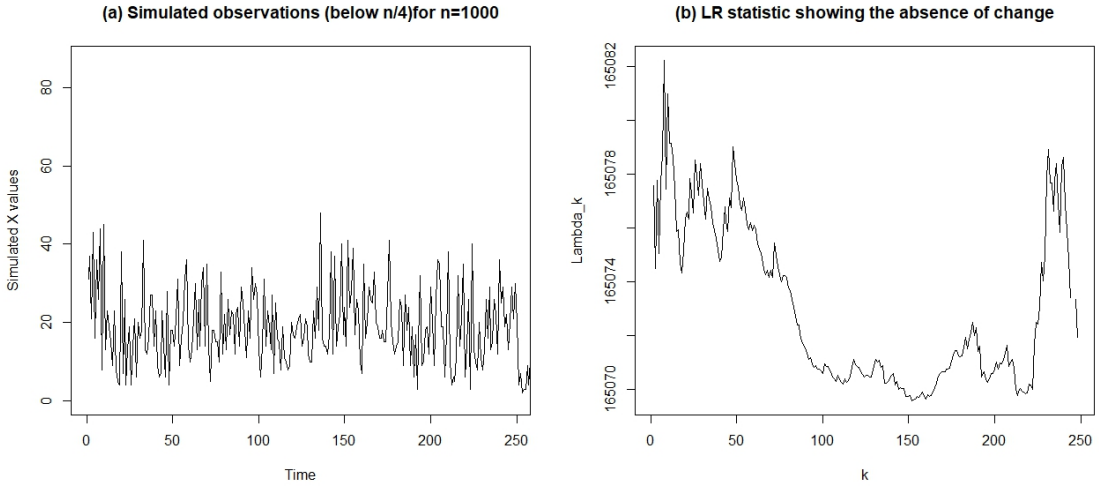

Figure 4.27: Graph of Likelihood Ratio Test Statistic Showing the Absence of Change in the Lower Partition for $n = 1000$

The NBMCP algorithm is applied to the middle and upper sub-partitions of the data set to check for any additional change points that may exist. Figures (4.28) and (4.29) represent the graphs of the likelihood ratio statistics which indicate that there are no changes detected between the 251^{th} and 750^{th} observations as well as between the 751^{th} and 1000^{th} observations.

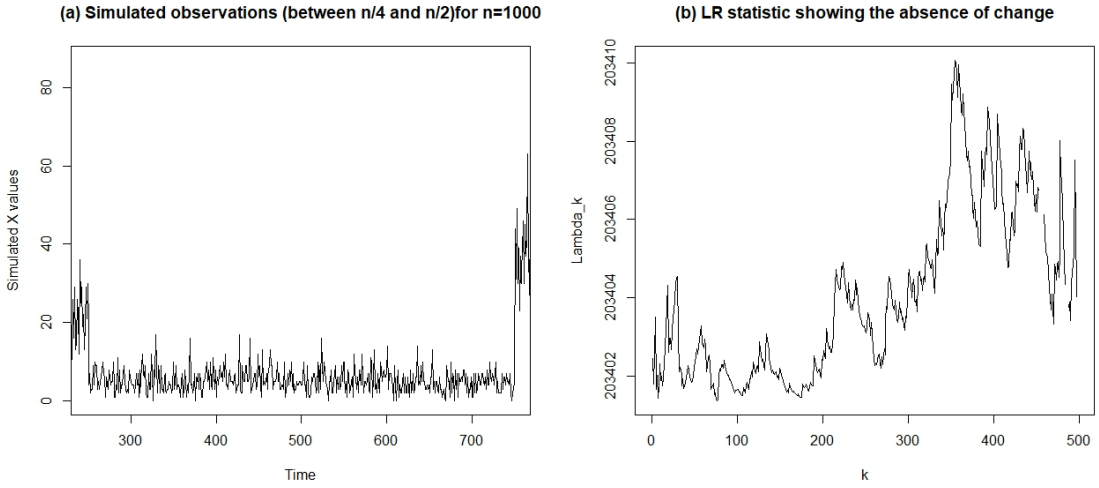

Figure 4.28: Graph of Likelihood Ratio Test Statistic Showing the Absence of Change in the Middle Partition for $n = 1000$

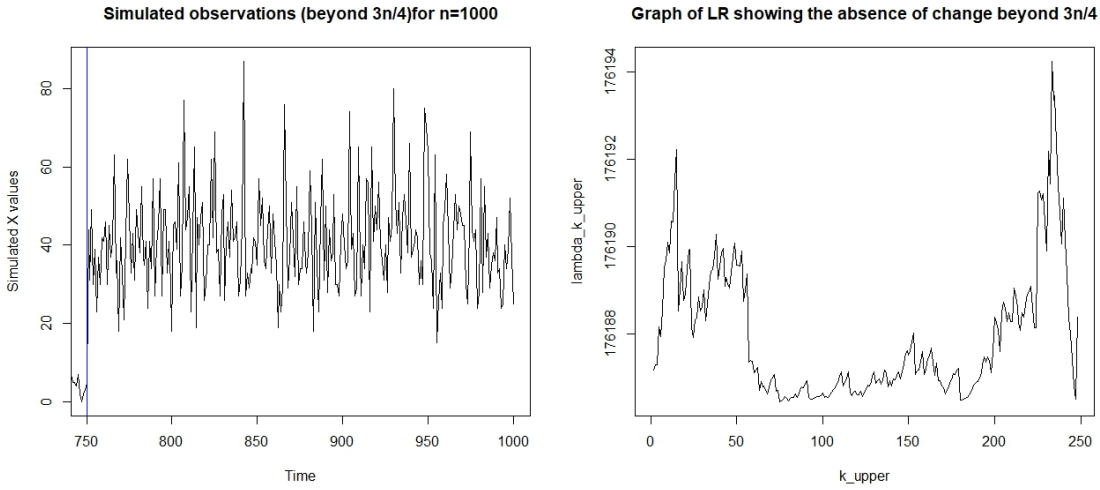

Figure 4.29: Graph of the Likelihood Ratio Statistic Showing the Absence of Change in the Upper Sub-partition for $n = 1000$

Given that no further changes are detected in the lower, middle and upper sub-partitions of the data set, there is no need to proceed with the change detection and location procedure. The algorithm comes to a halt and only two change points are reported. Table 4.5 gives a summary of the estimated change points for the different sample sizes and true change locations considered for the foregoing single change-point scenario. The estimation error was calculated as the difference between the estimated change point and the true changepoint. The results showed that the algorithm detects and

locates changes with very little ($\Delta = \pm 1$) or no ($\Delta = 0$) error. Changes located further away from the initial data point were more accurately estimated for small sample sizes.

Table 4.5: Actual Versus Estimated Changepoints Across Sample Sizes and Change Locations, for a Single-Change Setting

Sample size (n)	**Position of change**	**Actual cpt** (k)	**Estimated cpt** ($\hat{k}$)	**Error** (Δ)
$n = 12$	$n/4$	3	2	1
	$n/2$	6	5	1
	$3n/4$	9	9	0
$n = 20$	$n/4$	5	4	1
	$n/2$	10	9	1
	$3n/4$	15	15	0
$n = 60$	$n/4$	15	14	1
	$n/2$	30	30	0
	$3n/4$	45	45	0
$n = 100$	$n/4$	25	24	1
	$n/2$	50	50	0
	$3n/4$	75	75	0
$n = 200$	$n/4$	50	50	0
	$n/2$	100	100	0
	$3n/4$	150	150	0
$n = 500$	$n/4$	125	125	0
	$n/2$	250	250	0
	$3n/4$	375	375	0

4.5 Algorithm Diagnostics

This section gives a summary of results for the simulation study and visual analysis showcasing performance metrics of the NBMCPA given different sample sizes, and location of change points. For each sample size considered, the true changepoints are set at positions $n/4$, $n/2$ and $3n/4$. The distribution of estimated change points is visualized using histograms and plots of the kernel density estimates. The vertical red dashed lines mark the position of the true change points on the kernel density curves.

4.5.1 When the Sample Size is $n = 20$

Figures (4.30), (4.31) and (4.32) show the distribution of the estimated change points for a small sample of size $n = 20$. From the histogram, the estimated change points are scattered at different position with a slight concentration around the true change point for cases where the true changes are located at position $n/4$ and $3n/4$. The data distributions displayed on the histograms in figures 4.30 and 4.32 reveal that the algorithm has a lower accuracy and precision for small sample sizes when the true changes are located near the endpoints of the dataset. The KDE curves show two relatively broad peaks with the higher peaks showing a higher concentration of the estimates around the true change points. The lower peaks an indication of false alarms in change detection. On the other hand, when the true change point is located at the middle of the dataset (at $n/2$), the histogram shows a peak at around the middle position, with fewer estimates in regions away from $n/2$. In addition, the KDE curve has a single peak near, but not coinciding with, the true change location. This result indicates that the model accuracy and precision is higher for changes located at $n/2$ compared to when changes are located at $n/4 and 3n/4$. However, there is a notable misclassification rate evidenced by the departure of the peak of the KDE curve form the true change point.

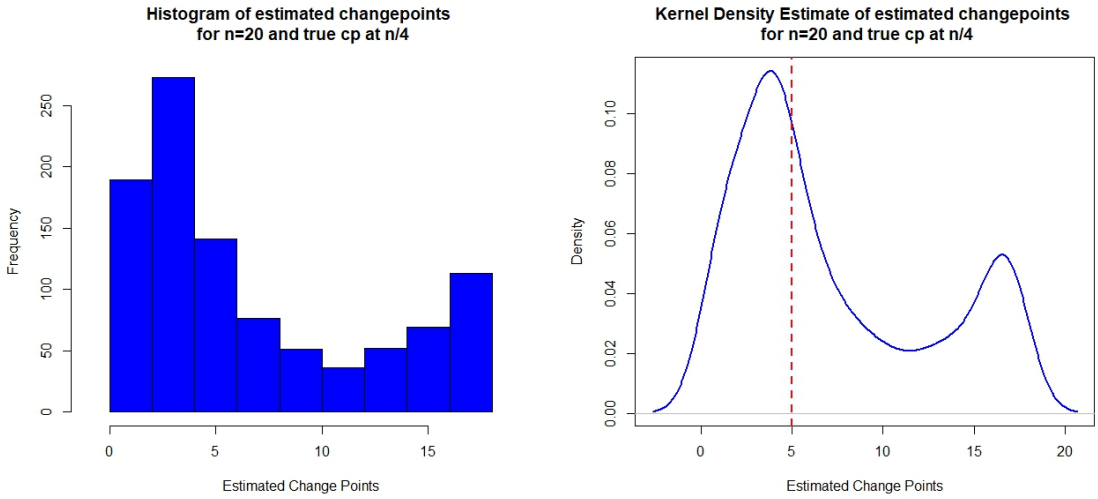

Figure 4.30: Distribution of Estimated Change Points for a Single True Change at $n/4$ for $n = 20$

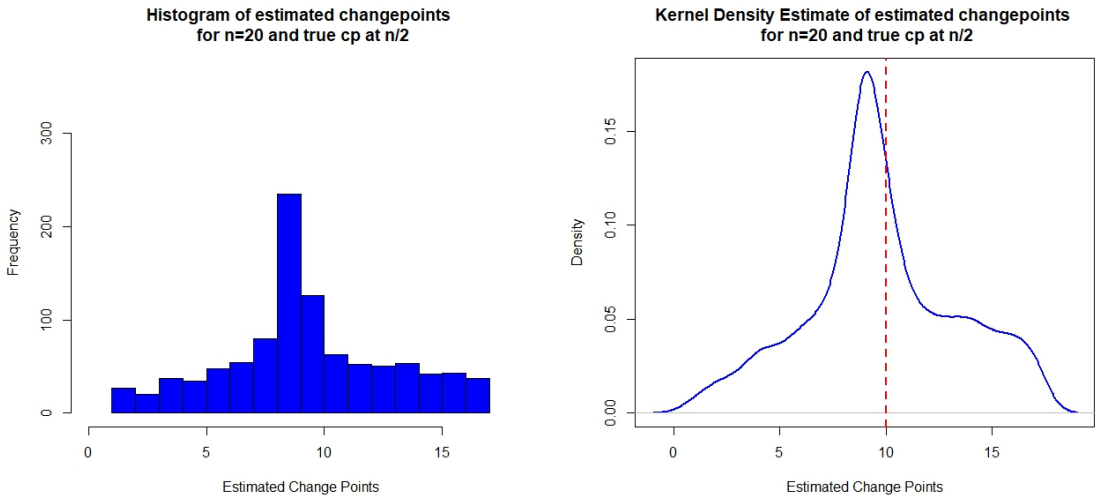

Figure 4.31: Distribution of Estimated Change Points for a Single True Change at $n/2$ for $n = 20$

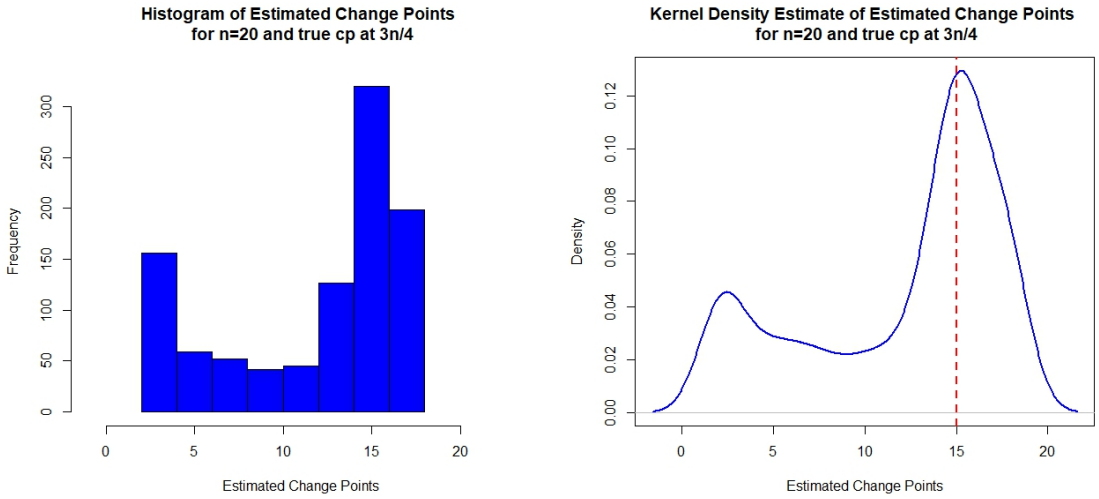

Figure 4.32: Distribution of Estimated Change Points for a Case of a Single True Change at $3n/4$ for $n = 20$

4.5.2 When the Sample Size is $n = 200$

Figures (4.33), (4.34) and (4.35) show the distribution of the estimated change points for a sample of size $n = 200$. The histogram shows that estimated change points are concentrated around the true change positions for all three locations, indicating a high precision and accuracy of the model across all positions of actual change points.This finding is further supported by the KDE curve which has a narrow peak in all three cases indicating that there is no significant bias or variability in the model's estimates for the given sample size. However, the distribution of estimated change points has longer tails to the right (for true change and $n/4$) and to the left (for true change and $3n/4$) pointing to the presence of false alarms and lower precision in the change point detection before and after the true change points respectively.

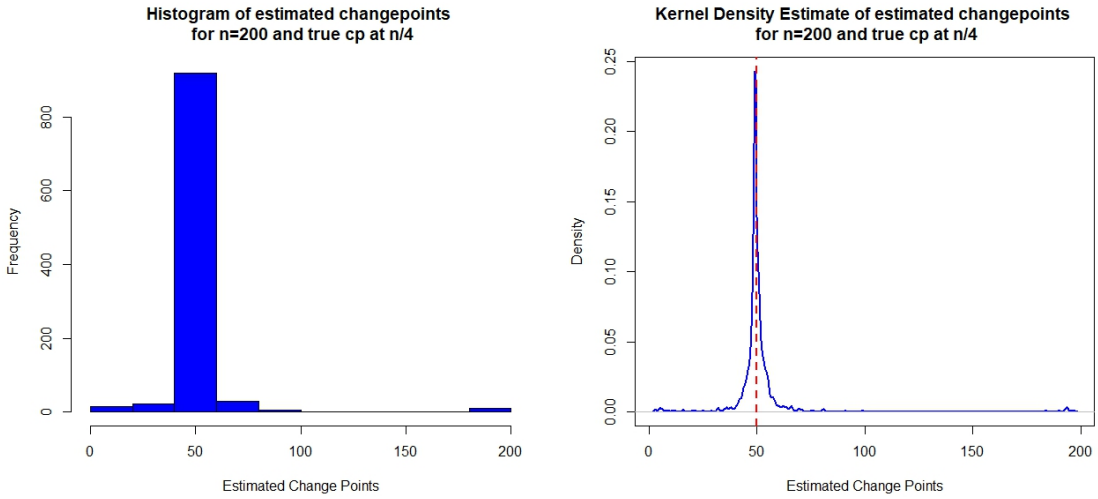

Figure 4.33: Distribution of Estimated Changepoints for a Case of a Single True Change at $n/4$ for $n = 200$

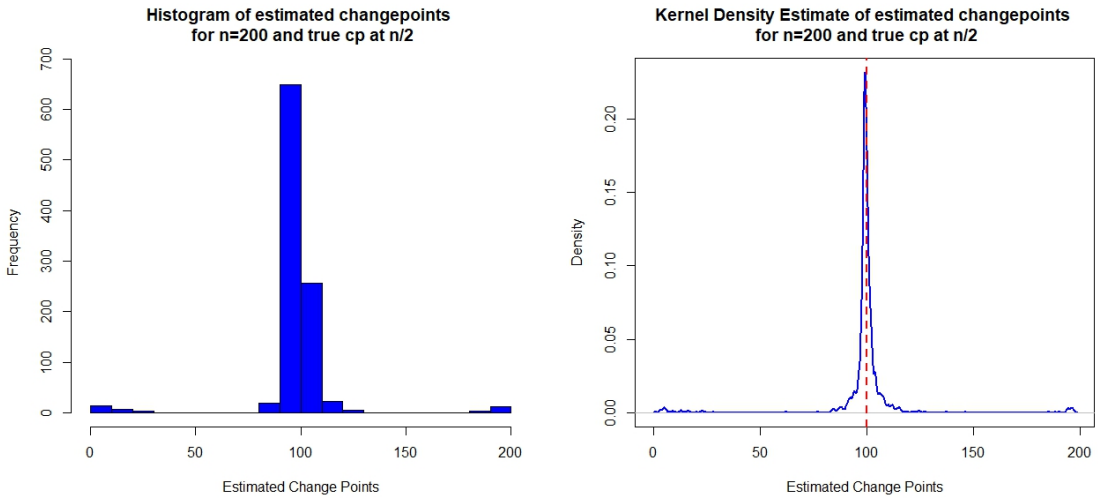

Figure 4.34: Distribution of Estimated Changepoints for a Case of a Single True Change at $n/2$ for $n = 200$

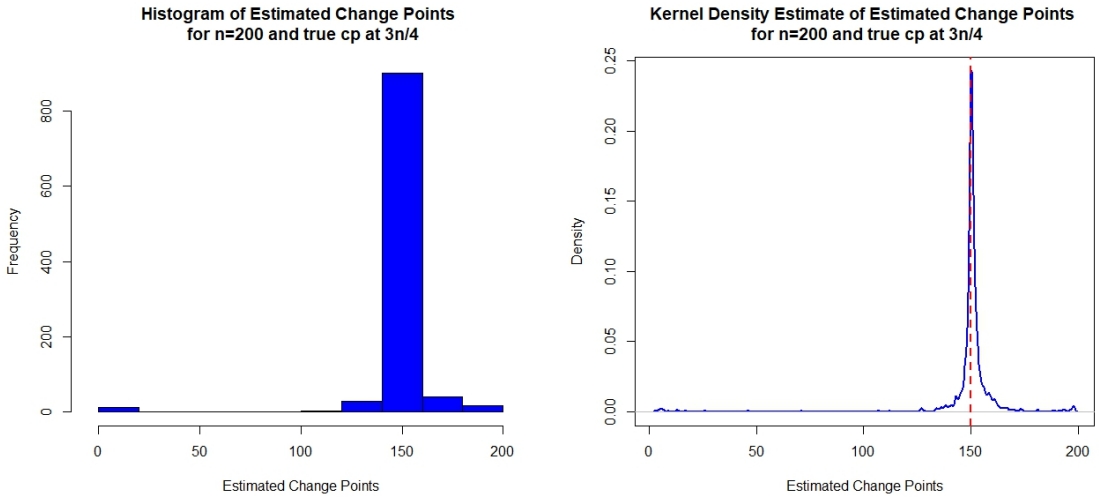

Figure 4.35: Distribution of Estimated Changepoints for a Case of a Single True Change at $3n/4$ for $n = 200$

4.5.3 When the Sample Size is $n = 500$

Figures (4.36), (4.37) and (4.38) show the distribution of the estimated change points for a sample of size $n = 500$. From the histogram, the estimated change points are concentrated around the true change positions for all three locations indicating a high precision and accuracy of the model. This finding is further supported by the KDE curve which has a narrow peak in all three cases indicating that there is no significant bias or variability in the model's estimates for the given sample size.

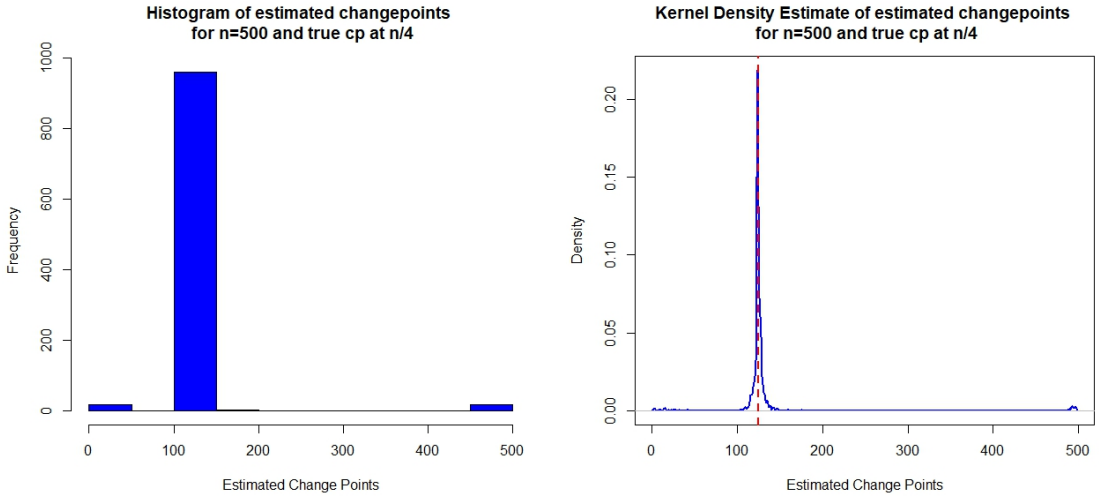

Figure 4.36: Distribution of Estimated Changepoints for a Case of a Single True Change at $n/4$ for $n = 500$

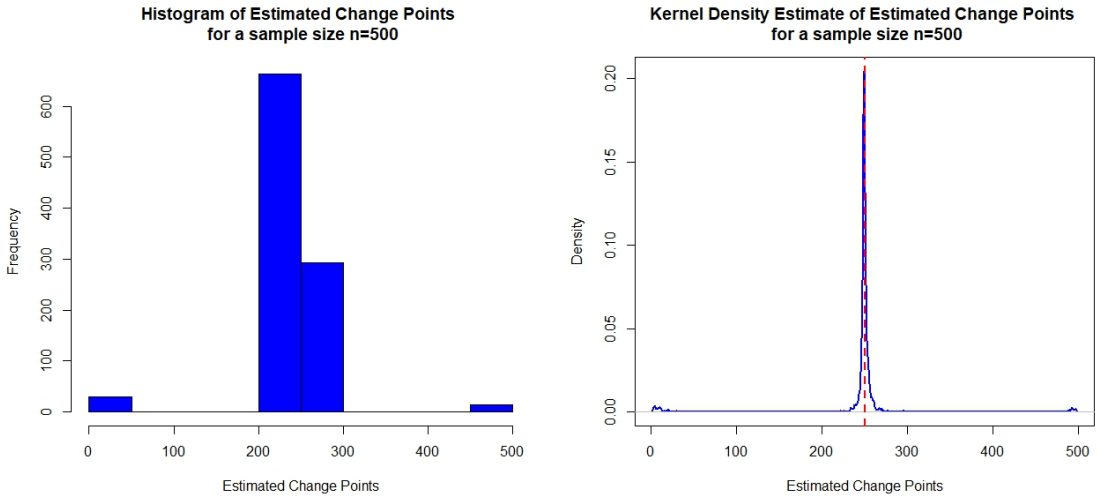

Figure 4.37: Distribution of Estimated Changepoints for a Case of a Single True Change at $n/2$ for $n = 500$

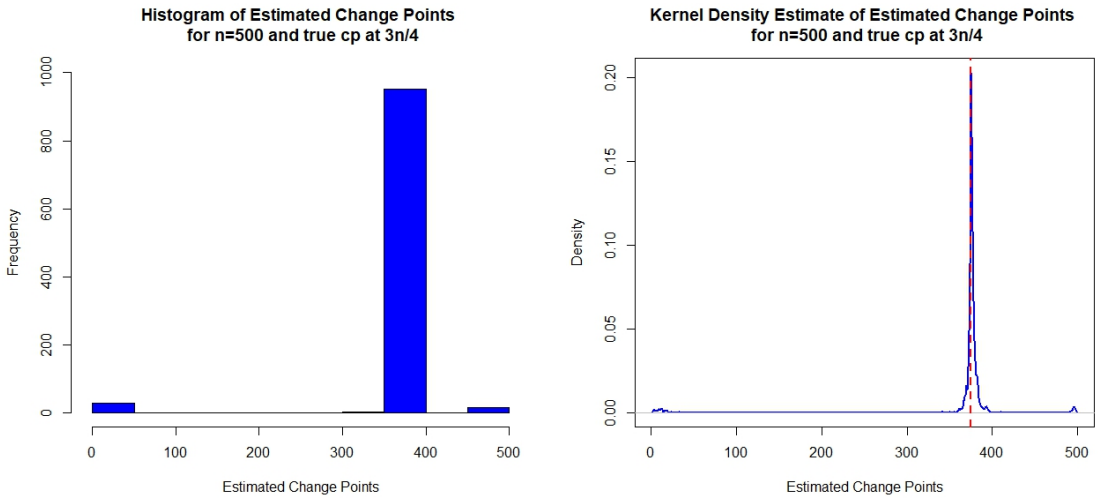

Figure 4.38: Distribution of Change Points for a Case of a Single Change at $3n/4$ for $n = 500$

Table 4.6 summarises the overall performance of the proposed Hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm (NBMCPA) using several complementary performance measures across different sample sizes and true change-point locations. Overall, the results demonstrate that the algorithm exhibits satisfactory detection performance over a wide range of simulation settings, with performance improving consistently as the sample size increases.

The observed improvement in precision, sensitivity and F1 score with increasing sample size is consistent with the theoretical properties of maximum likelihood estimation. Larger samples contain greater statistical information about the underlying Negative Binomial distribution, resulting in more stable parameter estimates and a better-defined likelihood surface around the true changepoint. Consequently, the likelihood ratio statistic is able to distinguish genuine structural changes from random sampling variability with greater confidence, leading to more accurate estimation of changepoint locations. This behaviour agrees with the asymptotic consistency and efficiency of maximum likelihood estimators discussed in Chapter Three.

Sensitivity reached a value of one for sample sizes of $n \geq 20$, indicating that the proposed algorithm successfully detected all simulated changepoints across the remaining experimental scenarios. This demonstrates that the likelihood ratio test, together with the asymptotically derived critical values, possesses adequate statistical power to detect true structural changes while maintaining a low false detection rate. At the same time, the false positive rate decreased steadily as the sample size increased (see figure 4.39, reflecting improved discrimination between genuine changepoints and random fluctuations. This behaviour is expected because sampling variability decreases with increasing sample size, thereby reducing the likelihood that random noise will produce spurious likelihood ratio maxima.

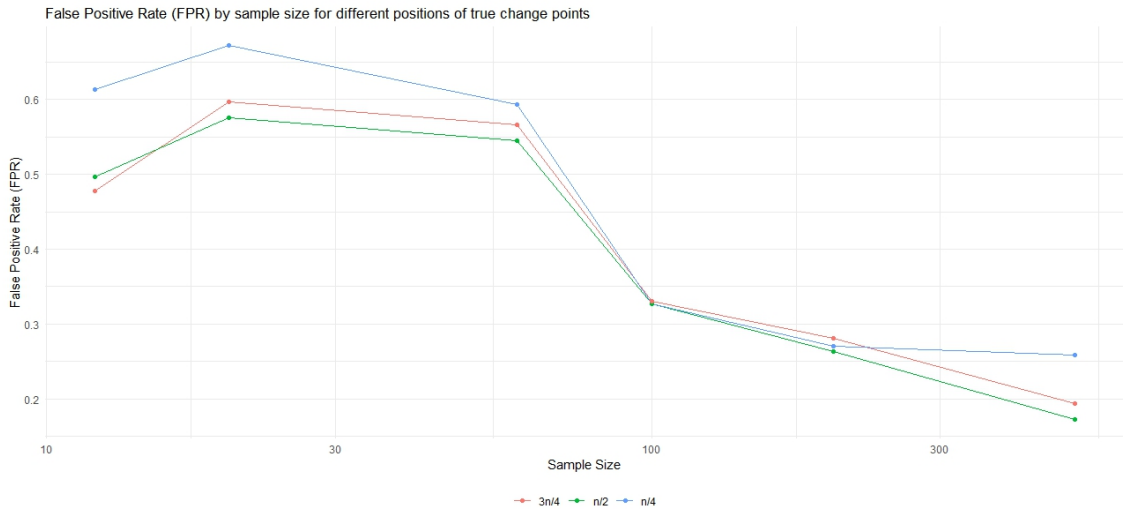

Figure 4.39: Graph of False Positive Rate as Sample Size Increases for Varying True Change Locations

The results further indicate that changepoints located near the centre of the observation period generally achieved slightly higher precision and F1 scores than those located closer to the beginning or end of the sequence. This phenomenon is well documented

in changepoint analysis and arises because central changepoints are supported by larger amounts of information on both sides of the partition. In contrast, changepoints near the boundaries produce relatively short segments, reducing the amount of information available for reliable estimation of the segment parameters and increasing estimation uncertainty.

Although slightly lower estimation accuracy was observed for the smallest sample sizes, this should not be interpreted as a limitation specific to the proposed algorithm. Rather, it reflects the inherent finite-sample behaviour of statistical changepoint estimation. With fewer observations, parameter estimates are naturally subject to greater sampling variability, making neighbouring candidate changepoints more difficult to distinguish. Similar behaviour has been reported for likelihood-based recursive segmentation methods, Binary Segmentation, PELT and Wild Binary Segmentation, where estimation accuracy improves progressively as the information content of the data increases.

Overall, the findings presented in Table 4.6 provide strong empirical support for the proposed algorithm. The combination of high sensitivity, increasing precision, declining false positive rates and improved F1 scores demonstrates that the algorithm provides reliable and accurate detection of structural changes while maintaining an appropriate balance between statistical power and Type I error control. These findings are consistent with likelihood theory and with previous studies on likelihood-based changepoint estimation (Gombay and Horváth, 1996; Bai and Perron, 2003; Killick et al., 2012; Fryzlewicz, 2014).

Table 4.6: Model Performance Metrics for Varying Sample Sizes and Change Locations

Sample size	True cpt	IQR AE	MedAE	MAE	Precision	Sensitivity	FPR	F1 Score
$n = 12$	$n/4$	3	2	2.53	0.385	0.096	0.614	0.154
	$n/2$	2	1	1.77	0.503	0.137	0.497	0.162
	$3n/4$	5	0	2.27	0.521	0.104	0.478	0.113
$n = 20$	$n/4$	6	3	4.29	0.328	1	0.672	0.494
	$n/2$	3	2	2.76	0.424	1	0.576	0.596
	$3n/4$	6	2	4.08	0.403	1	0.597	0.574
$n = 60$	$n/4$	7	2	7.17	0.407	1	0.593	0.579
	$n/2$	4	2	4.79	0.455	1	0.545	0.625
	$3n/4$	6	2	6.27	0.434	1	0.566	0.605
$n = 100$	$n/4$	4	2	6.14	0.673	1	0.327	0.805
	$n/2$	4	2	5.38	0.695	1	0.327	0.820
	$3n/4$	5	2	5.52	0.669	1	0.331	0.802
$n = 200$	$n/4$	3	1	4.94	0.730	1	0.270	0.844
	$n/2$	3	1	6.43	0.737	1	0.263	0.849
	$3n/4$	3	1	6.35	0.719	1	0.281	0.837
$n = 500$	$n/4$	3	1	11.82	0.741	1	0.259	0.851
	$n/2$	3	1	13.21	0.820	1	0.172	0.906
	$3n/4$	4	1	14.96	0.806	1	0.194	0.893

4.6 Summary of Simulation findings

The simulation study considered a single change detection problem for datasets with sizes ranging between $n = 12$ and $n = 500$ while controlling for the magnitude of change. Given the foregoing results, the NBMCPA consistently detects and estimates changes points across varying sample sizes and true change point locations. In particular, model's ability to accurately locate the true change point varies with depending on the location of the true change point, relative to the size of the dataset, similar to the results discussed in Truong et al. (2020).

In this study, the model performance was compared across three different true change point locations: at $n/2, n/4$, and $3n/4$. The simulation results revealed that the algorithm has highest precision when the true changepoint is located at the middle (point $n/2$) of the dataset consistent with Mundia et al. (2014) ; Nyambura et al. (2016); Aminikhanghahi and Cook (2017). A symmetrical distribution with equal amount of data either side of the true change point makes it easier for the model to detect the change point because the magnitude of change in the data before and after the true change point is likely to be roughly similar, allowing the model to detect subtle abrupt shifts.

When the true changepoint is located at position $n/4$ or $3n/4$, there is a data imbalance such that the true change point is located closer to the beginning (at $n/4$) or end (at $3n/4$)of the dataset. The smaller data segments before or after the change may have less data to provide evidence of the change, contingent on the sample size, which could result in lower detection accuracy.

The observed reduction in changepoint estimation accuracy for smaller sample sizes is consistent with established statistical theory and should not be interpreted as a limitation unique to the proposed Hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm. Changepoint estimation is fundamentally an inference problem in which the amount of information available for detecting structural changes depends directly on the sample size. When the number of observations is small, parameter estimates are subject to greater sampling variability, resulting in increased uncertainty in the likelihood ratio statistic and, consequently, less precise estimation of changepoint locations.

In particular, small samples provide relatively limited information for distinguishing genuine structural changes from random fluctuations in the observed count process. As a result, neighbouring candidate changepoints may produce similar likelihood values, making it more difficult to identify the true changepoint with high precision. This phenomenon may lead to slight misclassification or small deviations between the true and estimated changepoint locations, particularly when the structural change is relatively weak or occurs near the boundaries of the observation period.

As the sample size increases, the amount of information available for parameter estimation also increases, leading to more stable maximum likelihood estimates and greater separation between the likelihood values corresponding to competing changepoint locations. Consequently, the likelihood ratio statistic becomes increasingly sensitive to genuine structural changes, resulting in improved estimation accuracy and reduced variability in the estimated changepoint locations. The simulation results presented in this study clearly demonstrate this behaviour, with estimation accuracy improving consistently as the sample size increases.

The observed improvement in performance with increasing sample size is consistent with the asymptotic properties of maximum likelihood estimation, whereby estimators become increasingly consistent and efficient as the sample size grows under the regularity conditions discussed in Chapter 3 van der Vaart (2000); Pawitan (2013). Similar finite-sample behaviour has been reported for other likelihood-based and recur-

sive changepoint detection methods in the literature, indicating that reduced estimation accuracy in small samples is an inherent characteristic of statistical changepoint estimation rather than a weakness specific to the proposed methodology.

From a practical perspective, these findings suggest that the proposed algorithm is particularly well suited to applications involving moderate to large datasets, where sufficient information is available to estimate both the model parameters and changepoint locations with high precision. Nevertheless, the simulation results demonstrate that even for smaller sample sizes, the algorithm maintains satisfactory performance while exhibiting the expected finite-sample variability common to likelihood-based estimation procedures.

In a nutshell, the model performance is optimal when the true changepoint is located at $n/2$ because the data is balanced, and the shift is clear on both sides of the change point, which results in a high precision and detection accuracy, both of which increase with sample size. Further, the rate of false alarms is highest for changes located quarterway, followed by changes at $3n/4$, and lowest for changes located midway through the dataset. This simulation study considered model performance under a single-change setting. However, as the NBMCPA is designed to detect a single change at a time, starting with the most pronounced, the results of this study may be generalized for the multiple changepoint setting. Further investigations can be done to investigate the algorithm performance where there are multiple subtle changes in close proximity to each other within small and medium datasets.

4.6.1 Algorithm Efficiency and Scalability

An important consideration in evaluating any changepoint detection procedure is its computational efficiency when applied to increasingly large datasets. The simulation results demonstrate that the proposed NBMCPA maintains high estimation accuracy while accommodating sample sizes of up to 500 observations without deterioration in performance. As discussed in Chapter Three, the recursive binary segmentation procedure requires approximately $O(n \log n)$ computations under balanced partitioning, making it computationally efficient for moderate and large datasets.

As the sample size increases, the computational time naturally increases because additional candidate changepoint locations must be evaluated. However, the recursive partitioning strategy substantially reduces the computational burden compared with exhaustive multiple-changepoint search procedures, which may require quadratic or higher

computational complexity. The proposed algorithm therefore achieves a favourable balance between computational efficiency and estimation accuracy.

From a statistical perspective, larger datasets not only increase computational requirements but also provide greater information for parameter estimation and hypothesis testing. Consequently, the simulation results indicate that the modest increase in computational effort is accompanied by substantial improvements in changepoint estimation accuracy, precision and stability. These findings suggest that the proposed algorithm is well suited for analysing long count-data sequences commonly encountered in epidemiology, environmental monitoring, finance and industrial process control.

4.6.2 Comparison with Existing Changepoint Detection Methods

The simulation study presented in this chapter evaluated the statistical performance of the proposed hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm under a range of sample sizes and changepoint configurations. Although the study did not include an empirical benchmarking exercise against other established changepoint detection algorithms such as Binary Segmentation (BinSeg), Pruned Exact Linear Time (PELT), Moving Sum (MOSUM) or Wild Binary Segmentation (WBS), the methodological characteristics of the proposed algorithm can be compared conceptually with these approaches.

Binary Segmentation (BinSeg) is one of the earliest and most widely used multiple changepoint detection procedures. Similar to the proposed algorithm, BinSeg recursively partitions a sequence after detecting a statistically significant changepoint. However, classical BinSeg is primarily developed for data satisfying relatively simple distributional assumptions and is commonly implemented for Gaussian observations or models in which the variance structure is comparatively straightforward. In contrast, the proposed NBMCPA was specifically developed within a likelihood framework for Negative Binomial count data, thereby explicitly accommodating over-dispersion, which is a common characteristic of epidemiological count processes such as COVID–19 infection data.

The Pruned Exact Linear Time (PELT) algorithm (Killick et al., 2012) adopts a global optimisation strategy by minimising a penalised cost function while employing pruning rules to achieve linear computational complexity under suitable conditions. PELT is computationally efficient and guarantees an optimal segmentation with respect to the chosen cost function. By comparison, the proposed algorithm employs recursive

likelihood ratio testing and stepwise recursive binary segmentation to sequentially identify statistically significant changepoints. Although recursive segmentation does not guarantee a globally optimal solution in all circumstances, it offers a transparent and statistically interpretable framework that naturally integrates likelihood ratio hypothesis testing with maximum likelihood estimation.

Similarly, Wild Binary Segmentation (Fryzlewicz, 2014) improves upon classical Binary Segmentation by examining multiple randomly selected subintervals, thereby enhancing the detection of closely spaced changepoints. MOSUM methods detect structural changes using moving-window statistics and are particularly effective for identifying local changes in long sequences. These methods differ fundamentally from the proposed methodology, which was designed specifically to estimate changepoints in over-dispersed count data while providing formal likelihood-based inference for statistical significance.

An important advantage of the proposed NBMCPA is its ability to model both equi-dispersed and over-dispersed count data within a unified Negative Binomial likelihood framework. Since the Poisson distribution is obtained as a limiting case of the Negative Binomial distribution, the proposed methodology provides a flexible modelling framework that extends beyond traditional Poisson-based changepoint procedures. Furthermore, the use of likelihood ratio testing and asymptotically derived critical values provides a principled statistical basis for changepoint detection and helps control false detections during the recursive segmentation process.

Nevertheless, the present study acknowledges that an empirical benchmarking study comparing the proposed algorithm with established methods such as BinSeg, PELT, MOSUM and Wild Binary Segmentation would provide additional insight into their relative computational efficiency, detection accuracy and robustness under different simulation scenarios. Such comparative studies were beyond the scope of the current research and therefore represent an important direction for future methodological investigation.

4.7 Estimating Change Points in COVID-19 Infections Data for Kenya and Identifying Assignable Causes of Change

4.7.1 Overview

The Novel COVID-19 disease was first detected in Wuhan, China in Dec 2019 and quickly spread to the rest of the world due to its highly communicable nature. The pandemic impacted nations heavily due to extreme symptoms, high death rate and lack of cure; No prior vaccines existed either. Kenya reported its first COVID-19 case on March 13, 2020. By July 30, 2020, the country had recorded 17,975 cases and 285 deaths. The Kenyan government implemented various health and non-health strategies to mitigate the pandemic's impact. GoK measures included: Contact tracing, compulsory quarantine of suspected and confirmed cases, establishing COVID-19 testing centers, closure/reopening of schools, mask-wearing, social distancing including in PSVs, ban/lifting of ban on public gatherings, closure of bars and restaurants, sanitizing/handwashing centers outside business premises, lock-down/stay at home orders. Of interest in this study is to determine whether the COVID-19 policies implemented/lifted by the National Government of Kenya had an influence on the spread of the disease (decrease or increase in number of daily infections—detected as changes in the infections curve).

4.7.2 Data Source, Description and Exploration

Secondary data on COVID-19 daily infections worldwide was obtained from the John Hopkins university website. The data spanned 3rd January 2020 - 6th September 2023 The dataset was made up of seven variables, namely: Date reported, Country code, Country, Number of new cases, Cumulative cases, Number of new deaths, and Cumulative deaths. Two variables showing the Date reported and Number of new cases for Kenya were extracted. Summary statistics for these data indicated evidence of over-dispersion ($D > 1$) in the count data (see table 4.7) No missing values were found in the COVID–19 data. Figure 4.40 shows a plot of the raw COVID-19 daily cases data.

Due to the high number of zero infections in the first three and last eight months, data were heavily right skewed and zero-inflated. From the histogram, the peak (mode) was on the left side of the plot, and the tail stretched to the right (toward higher values).There

Table 4.7: Summary Statistics for the COVID-19 Daily Cases Dataset

Minimum	Maximum	Range	N	Mean	Variance	Dispersion Index (D)
0	3749	3749	1343	256.11	195,095.89	761.76

Note. N denotes the number of daily observations, and the dispersion index (D) is computed as the ratio of the sample variance to the sample mean.

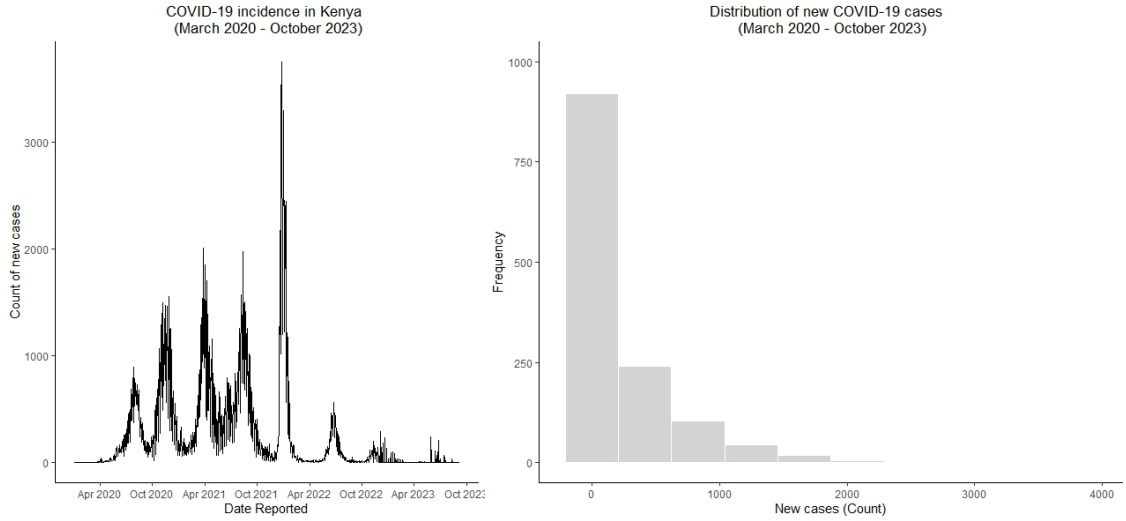

Figure 4.40: Distribution of Raw COVID-19 Daily New Cases

were a few large values pulling the distribution toward the right. Since the first case in Kenya confirmed in March 2020, hence the prior zero values were dropped. Further, there was no need to model infections when the disease was already contained with daily infections dropping to zero. As such, data were truncated to cover the period March 2020 - Aug 2021 (approximately 500 days). Figure 4.41 shows the distribution of the truncated data series for the study period.

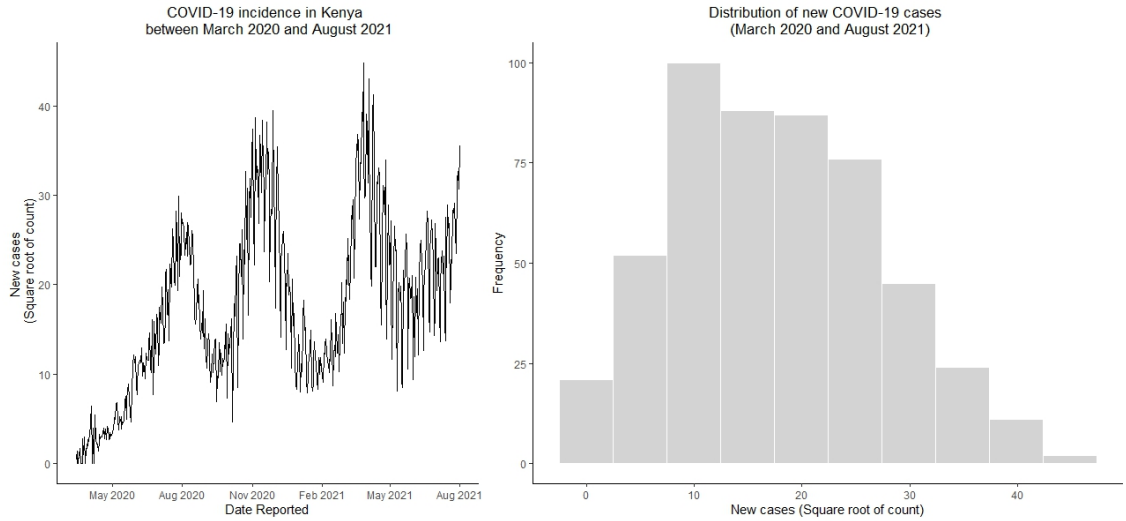

Figure 4.41: Distribution of Raw COVID-19 Daily New Cases Between March 2020 and August 2021

4.7.3 Data Transfromation

Different transformation methods were applied to the data including square-root transform, log transform, log $(x + 1)$ transform, z-score standardization, min-max scaling, and differencing. The square-root transformation was found to work best since it naturally handles zero values, stabilizes variance moderately, and is not too stringent on large values. Further, the shape of the original data distribution was preserved.

In addition to the square-root transform, a 7-point moving average was computed for

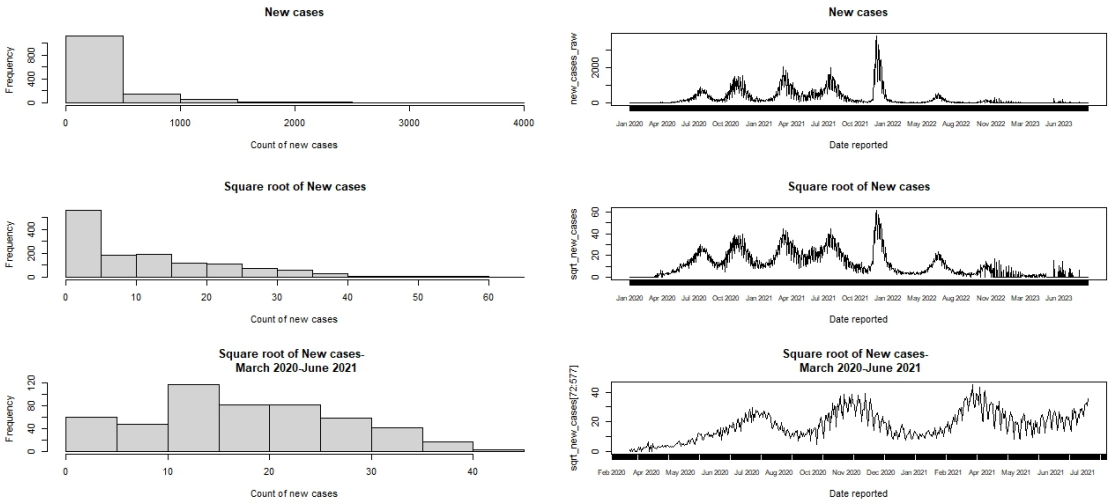

Figure 4.42: Distribution of Square-root Transformed COVID-19 Daily New Cases

the daily count data series to reduce random noise, stabilize variance, improve signal-to-noise ratio, and highlight trends by removing weekly patterns and end-of-week reporting effects. The moving averaging made it easier to detect and estimate change points on the infections curve. Figures 4.42 and 4.43 illustrate the effect of the preprocessing procedures applied to the COVID–19 daily infection data prior to changepoint estimation. Figure 4.42 shows the distribution of the square-root transformed daily case counts, while Figure 4.43 presents the same transformed series together with a 7-day moving average used to smooth short-term fluctuations.

The square-root transformation substantially reduced the strong right-skewness observed in the raw COVID–19 case counts while preserving the overall temporal pattern of the epidemic. Unlike logarithmic transformations, the square-root transformation naturally accommodates zero-valued observations without requiring arbitrary adjustments such as $\log(x + 1)$. Furthermore, it moderates the influence of extremely large daily case counts, thereby reducing heteroscedasticity while retaining the essential characteristics of the original data. Consequently, the transformed series provides a more stable basis for likelihood estimation under the proposed Negative Binomial changepoint model.

Although the transformation reduced variability, considerable day-to-day fluctuations remained due to reporting delays, weekly reporting cycles and other random sources of variation that are characteristic of epidemiological surveillance data. To further improve the signal-to-noise ratio, a 7-day moving average was superimposed on the transformed series, as shown in Figure 4.43. The moving average smooths short-term irregularities while preserving the underlying long-term trend, making structural changes in the infection process more readily identifiable.

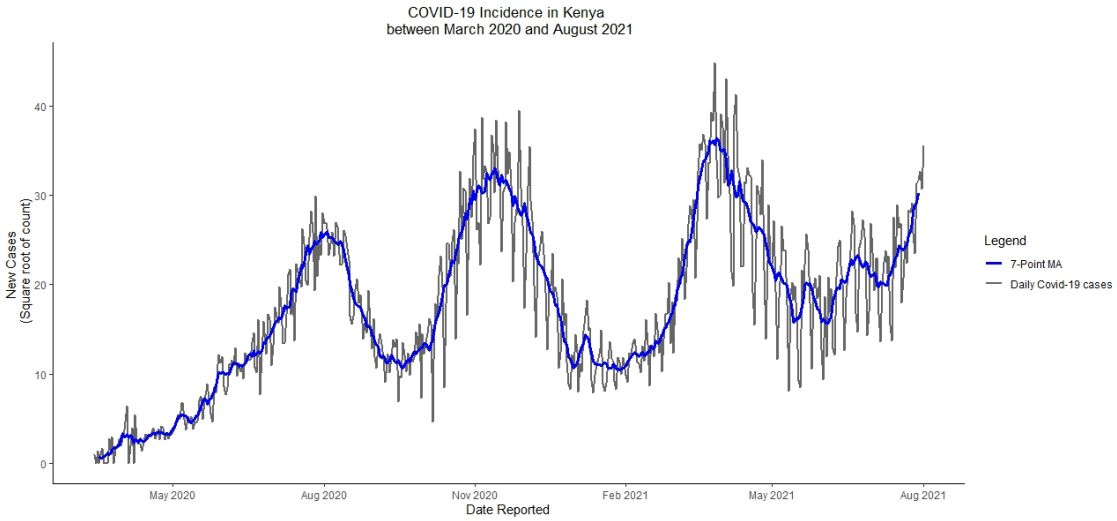

Figure 4.43: Squareroot Transformed Data with a 7-point Moving Average Superimposed

The smoothed trend clearly reveals the successive epidemic waves experienced in Kenya during the study period and provides a clearer visual representation of the underlying changes in transmission dynamics. By reducing random variability without substantially altering the long-term structure of the series, the moving average facilitates more reliable estimation of changepoints by preventing the likelihood ratio procedure from responding to isolated daily fluctuations. Consequently, the combination of square-root transformation and moving average smoothing produced a data series that retained evidence of over-dispersion while providing a stable and interpretable input for the proposed algorithm.

The adequacy of the preprocessing procedure was further confirmed by the summary statistics presented in Table 4.8, which show that the transformed series remained over-dispersed (dispersion index $D > 1$). This finding justified the continued use of the Negative Binomial modelling framework while ensuring that the transformed data were more suitable for robust likelihood-based changepoint estimation.

Table 4.8 gives the summary statistics for the transformed, averaged and truncated data series.

Table 4.8: Summary Statistics for the Transformed COVID-19 Daily Cases Dataset

Minimum	Maximum	Range	Mean	Variance	Dispersion index (D)	Sample size (n)
1	36	35	17.64	75.79	4.30	500

4.7.4 Detection and Estimation of Changepoints under the NBMCPA

Figure 4.44 shows the first estimated changepoint in the data series dated as 25/05/2020. The change detected was found to be statistically significant as shown by the peak of the LRT statistic curve which lay above the critical value at the 5% level. The estimated changepoint was found to coincide with the start of the first distinctive wave of viral infection in the country.

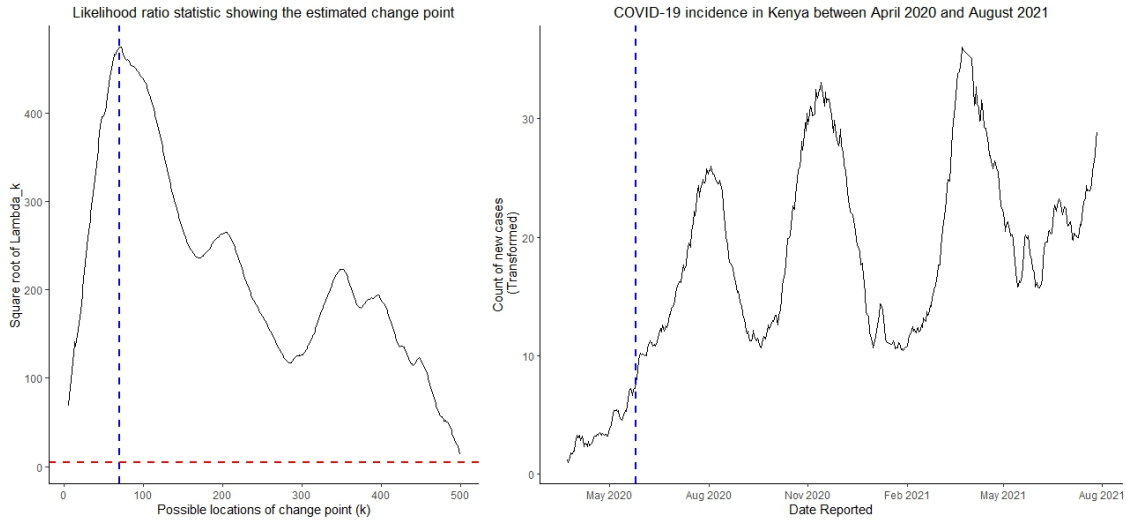

Figure 4.44: Estimate of the First Changepoint in the COVID-19 Infections Trend

Following the identification of the first changepoint, the data was segmented at the changepoint and the subsequent partitions tested for existence of change under the recursive binary segmentation approach discussed in chapter 3. Another two statistically significant changes in the daily new cases trend were detected and estimated at 18/10/2020 and 01/03/2021 respectively. Figure 4.45 shows the first three changepoints detected in the COVID-19 infections trend in Kenya. The three change points were found to correspond to the start of the first three waves of infection experienced in the country.

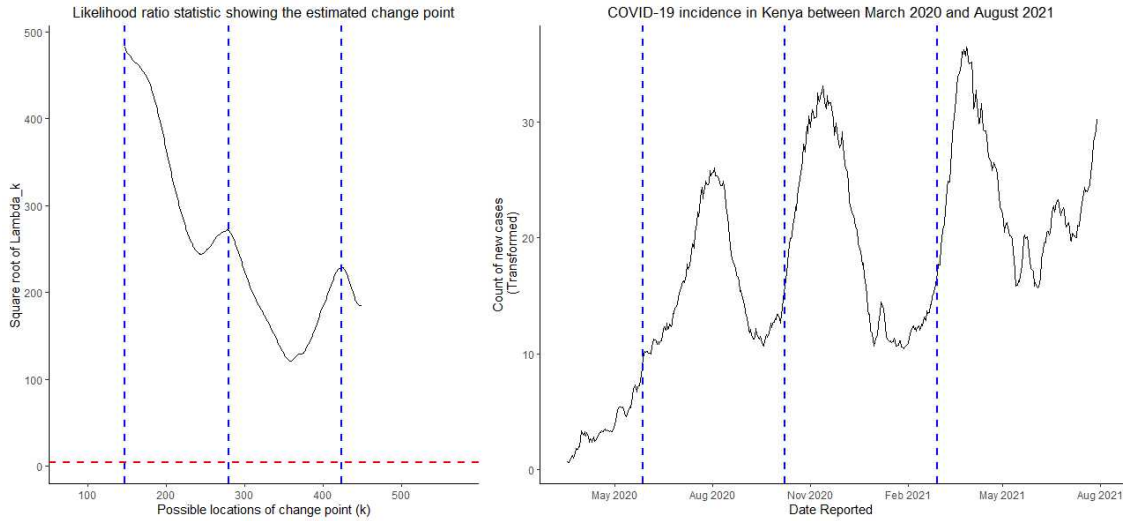

Figure 4.45: Estimates of the First Three Changepoints in the COVID-19 Infections Trend

4.7.5 Possible Assignable Causes of Change

The proposed hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm identifies statistically significant structural changes in the observed COVID–19 infection counts based solely on the observed data and the underlying likelihood framework. The algorithm does not incorporate information on government interventions, epidemiological events or behavioural changes during the estimation process. Consequently, the identification of assignable causes constitutes a post-estimation interpretation exercise intended to provide epidemiological context for the detected changepoints rather than evidence of direct causal relationships.

Following the estimation of statistically significant changepoints, the corresponding calendar dates were compared with official COVID–19 policy timelines, Ministry of Health situation reports, government press releases and documented epidemiological events occurring during the study period. The objective was to identify events that were temporally associated with the detected structural changes and whose expected epidemiological effects were consistent with the observed changes in infection trends. It should therefore be emphasized that the relationships presented in this section represent plausible temporal associations rather than definitive causal effects.

Table 4.9 summarises the estimated changepoints, their corresponding calendar dates, the direction of change in the infection process and the major epidemiological or policy-related events that were temporally associated with each detected changepoint.

Table 4.9: Estimated Changepoints and Associated Epidemiological or Policy Factors

Change point	Date	Direction	Plausible Epidemiological or Policy-Related Association
CP_1	25 May 2020	Increase	Early phase of the pandemic characterised by limited understanding of disease transmission, evolving surveillance systems, constrained testing capacity, challenges in contact tracing and reduced health-seeking behaviour following mandatory quarantine policies.
CP_2	18 Oct 2020	Increase	Gradual reopening of educational institutions, reopening of bars and restaurants under revised public health guidelines, increased population mobility and relaxation of selected containment measures.
CP_3	19 Dec 2020	Decrease	Implementation of nationwide curfew measures, suspension of political gatherings and reinforcement of public health regulations during the December holiday period.
CP_4	1 Mar 2021	Increase	Progressive relaxation of selected containment measures, increased social interaction, emergence of more transmissible SARS-CoV-2 variants and the onset of the third epidemic wave reported in Kenya.

Source: Compiled from Kenya Ministry of Health COVID–19 Situation Reports (2020–2021), Presidential Executive Orders and Government of Kenya public health directives.

The estimated changepoints exhibit close temporal correspondence with several major public health interventions and epidemiological developments documented during the COVID–19 pandemic in Kenya. For example, the first detected changepoint (25 May 2020) occurred during the early stages of the pandemic when surveillance systems, testing capacity and contact tracing mechanisms were still being established. During

this period, considerable uncertainty surrounded disease transmission, and testing was largely targeted toward symptomatic individuals and identified contacts. Consequently, the observed increase in reported infections is consistent with the combined effects of expanding community transmission, progressive improvements in case detection and evolving surveillance systems rather than any single identifiable event.

The second estimated changepoint (18 October 2020) coincided with the phased reopening of schools, reopening of bars and restaurants and gradual relaxation of selected containment measures. These developments were associated with increased population mobility and interpersonal contact, factors that have been identified in epidemiological studies as contributing to increased opportunities for disease transmission. The observed upward shift in infection counts is therefore consistent with the expected epidemiological consequences of increased social mixing, although other contemporaneous factors, including behavioural adaptation and changes in testing practices, may also have contributed.

The third changepoint (19 December 2020) was temporally associated with the reinforcement of nationwide curfew measures and restrictions on political gatherings. These interventions were intended to reduce population mobility and limit opportunities for community transmission. The subsequent decline in reported infections is consistent with the expected effects of non-pharmaceutical interventions reported in both Kenyan Ministry of Health surveillance reports and international epidemiological studies. Nevertheless, the observed reduction cannot be attributed exclusively to these interventions because seasonal behavioural changes, reporting practices and other unmeasured factors may also have influenced the infection dynamics.

The fourth estimated changepoint (1 March 2021) corresponded with the emergence of Kenya's third epidemic wave, during which increased community transmission was reported alongside the detection of more transmissible viral variants and gradual relaxation of some public health restrictions. Government surveillance reports documented a sustained increase in confirmed cases during this period, supporting the temporal association between the detected changepoint and the changing epidemiological situation.

Overall, the detected changepoints correspond closely with major epidemiological and policy developments reported during the COVID–19 pandemic in Kenya. These findings suggest that the proposed NBMCPA successfully identified statistically significant structural changes that are temporally consistent with documented public health events.

However, because the analysis is based on observational surveillance data, the estimated relationships should be interpreted as plausible temporal associations rather than evidence of direct causality. Multiple concurrent interventions, changes in testing capacity, behavioural responses, viral evolution and other unmeasured factors are likely to have jointly influenced the observed infection trends. Consequently, the principal contribution of the proposed methodology is the statistically rigorous identification of the timing of structural changes, while the interpretation of potential assignable causes provides valuable epidemiological context for understanding the observed patterns.

4.8 Determining the Average Lag Between Policy Dates and Calendar Change Points in the COVID-19 Infections Trend in Kenya

4.8.1 Overview

This section presents the results of the policy–change point and the policy–peak lag analyses, which examined the temporal relationship between COVID-19 policy implementation dates, corresponding changepoints, and the subsequent peaks in the infection trend. The objective was to determine how quickly observable changes in the epidemic curve followed major and minor government interventions.

4.8.2 Data Source and Policy Classification

The COVID-19 measures and policies enforced by the Government of Kenya to contain the pandemic were obtained second hand from the Ministry of Health press releases, gazetted legal notices under the Public Health Act and Public Order Act, and presidential addresses made during the study period. A set of 12 GoK COVID-19 measures were considered between March 2020 and October 2021. The COVID-19 policy timeline presented in Table 4.10 was compiled from multiple authoritative sources to ensure accuracy and completeness. Primary data on policy implementation dates were obtained from official communications issued by the Ministry of Health Kenya Ministry of Health Kenya (2021) and formal government directives, including presidential addresses and gazetted legal notices Government of Kenya (2021). These sources provide the most direct record of intervention timing and legal enforcement.

Table 4.10: COVID-19 Policy Measures in Kenya Between March 2020 and October 2021

Date	Policy	Classification	Reason / Justification
2020-03-15	Travel restrictions, school closure	Major	Nationwide measures altering mobility
2020-03-27	Night curfew imposed	Major	Movement control
2020-04-06	Nairobi/Mombasa lockdown	Major	Inter-county spread limitation
2020-06-06	Phased reopening announced	Major	Relaxation of national measures
2020-07-06	Domestic flights resume	Minor	Sector-specific easing
2020-08-01	International flights resume	Major	Border reopening
2020-09-28	Bars reopen, curfew shortened	Major	Economic reopening
2020-11-04	Restrictions tightened	Major	Second-wave containment
2021-03-26	Partial lockdown (5 counties)	Major	Regional but epidemiologically significant
2021-05-01	Restrictions eased	Major	Nationwide impact
2021-08-18	Curfew extended	Minor	Continuation, not a new measure
2021-10-20	Curfew lifted nationwide	Major	National reopening

Note: Policy dates were compiled from World Health Organization (2020–2021) COVID-19 situation reports for Kenya, which document official Government of Kenya interventions.

The policies were classified into major and minor according to the classification logic described in table 3.2 and results summarized in table 4.10.

To enhance reliability and minimize potential reporting inconsistencies, the compiled policy dates were cross-validated against international datasets and reports. In particular, the World Health Organization (2020–2021) situation reports for Kenya were used to corroborate major public health interventions, as these reports synthesize country-level policy actions based on official government submissions. Additionally, the Oxford COVID-19 Government Response Tracker discussed in Hale et al. (2021) was utilized to verify the timing and classification of key interventions within a standardized global framework.

The integration of these sources ensures both internal validity and external comparability of the policy timeline. While national sources provide detailed and context-specific information, international datasets such as the WHO reports and the OxCGRT enhance

consistency and allow the policy measures to be situated within the broader global response to the COVID-19 pandemic. This triangulated policy dataset forms the basis for the subsequent policy–change point lag analysis, enabling a robust assessment of the temporal relationship between government interventions and observed structural changes in the COVID-19 infection trajectory.

The estimated change points exhibit a clear temporal relationship with major government interventions, although with varying lag durations. The first change point (25 May 2020) occurs approximately 7–8 weeks after the introduction of stringent containment measures in March and April 2020, suggesting a delayed but cumulative policy effect, likely influenced by limited early testing capacity and reporting delays. The second change point (18 October 2020) appears about 20 days after the reopening of bars and relaxation of curfew measures, aligning closely with the expected epidemiological lag and indicating a rapid increase in transmission following social reopening. The third change point (19 December 2020) follows the reintroduction of restrictions in early November by approximately six weeks, reflecting a slower response potentially shaped by behavioral fatigue and increased mobility during the festive season. Overall, the results demonstrate that policy impacts on infection dynamics are not instantaneous but manifest after context-dependent delays.

4.8.3 Choice of Window and Results of Lag Analysis

A window of 21 days was initially used to assess the lag between each major policy date and estimated changepoint through forward-matching criterion. However, the results showed that only one changepoint was within 21 days of the immediate previous policy, while others were outside the set window. This finding indicated that either policies took more than 3 weeks to show effects, or the changepoints may not have been directly attributed to the policies due to other overriding factors. Therefore, the final lag analysis was based on the reverse matching approach where last preceding policy date for each of the changepoints estimated in the infections trend, but using a window W=45 days. Three out of the four changepoints were reverse-matched to preceding policy dates within the set window, resulting in three lag values available for use in the analysis.
The observed median lag was 45 days while the mean lag was 38 days with a standard deviation of 15 days (see table 4.11). The average lag was quite large numerically, but there was a lack of precision, due to the small sample, to claim that the population mean is statistically different from 0.

Table 4.11: Descriptive Statistics of Policy-Changepoint Lags

Statistic	Value (days)
Mean	38
Median	45
Range	20 – 49
Standard deviation	15.71
95% CI for mean	20.2 – 55.7

4.8.4 Results of the Peak Analysis

4.8.4.1 Policy–Peak Alignment

Twelve government policy actions were evaluated in this analysis. Of these, ten policies had identifiable peaks within the permissible matching window, while two policies (implemented on 18 August 2021 and 20 October 2021) did not correspond to any detectable peak. Table 4.12 summarizes the matched peak dates and the corresponding number of days between policy implementation and the subsequent observed peak.
Overall, the lag between the introduction of a policy and the nearest subsequent infection peak ranged from 2 to 19 days for the earliest policies, and from 5 to 11 days for most of the later policies. These lags represent the temporal distance between each policy action and the maximum point in the smoothed epidemic curve during the corresponding period. The variability in lag estimates reflects differences in intervention type, timing relative to epidemic wave progression, and the prevailing transmission dynamics.

Table 4.12: Policy Dates, Matched Peaks, and Lag in Days

Policy Date	Matched Peak	Lag (days)
2020-03-15	2020-04-03	19
2020-03-27	2020-04-03	7
2020-04-06	2020-04-11	5
2020-06-06	2020-06-11	5
2020-07-06	2020-07-16	10
2020-08-01	2020-08-06	5
2020-09-28	2020-10-04	6
2020-11-04	2020-11-06	2
2021-03-26	2021-03-31	5
2021-05-01	2021-05-12	11

4.8.4.2 Descriptive Statistics for Policy-to-Peak Lags

Across the ten policies with valid matches, the mean peak lag was 7.5 days, with a standard deviation of 4.81 days. The median lag fell within the range of 5–7 days, indicating that, on average, epidemic peaks tended to occur approximately one week after the introduction of a policy measure. Although this provides a preliminary indication of short-term peak responses, the wide spread in lag values highlights the influence of contextual factors and underscores the need for cautious interpretation.

Table 4.13: Summary Statistics for Peak-Policy Lag (in Days)

Statistic	Value
Number of policies with valid matches	10
Mean lag (days)	7.5
Standard deviation (days)	4.81
Median lag (days)	5–7

These descriptive estimates capture only the alignment of policy dates with local peaks and do not necessarily reflect longer-term changes in transmission rates or the overall epidemic trajectory. They should therefore be interpreted as descriptive indicators rather than definitive causal measures.

The mean lag was adopted as the principal summary statistic because it provides a simple and intuitive measure of the average time required for changes in public health policy to become observable in the reported COVID–19 infection data. Given that multiple policy interventions were implemented throughout the study period, the mean offers a convenient overall measure of the typical response time while facilitating comparison across different interventions.

Recognising that response times may vary substantially owing to differences in intervention type, public compliance, disease incubation period, testing practices and reporting delays, the mean lag was interpreted together with measures of variability, including the standard deviation and observed range. These complementary statistics demonstrate that although the average response time was approximately 38 days, individual policy effects exhibited considerable variation. Consequently, the mean should be interpreted as a summary measure of the overall temporal association rather than as a fixed delay applicable to every intervention.

The observed variability is epidemiologically plausible because different interventions are expected to influence transmission dynamics through different mechanisms and over different timescales. Furthermore, changes in reported infections reflect not only transmission but also delays associated with symptom onset, testing behaviour, laboratory confirmation and surveillance reporting. Accordingly, the estimated lag represents the combined effect of these processes rather than the biological effect of the intervention alone. Although the lag analysis provides useful descriptive insight into the temporal relationship between policy implementation and observed structural changes, it does not establish causal effects. Future research may extend the present work by combining the proposed changepoint estimation framework with interrupted time-series analysis, cross-correlation analysis, transfer function models or Bayesian dynamic regression models to investigate both the timing and magnitude of intervention effects. Such approaches would provide complementary evidence regarding the effectiveness of individual public health measures while building upon the statistically robust changepoint estimates obtained using the proposed NBMCPA.

4.8.4.3 Prediction of the Next Expected Peak

Using the empirical distribution of observed policy-to-peak lags, the next expected epidemic peak was estimated to occur on 9 August 2021. A prediction window, derived from the estimated date plus or minus one standard deviation, produced an interval of 6–11 August 2021. This prediction is consistent with the approximate one-week lag pattern observed in earlier periods.

However, this forecast should be interpreted with caution. It is based on a limited number of lag observations and does not account for important epidemiological and behavioural factors such as emerging variants, vaccination coverage, or changes in population mobility and compliance. As such, the prediction serves as an indicative short-term estimate rather than a formal epidemiological forecast.

CHAPTER FIVE
SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter presents the conclusions drawn from the findings of the study and discusses their implications in relation to the research objectives. The conclusions are based on the theoretical developments, simulation results and empirical application presented in the preceding chapters. Rather than providing a summary of the work undertaken, the chapter highlights the principal findings that emerged from the study, the methodological and practical contributions of the proposed hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm, and the significance of these findings for changepoint analysis of count data. The chapter further discusses potential applications of the proposed methodology and identifies areas requiring further investigation to advance both the theoretical and applied aspects of changepoint detection in over-dispersed count processes.

5.2 Conclusions

The principal objective of this study was to develop a likelihood-based multiple changepoint algorithm capable of detecting structural changes in count data while explicitly accommodating over-dispersion through the Negative Binomial distribution. The study was motivated by the observation that many existing changepoint detection procedures are based on Poisson or Gaussian assumptions and therefore perform less effectively when applied to over-dispersed count data commonly encountered in epidemiology, public health, environmental monitoring and industrial quality control.

The conclusions of the study are presented according to the specific research objectives.

5.2.1 Development of the Hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm

The study successfully developed a likelihood-based multiple changepoint detection algorithm founded on the Negative Binomial probability model and recursive likelihood

ratio testing. The proposed methodology integrates maximum likelihood estimation, asymptotically derived likelihood ratio tests and Stepwise Recursive Binary Segmentation to estimate multiple changepoints in count data exhibiting over-dispersion.

Unlike conventional Poisson-based changepoint procedures, the proposed methodology accommodates both equi-dispersed and over-dispersed count processes within a unified statistical framework. Since the Poisson distribution represents a limiting case of the Negative Binomial distribution, the proposed algorithm provides a flexible modelling framework that extends naturally to a wider class of count data without sacrificing statistical rigour. The study therefore contributes a robust likelihood-based methodology for analysing structural changes in stochastic count processes characterised by heterogeneous variability.

5.2.2 Determination of Critical Values for the Likelihood Ratio Test

The study successfully established asymptotic critical values for the likelihood ratio test used to determine the statistical significance of candidate changepoints. The critical values were derived by extending the asymptotic likelihood theory developed by Gombay and Horváth to the Negative Binomial framework adopted in this study.

The resulting hypothesis testing procedure provides a statistically principled mechanism for distinguishing genuine structural changes from random variation while maintaining appropriate control of false detections. Consequently, the proposed algorithm combines efficient changepoint estimation with formal statistical inference, thereby enhancing the reliability and interpretability of the detected changepoints.

5.2.3 Performance of the Proposed Algorithm under Simulation

The simulation study demonstrated that the proposed algorithm consistently identified true changepoints across a wide range of sample sizes and changepoint locations while maintaining low false detection rates. Estimation accuracy improved progressively with increasing sample size, reflecting the increased statistical information available for likelihood estimation and hypothesis testing. Similarly, changepoints located near the centre of the observation period were estimated more accurately than those occurring near the boundaries because the resulting data partitions contained more balanced information for parameter estimation.

These findings are consistent with established likelihood theory and the asymptotic properties of maximum likelihood estimation, indicating that the observed behaviour reflects fundamental statistical characteristics rather than limitations of the proposed methodology. Overall, the simulation results demonstrate that the proposed HL-NBMCPA provides reliable and stable changepoint estimates under a variety of realistic data configurations.

5.2.4 Application to COVID–19 Infection Data in Kenya

Application of the proposed algorithm to daily COVID–19 infection counts in Kenya identified four statistically significant changepoints corresponding to major structural shifts in the progression of the epidemic. Comparison of the estimated changepoint dates with documented public health interventions and epidemiological developments revealed close temporal correspondence between the detected structural changes and several important events occurring during the study period.

However, the study recognises that the proposed algorithm estimates the timing of statistically significant structural changes solely from the observed count data and does not explicitly model the effects of public health interventions. Consequently, the observed correspondence between estimated changepoints and policy measures should be interpreted as evidence of plausible temporal association rather than direct causality. The findings nevertheless demonstrate that the proposed methodology provides a valuable statistical framework for identifying periods during which substantial changes in epidemic dynamics occurred.

5.2.5 Lag Analysis between Policy Implementation and Structural Changes

The lag analysis provided an interpretable measure of the temporal relationship between public health interventions and the estimated changepoints in COVID–19 infection trends. The estimated average lag suggested that observable structural changes in reported infection counts generally occurred several weeks after implementation of major policy interventions, although considerable variability existed among individual interventions.

The study emphasises that the estimated lag represents a descriptive measure of temporal association rather than a causal estimate of intervention effectiveness. Nevertheless, the findings provide useful evidence regarding the typical timescales over which

changes in epidemic dynamics became observable in surveillance data, thereby contributing information that may support planning and evaluation of future public health responses.

5.2.6 Overall Contribution of the Study

This study makes four principal contributions to knowledge.

First, it extends likelihood-based changepoint methodology by developing a unified statistical framework capable of analysing both equi-dispersed and over-dispersed count data through the Negative Binomial distribution.

Second, it contributes to the theoretical development of changepoint analysis by extending asymptotic likelihood ratio testing procedures to the proposed Negative Binomial multiple changepoint framework and establishing corresponding critical values for statistical inference.

Third, the study provides a computationally efficient recursive estimation algorithm that combines maximum likelihood estimation, likelihood ratio testing and Stepwise Recursive Binary Segmentation to detect multiple changepoints while maintaining statistical interpretability and computational feasibility.

Finally, the study demonstrates the practical applicability of the proposed methodology through analysis of Kenya's COVID–19 infection data, illustrating how statistically estimated changepoints can be integrated with epidemiological evidence to improve understanding of structural changes in infectious disease surveillance data. Beyond the immediate application to COVID–19, the proposed methodology provides a general framework that can be adapted to numerous problems involving over-dispersed count data in epidemiology, environmental science, finance, manufacturing and other disciplines where accurate detection of structural changes is required.

5.3 Applications of the Results

The findings of this study have both methodological and practical applications in the analysis of stochastic count data characterised by over-dispersion. The proposed Hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm (HL-NBMCPA) provides a statistically rigorous framework for detecting structural changes

while accommodating the variability commonly observed in count processes. Consequently, the methodology may be applied across a wide range of disciplines where accurate identification of changes in count data is required.

Within public health and epidemiology, the proposed algorithm provides an objective statistical tool for identifying significant changes in disease incidence over time. The methodology can support routine surveillance by assisting researchers and public health authorities in identifying periods during which transmission dynamics change significantly. Such information may complement existing surveillance systems by facilitating timely evaluation of disease trends and assessment of the temporal association between interventions and observed changes in infection patterns.

Beyond infectious disease surveillance, the proposed methodology may be applied to other forms of over-dispersed count data encountered in environmental monitoring, industrial quality control, finance, transportation, insurance and ecological studies. In such applications, reliable identification of structural changes can improve monitoring systems, enhance risk assessment and support evidence-based decision making.

The study also contributes to statistical methodology by extending likelihood-based changepoint analysis to the Negative Binomial framework. The proposed algorithm provides a foundation upon which future computational implementations may be developed, including software packages for commonly used statistical computing environments such as `R` and Python. Such implementations would facilitate wider adoption of the methodology by researchers and practitioners working with count data.

Overall, the proposed methodology offers a flexible and statistically interpretable framework for detecting multiple changepoints in over-dispersed count processes and has the potential to support both methodological research and practical decision-making in a broad range of scientific disciplines.

5.4 Recommendations for Future Research

Although the proposed hybrid Likelihood-Based Negative Binomial Multiple Changepoint Algorithm demonstrated satisfactory theoretical and empirical performance, the study has identified several opportunities for further methodological development and practical application.

First, future studies should consider extending the proposed methodology to online or sequential changepoint detection. The current algorithm was developed for retrospective analysis in which the complete dataset is available before estimation. However, many practical applications, particularly disease surveillance, financial monitoring and industrial process control, require changepoints to be detected as new observations become available. Developing an online version of the proposed algorithm would facilitate real-time monitoring and early warning systems.

Second, future research should investigate multivariate changepoint models capable of simultaneously analysing multiple correlated count processes. During disease outbreaks, several epidemiological indicators such as confirmed cases, hospital admissions, intensive care unit occupancy, mortality and vaccination uptake evolve concurrently. Extending the proposed likelihood framework to multivariate count data would provide a more comprehensive understanding of structural changes across related surveillance indicators.

Third, the present study focused on over-dispersed count data modelled using the Negative Binomial distribution. Future studies should extend the methodology to accommodate under-dispersed count data, which may arise in manufacturing processes, ecological studies and highly regulated monitoring systems. Such extensions may consider alternative count distributions, including the Conway–Maxwell–Poisson (COM-Poisson), Generalized Poisson and Double Poisson distributions.

Fourth, further research should investigate the theoretical properties of the proposed estimator in greater detail. Although the methodology is founded on established likelihood theory and asymptotic inference, formal proofs of estimator consistency, asymptotic normality, convergence properties and statistical efficiency under the proposed recursive estimation framework would strengthen the theoretical foundation of the algorithm and provide additional insight into its statistical behaviour.

Fifth, future studies should undertake comprehensive comparative benchmarking of the proposed algorithm against other established changepoint detection procedures, including Pruned Exact Linear Time (PELT), Binary Segmentation (BinSeg), Wild Binary Segmentation (WBS) and Moving Sum (MOSUM) methods. Such comparisons would enable evaluation of relative detection accuracy, computational efficiency, robustness and scalability under different simulation scenarios and application domains.

Sixth, the computational implementation of the proposed methodology could be enhanced through parallel computing and optimized software development. Develop-

ing dedicated implementations in statistical software such as R or Python would improve computational efficiency and facilitate wider adoption of the methodology by researchers and practitioners working with large count datasets.

Finally, although the present study demonstrated the usefulness of the proposed methodology using COVID–19 infection data from Kenya, future applications should consider other epidemiological, environmental, industrial and financial datasets to further evaluate the generalisability of the algorithm. Applying the methodology across diverse application areas would provide additional evidence regarding its robustness and practical utility under varying data characteristics.

Collectively, these research directions provide opportunities for extending both the theoretical and practical contributions of the proposed NBMCPA while advancing the broader field of changepoint analysis for count data.

5.4.1 Policy and Public Health Recommendations

Based on the observed lag between COVID-19 policy implementation and the corresponding peaks in infection trends, the following recommendations are proposed:

1. **Implement interventions proactively.** Given that policies take approximately 3–7 weeks to yield detectable effects, public health measures should be enacted early, ideally before infections escalate to critical levels. Proactive implementation increases the likelihood of moderating transmission and preventing healthcare system overload.

2. **Maintain policy measures long enough to observe their effects.** Because of the substantial lag between policy introduction and measurable epidemiological response, premature relaxation may falsely appear justified. Decision-makers should allow at least one full lag period (approximately 4–6 weeks) before evaluating the effectiveness of any intervention.

3. **Strengthen continuous monitoring and real-time surveillance.** Ongoing collection of infection, testing, hospitalization, and mobility data is essential for assessing policy impact. Continuous monitoring supports timely adjustments to interventions and enables more accurate characterization of policy effectiveness.

4. **Update peak predictions dynamically as new data become available.** Predictions based on small samples are inherently uncertain. As additional policy

implementations and peaks occur, predictive models should be recalibrated. Incorporating diverse modelling approaches such as lag–based analysis, compartmental models, and mobility–informed models may enhance predictive accuracy.

5. **Tailor interventions based on policy type and expected lag.** Different interventions may exhibit distinct lag structures. For example, comprehensive lockdowns may influence transmission more rapidly than localized or minor restrictions. Future analyses should disaggregate interventions by type and intensity to determine which policies achieve faster or more substantial effects.

6. **Communicate expected delays clearly to the public and stakeholders.** Transparency regarding the inherent delay in policy impact is essential for managing public expectations and ensuring compliance. Communicating that policy effects may not be immediately observable can reduce frustration and bolster public trust.

7. **Expand empirical research to improve future policy planning.** Increasing the number of policy–peak observations will improve statistical power and reliability. Collecting detailed information on policy scope, enforcement, and population adherence will facilitate a more robust evaluation of intervention effectiveness and strengthen evidence-based policy making.

REFERENCES

Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367.

Anastasiou, A., Cribben, I., and Fryzlewicz, P. (2022). Cross-validatory selection of the number of change-points in high-dimensional time series via the lasso. *Journal of the American Statistical Association*, 117(538):903–915.

Anastasiou, A. and Fryzlewicz, P. (2022). Detecting multiple generalized change-points by isolating single ones. *Metrika*, 85(2):141–174.

Aue, A. and Kirch, C. (2023). Change point analysis for time series: A review. *Annual Review of Statistics and Its Application*, 10:1–25.

Baranowski, R., Chen, Y., and Fryzlewicz, P. (2019). Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):649–672.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression Analysis of Count Data*. Cambridge University Press, New York, 2 edition.

Chander, A., Singh, P., and Kumar, V. (2024). Automatic change point detection in time series: A review. *Expert Systems with Applications*, 235:121345.

Chen, J. and Gupta, A. K. (2012). *Parametric Statistical Change Point Analysis*. Birkhäuser, Boston.

Chen, Y., Li, X., and Wang, L. (2023). Deep learning for changepoint detection in overdispersed count data. *Neural Computation*, 35(4):678–706.

Chen, Y.-C. (2017). A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187.

Cho, H. (2016). Change-point detection in panel data via double CUSUM statistic. *Electronic Journal of Statistics*, 10(2):2000–2038.

Coughlin, S. S., Yiğiter, A., Xu, H., Berman, A. E., and Chen, J. (2021). Early detection of change patterns in covid-19 incidence and the implementation of public health policies: A multi-national study. *Public Health in Practice*, 2:100064.

Dehning, J., Zierenberg, J., Spitzner, F. P., et al. (2020). Inferring change points in the spread of covid-19 reveals the effectiveness of interventions. *Science*, 369(6500):eabb9789.

Dette, H. and Quanz, P. (2023). Detecting relevant changes in the spatiotemporal mean function. *Journal of Time Series Analysis*, 44(5–6):505–532.

Eichinger, B. and Kirch, C. (2018). A mosum procedure for the estimation of multiple random change points. *Bernoulli*, 24(1):526–564.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., and Bhatt, S. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820):257–261.

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

Fryzlewicz, P. and Subba Rao, S. (2014). Multiple-change-point detection for auto-regressive conditional heteroscedastic processes. *JRSS B*, 76(5):903–924.

Gichuhi, A. W. (2008). *Nonparametric changepoint analysis for Bernoulli random variables based on neural networks*. PhD thesis, Technische Universität Kaiserslautern.

Gombay, E. and Horvath, L. (1990). Asymptotic distributions of maximum likelihood tests for change in the mean. *Biometrika*, 77(2):411–414.

Gombay, E. and Horvath, L. (1996). On the rate of approximations for maximum likelihood tests in change-point models. *Journal of Multivariate Analysis*, 56(1):120–152.

Government of Kenya (2020–2021). Presidential addresses and legal notices on covid-19 containment measures. Includes Public Health Act and Public Order Act directives.

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., and Tatlow, H. (2021). A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nature Human Behaviour*, 5(4):529–538.

Hartl, T., Walde, J., and Hohle, M. (2022). Bayesian change point analysis for the impact of COVID-19 policy interventions. *Statistics in Medicine*, 41(15):2798–2815.

Haug, N., Geyrhofer, L., et al. (2020). Ranking the effectiveness of worldwide covid-19 government interventions. *Nature Human Behaviour*, 4(12):1303–1312.

Haynes, K., Eckley, I. A., and Fearnhead, P. (2017). Computationally efficient change-point detection for a range of penalties. *Journal of Computational and Graphical Statistics*, 26(1):134–143.

Horvath, L. and Huskova, M. (2012). Change-point detection in panel data. *Journal of Time Series Analysis*, 33(4):631–648.

Hsiang, S., Allen, D., Annan-Phan, S., Bell, K., Bolliger, I., Chong, T., and Wu, J. (2020). The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature*, 584(7820):262–267.

Kang, J.-H. and Kim, T.-H. (2019). A comparison of kernel density estimation methods for change-point detection. *Journal of Statistical Computation and Simulation*, 89(12):2265–2284.

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

Knoblauch, J. and Damoulas, T. (2023). Scalable bayesian change point detection for count data. *Bayesian Analysis*, 18(2):431–458.

Korkas, K. K. and Fryzlewicz, P. (2017). Multiple change-point detection for non-stationary time series using wild binary segmentation. *Statistica Sinica*, 27(1):287–311.

Küchenhoff, H., Günther, F., Höhle, M., and Bender, A. (2021). Analysis of the early COVID-19 epidemic curve in Germany by regression models with change points. *Epidemiology & Infection*, 149:e29.

Lee, S. and Li, B. (2023). Changepoint detection in overdispersed count data using the generalized Poisson distribution. *Journal of Statistical Computation and Simulation*, 93(4):621–643.

Li, Y., Wang, X., and Wang, H. (2021). Estimating the delay between policy implementation and epidemiological impact with a change point model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(4):972–990.

Liu, S., Zhou, C., and Wang, L. (2022). Support vector regression-based changepoint detection for overdispersed count data. *Computational Statistics & Data Analysis*, 174:107537.

Lord, D., Park, B.-J., and Model, P.-G. (2012). Negative binomial regression models and estimation methods. *Probability Density and Likelihood Functions*.

Loschi, R. H. and Cruz, F. R. (2005). Bayesian identification of multiple change points in Poisson data. *Advances in Complex Systems*, 8(4):465–482.

Ludwig, J. A. and Reynolds, J. F. (1988). *Statistical ecology: A primer in methods and computing*. Wiley.

Luo, Y. and Zhang, H. (2022). Bayesian multiple changepoint detection in overdispersed count data with joint mean-dispersion modeling. *Statistics and Computing*, 32(5):78.

Mazza, A. and Punzo, A. (2011). Discrete beta kernel graduation of age-specific demographic indicators. In Ingrassia, S., Rocci, R., and Vichi, M., editors, *New perspectives in statistical modeling and data analysis*, pages 127–134. Springer.

Meier, A., Kirch, C., and Cho, H. (2021). MOSUM-based change point detection for time series with non-stationary variance. *Journal of Time Series Analysis*, 42(3):312–332.

Ministry of Health Kenya (2020–2021). Covid-19 press releases and situation reports. Official daily reports and policy announcements.

Mundia, S., Gichuhi, A., and Kihoro, J. (2014). The power of likelihood ratio test for a change-point in binomial distribution. *Journal of Agriculture, Science and Technology*, 16(3):105–123.

Nyambura, S., Mundia, S., and Waititu, A. (2016). Estimation of change point in Poisson random variables using the maximum likelihood method. *American Journal of Theoretical and Applied Statistics*, 5(4):219–224.

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

Pavlidis, N. G., Tasoulis, D. K., and Adams, N. M. (2020). Kernel-based change point detection. *Pattern Recognition*, 108:107559.

Pawitan, Y. (2013). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, Oxford, 2 edition.

Scott, S. L. and Varian, H. R. (2014). Predicting the present with Bayesian structural time series. *International Journal of Forecasting*, 30(4):1045–1056.

Severini, T. A. (2021). *Likelihood Methods in Statistics*. Chapman and Hall/CRC, Boca Raton, FL, 2 edition.

Suh, M.-S. and Kim, C. (2015). Change-point analysis of tropical night occurrences for five major cities in Republic of Korea. *Advances in Meteorology*, 2015:742543.

Tariq, A., Lee, Y., and Chowell, G. (2023). Real-time changepoint detection for overdispersed count data: An application to COVID-19 surveillance. *Epidemics*, 42:100667.

Truong, C. et al. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

van den Burg, G. J. J. and Williams, C. K. I. (2020). An evaluation of change point detection algorithms. *arXiv preprint arXiv:2003.06222*.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK.

Villarini, G., Smith, J. A., and Vecchi, G. A. (2013). Changing frequency of heavy rainfall over the central United States. *Journal of Climate*, 26(1):351–357.

Wambui, G. D., Waititu, G. A., and Wanjoya, A. (2015). The power of the pruned exact linear time (PELT) test in multiple changepoint detection. *American Journal of Theoretical and Applied Statistics*, 4(6):581–586.

Wang, T. and Samworth, R. J. (2023). Wild binary segmentation 2.0: A powerful and flexible method for multiple change-point detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 85(2):315–342.

Wasserman, L. (2013). *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York, 2 edition.

Wilken, B. A. (2007). Change-point methods for overdispersed count data. Master's thesis, Air Force Institute of Technology.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, 27(8):1–25.

Zhao, Z. and Yau, C. Y. (2022). Bayesian multiple changepoint detection for overdispersed count data using a negative binomial likelihood. *Journal of Computational and Graphical Statistics*, 31(2):456–469.

Zhu, X., Liu, J., and Li, Y. (2021). Deep learning for changepoint detection in time series. *Neural Networks*, 138:175–188.