Predictive Statistical Modelling Approaches for Events in Discrete Space and Time

A Thesis

submitted to
Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

by

Shubhankar Sahu

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

April, 2024

Supervisor: Prof. Ankur Sinha

Certificate

This is to certify that this dissertation entitled Predictive Statistical Modelling Approaches for Events in Discrete Space and Time towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Shubhankar Sahu under the supervision of Prof. Ankur Sinha, Associate Professor, Operations and Decision Sciences , IIM Ahmedabad , during the academic year 2023-2024.

Prof. Ankur Sinha

Committee:

Prof. Ankur Sinha

Prof. Anindya Goswami

Declaration

I hereby declare that the matter embodied in the report entitled Predictive Statistical Modelling Approaches for Events in Discrete Space and Time are the results of the work carried out by me at the Department of Operations and Decision Sciences , IIM Ahmedabad, under the supervision of Prof. Ankur Sinha and the same has not been submitted elsewhere for any other degree.Wherever others contribute, every effort is made to indicate this clearly, with due reference to the literature and acknowledgement of collaborative research and discussions.

Shubhankar Sahu

Reg. No.-20191090

This thesis is dedicated to my parents and family for their unwavering support and belief in me.

Acknowledgments

I am immensely grateful to Professor Ankur Sinha and Professor Karthik Sriram for their mentorship and invaluable guidance throughout this project. Their continuous support and insightful perspectives have been instrumental in shaping the direction and quality of this thesis. I am deeply indebted to them for dedicating their precious time and sharing their vast knowledge and expertise, which has been pivotal in enriching the depth and rigour of the work. I would like to thank Professor Anindya Goswami whose insightful feedback from time to time have played a crucial role in refining and strengthening this work's theoretical and practical aspects.

I would especially like to thank Sanket for spending time discussing the project and giving me valuable advice and for his continuous motivation. I would also like to thank Ankit for his valuable advice on the project. I would also like to extend my heartfelt appreciation to my family, especially my mother, whose unconditional love, encouragement, and understanding have been an unwavering source of motivation. I would also like to thank my friends at IISER, Samarth, Aniket, Mehul, Varun, Raghav, Rajeet, Vedant, Hrishikesh, and Arindam, for making my time at IISER memorable. I would also like to thank Kapil Chandak, who encouraged me to pursue my thesis at IIM Ahmedabad and has provided support throughout. I'm grateful to Shreyas, Neeraj, and Ashish for our valuable discussions about my project.

Finally, I would like to acknowledge the support and resources provided by IIM Ahmedabad,and IISER Pune which have played a vital role in facilitating this project and enabling its successful completion.

Abstract

This thesis investigates the use of predictive modelling to improve decision-making for events that are distributed over discrete space and occur in discrete time. We explore three main statistical methods: Kernel Density Estimation (KDE), Bayesian methods, and Maximum Likelihood Estimation (MLE), known for their effectiveness in analyzing data and applicability to continuous and discrete time scenarios. A key contribution is developing an advanced KDE model that features adaptive bandwidth and weighted approaches for more accurate density estimates. We focus our modelling approaches on coal production data from various mines, we demonstrate how these predictive models can address uncertainties and periodic trends in mine output. Accurately forecasting mine production can enhance resource allocation and inventory management. The thesis further extends to prescriptive analytics by incorporating mathematical optimization to allocate coal optimally from mines to power plants. This involves formulating and solving a mathematical model with real-world constraints to find the best allocation solution. By combining predictive and prescriptive analytics, this research provides a comprehensive approach to solving complex problems in the mining industry, aiming to improve resource allocation, inventory management, and overall operational efficiency. The findings have the potential to significantly impact the sectors where future events can occur only in discrete space and time, leading to more informed and data-driven decisions.

Contents

Chapter 1

Introduction

Predictive modelling applies both statistical techniques and machine learning algorithms to predict outcomes. It is often used to predict future events based on historical data, but it can also be applied to any type of unknown event, regardless of when it occurred. For example, predictive models are often used to detect crimes and identify suspects after the crime has taken place. Predictive modelling involves formulating mathematical frameworks that identify patterns and relationships within data, thereby enabling the prediction of outcomes. Finally, various performance metrics are calculated to validate the accuracy and effectiveness of the predictive models.

Predictive modelling is relevant across a broad spectrum of industries, leaving virtually no sector unaffected by its applications. In the financial industry, it is instrumental for assessing credit risk, forecasting market trends, and identifying fraudulent transactions, thereby safeguarding financial stability (Khandani et al., 2010) (8). Healthcare benefits from predictive models through disease outbreak prediction, patient outcome forecasting, contributing significantly to public health and efficiency (Ijaz et al.,2020) (9) . In agriculture, predictive modelling aids farmers by forecasting crop yields, thereby facilitating more informed decisions regarding resource allocation, irrigation, and pest management (Satpathi et al.,2023) (10) Marketing and retail sectors utilize these models to decode consumer behaviour, optimize pricing strategies, and enhance supply chain management, driving sales and customer satisfaction .

Predictive Modelling and Probability Density estimation are very similar. While density

estimation techniques often involve finding the parameters (e.g., mean, variance) of a probability distribution that best fit the observed data and capture its underlying uncertainty and variability, The primary goal of predictive modelling is to predict new outcomes based on patterns found in historical data by incorporating multiple and complex relationships between variables and. For the scope of our thesis, we treat both of them the same and try to estimate the underlying probability distribution of the data for prediction purposes.

This thesis focuses on three main statistical approaches: Kernel Density Estimation (KDE), Bayesian methods, and Maximum Likelihood Estimation (MLE). These methodologies stand out for their robustness in capturing complex relationships within data and their adaptability to continuous and discrete time scenarios. Kernel Density Estimation is a nonparametric density estimation technique that offers a flexible and data-driven approach to estimating the underlying distribution without making strong assumptions about its shape.(Silverman, 1986)(1) (Wand and Jones, 1994)(2)

In recent years, there has been increasing use of Bayesian methods for density estimation primarily because of their ability to handle complex data structures and incorporate prior knowledge with observed data, enhancing model accuracy over time, a feature especially beneficial in dynamic systems (Gelman et al., 1995) (11).MLE provides a framework for estimating model parameters that maximize the likelihood of observing the given data, ensuring precision in predictions

Central to the study is developing and applying a novel KDE model in discrete space and time that incorporates adaptive bandwidth and assigned weights approaches, ensuring more precise and contextually relevant density estimates. This model illustrates how we can blend traditional statistical techniques with innovative adaptations, catering to the specificities of real-world data.

We explore the theoretical framework of these techniques and examine their practical implementations by considering a specific case of coal production from mines. These coal mines are distributed spatially over discrete space, and we are trying to model their activity, or the amount of coal they produce each day, which is also in discrete quantities. The production patterns of mines are often complex, sometimes periodic and exhibit inherent uncertainties, making it challenging to rely on traditional parametric models that assume specific distributional forms. Accurate forecasting of mine production can aid in efficient planning, resource allocation, and decision-making processes. One can extend this study to

several other types of data, which consist of events happening that are distributed in discrete space and time.

The final component of this thesis involves mathematical optimization. After obtaining probabilities of coal production from mines using predictive modelling techniques, the focus shifts to prescriptive analytics and determining the efficient allocation of coal from these mines to different electricity-generating power plants. Mathematical optimization is a powerful tool that allows finding the best solution to a problem, subject to various constraints representing real-world limitations or requirements. In this case, the objective is to optimize the allocation of coal from mines to power plants in the most efficient manner possible.

The optimization problem is typically formulated as a mathematical model, often involving linear or nonlinear programming techniques. Various optimization algorithms and solvers are then employed to find the optimal solution that maximizes or minimizes the objective function while satisfying all the specified constraints. These algorithms may use techniques such as simplex methods, interior point methods, or metaheuristics like genetic algorithms or simulated annealing, depending on the complexity and nature of the problem. Our thesis does not discuss the algorithms used to solve this optimisation problem.

By incorporating predicted coal production probabilities from the predictive modelling component, the optimization model can account for uncertainties and variabilities in the supply side. This can lead to more robust and reliable coal allocation plans that adapt to potential disruptions or fluctuations in mine production. However, the optimisation model is still simple and not subjected to complex constraints. A more robust framework involves stochastic optimisation for handling uncertainties, which is beyond the scope of this thesis.

Ultimately, the mathematical optimization component of this thesis aims to facilitate more efficient resource allocation, improved inventory management, and better decision-making processes in the mining industry by developing accurate and robust models that consider both predictive analytics and prescriptive analytics.

Chapter 2

Density Estimation

We introduce the basic preliminaries for understanding probability density estimation in this chapter. This also describes the existing paramteric and non paramteric approaches for density estiation.

2.1 Probability density functions

Definition 1(Probability Density Function): For a continuous random variable X, its probability density function (pdf) $f(x)$ is characterized as the derivative of the cumulative distribution function (CDF) $F(x)$, denoted by:

$$f(x) = \frac{d}{dx}F(x). \tag{2.1}$$

Proposition 1: Given a pdf $f(x)$ for a random variable X, the CDF $F(x)$ can be obtained through integration:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt. \tag{2.2}$$

Theorem 1 (Properties of pdf): The pdf $f(x)$ of a random variable X satisfies the following properties:

1. Non-negativity: $f(x) \geq 0$ for all x.

2. Normalization: $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

The probability that X falls within the interval $a < X \leq b$ is found using $f(x)$:

$$\Pr(a < X \leq b) = \int_a^b f(x)\,dx = [F(x)]_a^b. \tag{2.3}$$

2.2 The Empirical Distribution Function

Definition 2 (Empirical Distribution Function): The empirical distribution function (EDF) for a sample $X_1, \ldots, X_n$ is defined as:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x), \tag{2.4}$$

where I is the indicator function, denoting 1 if $X_i \leq x$ and 0 otherwise.

The EDF $\hat{F}(x)$ serves as a non-parametric estimator of the CDF $F(x)$, placing equal weight $\frac{1}{n}$ at each observation X_i. This estimator is pivotal in understanding the distribution of data without assuming a specific underlying probability distribution.

The CDF $F(x)$ encapsulates the entirety of a random variable's probability structure, applicable across both continuous and discrete scenarios. Unlike the PDF or PMF, which are specific to the nature of the variable, the CDF remains a universal descriptor. Moreover, there are cases where neither PDF nor PMF exist.

2.3 Density estimation Approaches

Suppose we have a set of observed data points $X_1, \ldots, X_n$ which are assumed to be a sample from an unknown probability density function. Density estimation is the problem of reconstructing the probability density function using the given data points. There are two major approaches towards it.

1. **Parametric Estimation**
   This approach assumes that the data is drawn from a known parametric family of probability distributions. The underlying density of the data is then usually estimated by the various known estimates of parameters.

2. **Non-Parametric Estimation**
   The basic idea of the nonparametric approach is to use data to infer the underlying distributions, and it does not assume any specific form of the distribution.

Nonparametric density estimation is a cornerstone of statistical analysis, providing a method to estimate the underlying probability density function of a dataset without assuming a specific parametric form. Unlike parametric methods, which presuppose a particular distribution shape (e.g., normal, exponential), nonparametric methods adapt to the data's intrinsic structure. This flexibility makes them invaluable for exploratory data analysis, especially when little is known about the data's underlying distribution.

2.4 Non parametric Approaches

2.4.1 Histogram

The histogram is one of the most basic and intuitive methods for density estimation. It involves partitioning the range of data into a set of contiguous intervals (or bins) and then counting the number of data points that fall into each bin. Mathematically, the histogram estimator $\hat{f}$ is defined as:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} I(x - X_i \in [0, h)). \tag{2.5}$$

Where: n is the number of data points, h is the width of the bins, $I(\cdot)$ is the indicator function (which is 1 if the condition inside is true, and 0 otherwise), and X_i are the data points. The choice of bin width h is crucial. A smaller h results in a "spikier" histogram that might capture noise, while a larger h might oversmooth the data.

2.4.2 The Naive Estimator

The naive estimator is another simple method, which places a rectangular "pulse" of width $2h$ and height $\frac{1}{2nh}$ centred at each data point. The mathematical formulation for the naive estimator is:

$$\hat{f}(x) = \frac{1}{2nh} \sum_{i=1}^{n} I(|x - X_i| < h). \tag{2.6}$$

Where n is the number of data points, h is half the width of the rectangular pulse, $I(\cdot)$ is the indicator function, and X_i are the data points. Like the histogram, the naive estimator's appearance is influenced by the choice of h. A smaller h may make the estimate too rough, while a larger h may over smooth it.

2.4.3 The Nearest Neighbour Method

The nearest neighbour class of estimators represents an attempt to adapt the amount of smoothing to the "local" density of data. The degree of smoothing is controlled by an integer k, which is considerably smaller than the sample size; typically $k \approx n^{1/2}$. Define the distance $d(x, y)$ between two points on the line to be $|x - y|$ in the usual way, and for each t define

$$d_1(t) \leqslant d_2(t) \leqslant \ldots \leqslant d_a(t)$$

to be the distances, arranged in ascending order, from t to the points of the sample. The kth nearest neighbour density estimate is then defined by

$$\hat{f}(t) = \frac{(k-1)}{2nd_k(t)}.$$

2.4.4 The Kernel Estimator

The kernel estimator refines the idea behind the naive estimator by using a smooth, non-negative function K (called the kernel) instead of a simple rectangular pulse. This results in a smoother and more adaptable estimate. Mathematically, the kernel estimator $\hat{f}$ is given

by:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{2.7}$$

Where n is the number of data points, h is the bandwidth (which determines the width or scale of the kernel), and K is the kernel function. The bandwidth h plays a crucial role: a small h captures more granular details, potentially leading to overfitting, while a large h may oversmooth the data and affecting the estimation's sensitivity to the data points. We adopt Silverman's rule of thumb for bandwidth selection, which is designed to minimize the mean integrated squared error (MISE) of the density estimate.

Silverman's Rule of Thumb

For a univariate density estimation, the optimal bandwidth h under the assumption that the data points being analyzed are distributed approximately normally (i.e., they follow a Gaussian distribution) is given by (Silverman, 1986)(1):

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-\frac{1}{5}}. \tag{2.8}$$

where: h is the bandwidth, $\hat{\sigma}$ is the standard deviation of the sample, n represents the size of the sample (5). A bandwidth value h is more robust when it enhances the estimation for distributions that exhibit significant skewness or have long tails. Such an improvement can be achieved by empirical means, which involves substituting the standard deviation $\hat{\sigma}$ with a parameter A as illustrated below:

$$A = \min\left(\frac{I\hat{Q}R}{\hat{\sigma}}, \frac{1}{1.34}\right), \tag{2.9}$$

where $I\hat{Q}R$ represents the estimated interquartile range.

The coefficient typically set at 1.06 can be adjusted to 0.9 to refine the model further.

Consequently, the revised bandwidth equation is:

$$h = 0.9 \min \left(\frac{I\hat{Q}R}{\hat{\sigma}}, \frac{1}{1.34} \right) n^{-\frac{1}{5}}. \tag{2.10}$$

This method, often referred to as the *normal distribution approximation* or the Gaussian approximation, is synonymous with *Silverman's rule of thumb*.This formula balances the trade-off between bias and variance in the estimation process, aiming for a density estimate that is neither too smooth (underfitting) nor too rough (overfitting).

Chapter 3

Bayesian Inference

Bayesian inference is a powerful statistical approach that provides a coherent framework for incorporating prior knowledge and observed data to estimate unknown parameters or make predictions. It is founded on Bayes' theorem, a fundamental result in probability theory that governs the process of updating beliefs based on evidence.

Before going into the specifics of Bayesian inference, it is essential to establish some preliminary concepts and notation. Let Θ denote the parameter space and $\theta \in \Theta$ be the unknown parameter(s) of interest. Furthermore, let $\mathcal{X}$ represent the sample space, and $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ be the observed data consisting of n independent and identically distributed (i.i.d.) samples from an underlying probability distribution $f(\mathbf{x} \mid \theta)$.

3.1 Bayesian Inference Components

The Bayesian inference framework comprises three fundamental components:

3.1.1 Prior Distribution

The prior distribution, denoted by $\pi(\theta)$, encapsulates our a priori knowledge or beliefs about the parameter(s) θ before observing any data. It quantifies the initial uncertainty or infor-

mation regarding the parameter(s) of interest. The choice of prior distribution can be based on subjective beliefs, previous studies, or established conventions within the field.

3.1.2 Likelihood Function

The likelihood function, denoted by $L(\mathbf{x} \mid \theta)$, represents the probability of observing the data $\mathbf{x}$ given the parameter(s) θ. It is a function of θ and provides information about the plausibility of different parameter values based on the observed data. The likelihood function is typically derived from the assumed probability distribution $f(\mathbf{x} \mid \theta)$.

3.1.3 Posterior Distribution

The posterior distribution, denoted by $\pi(\theta \mid \mathbf{x})$, is the updated distribution of the parameter(s) θ after incorporating the observed data $\mathbf{x}$. It combines the prior knowledge encoded in the prior distribution and the information from the data through the likelihood function.

3.2 Bayes' Theorem

Bayes' theorem lies at the heart of Bayesian inference and provides a principled way to calculate the posterior distribution from the prior distribution and the likelihood function. It is given by:

$$\pi(\theta \mid \mathbf{x}) = \frac{L(\mathbf{x} \mid \theta)\pi(\theta)}{m(\mathbf{x})} \quad . \tag{3.1}$$

where $m(\mathbf{x})$ is the marginal distribution of the data, also known as the normalizing constant, and is defined as:

$$m(\mathbf{x}) = \int_{\Theta} L(\mathbf{x} \mid \theta)\pi(\theta)\, d\theta \quad . \tag{3.2}$$

The marginal distribution $m(\mathbf{x})$ ensures that the posterior distribution $\pi(\theta \mid \mathbf{x})$ is a proper probability density function that integrates to unity over the parameter space Θ.

In the subsequent sections, we will explore various aspects of Bayesian inference, including prior selection, conjugate priors, computational techniques for posterior estimation, and applications in predictive modeling.

3.3 Derivation of the Posterior Distribution

The posterior distribution $\pi(\theta \mid x)$ can be derived using Bayes' theorem by applying the concept of "Posterior $\propto$ Likelihood $\times$ Prior."

$$\pi(\theta \mid x) \propto L(\theta \mid x)\pi(\theta) \quad . \tag{3.3}$$

This proportionality holds because the normalizing constant $m(x)$ does not depend on θ and can be treated as a constant when updating the prior distribution using the likelihood function.

To obtain the full posterior distribution, we need to normalize the right-hand side of the above equation so that it integrates to 1 over the parameter space:

$$\pi(\theta \mid x) = \frac{L(\theta \mid x)\pi(\theta)}{\int L(\theta \mid x)\pi(\theta)\, d\theta} \quad . \tag{3.4}$$

The denominator in the above equation is the marginal distribution $m(x)$, which acts as a normalizing constant to ensure that the posterior distribution is a valid probability density function (PDF) or probability mass function (PMF), depending on the nature of the parameter(s) θ.

If the prior distribution and the likelihood function belong to the same family of distributions (e.g., both are Normal distributions or both are Gamma distributions), the resulting posterior distribution will also belong to that family. This property, known as conjugacy, simplifies the calculation and interpretation of the posterior distribution.

In cases where conjugacy does not hold, numerical methods or approximations may be required to obtain the posterior distribution.

3.4 Bayesian Updating and Sequential Learning

Bayesian inference allows us to sequentially update the posterior distribution of a parameter as new data becomes available. This process is a key aspect of Bayesian inference and is particularly useful in scenarios where data comes in incrementally.

Initial Posterior with x_1 Only:

Given the likelihood $L(x_1 \mid \theta)$ and a prior distribution $\pi(\theta)$, the posterior distribution of θ after observing x_1 alone is given by Bayes' theorem:

$$\pi(\theta \mid x_1) = \frac{L(x_1 \mid \theta)\pi(\theta)}{m(x_1)}. \tag{3.5}$$

where $m(x_1)$ is the marginal likelihood of x_1, which can be calculated as:

$$m(x_1) = \int L(x_1 \mid \theta)\pi(\theta)\, d\theta. \tag{3.6}$$

Update with x_1, x_2 Using the Previous Posterior:

Now, suppose you obtain an additional observation x_2. You want to update your belief about θ using the new data x_2 and the previous posterior $\pi(\theta \mid x_1)$ as the new prior. Apply Bayes' theorem again:

$$\pi(\theta \mid x_1, x_2) = \frac{L(x_2 \mid \theta, x_1)\pi(\theta \mid x_1)}{m(x_2 \mid x_1)}. \tag{3.7}$$

And $m(x_2 \mid x_1)$, the marginal likelihood of x_2 given x_1, is calculated as:

$$m(x_2 \mid x_1) = \int L(x_2 \mid \theta)\pi(\theta \mid x_1)\, d\theta. \tag{3.8}$$

This procedure allows you to sequentially update the posterior distribution of θ as you receive more data, using the posterior from the previous step as the prior for the next step. This process demonstrates the recursive nature of Bayesian updating, where the posterior from previous data becomes the prior for the next update.

3.5 Known Families of Distributions

In Bayesian inference, certain families of distributions play an important role due to their mathematical properties and the ease of computation they provide. These families of distributions are known as conjugate prior distributions. When the prior distribution and the likelihood function belong to the same family of distributions, the resulting posterior distribution will also belong to that family, simplifying the calculation and interpretation of the posterior distribution. Some commonly used conjugate prior distributions and their corresponding likelihood functions are as follows:

3.5.1 Normal-Normal Conjugate Family

Suppose the likelihood function is a Normal (Gaussian) distribution with known variance σ^2, and the prior distribution on the mean μ is also a Normal distribution. In that case, the posterior distribution for μ will also be a Normal distribution.

3.5.2 Gamma-Poisson Conjugate Family

Suppose the likelihood function is a Poisson distribution, and the prior distribution on the rate parameter λ is a Gamma distribution. In that case, the posterior distribution for λ will also be a Gamma distribution.This will also work incase of likelihood is Exponential distribution.

Table 3.1: Some common prior -likelihood conjugate pairs

Prior	Likelihood	Posterior Prameters
Gamma$(\lambda;\alpha,\beta)$ $p(\lambda)=\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$	Exponential$(x;\lambda)$ $p(x)=\lambda e^{-\lambda x}; x\in[0,\infty)$	$\alpha+n, \beta+\sum_{i=1}^n x_i$
Gamma$(\lambda;\alpha,\beta)$ $p(\lambda)=\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$	Poisson$(x;\lambda)$ $p(x)=\frac{\lambda^x e^{-\lambda}}{x!}; x\in\{0,1,...\}$	$\alpha+\sum_{i=1}^n x_i, \beta+n$
Beta$(\theta;a,b)$ $p(\theta)=C\theta^{a-1}(1-\theta)^{b-1}$ $C=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$	Bernoulli$(x;\theta)$ $p(x=1)=\theta=1-p(x=0);$ $x\in\{0,1\}$	$a+\sum_{i=1}^n x_i,$ $b+n-\sum_{i=1}^n x_i$
Normal$(\mu;\mu_0,\sigma_0^2)$ $p(\mu)=\frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$	Normal$(x;\mu,\sigma^2)$ $p(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}; x\in\mathbb{R}$	$\mu_n=\frac{\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n x_i}{\sigma^2}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}, \sigma_n^2=\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$

In all cases in the table, the data is assumed to consist of n points $x_1, ..., x_n$.

3.5.3 Beta-Binomial Conjugate Family

If the likelihood function is a Binomial distribution, and the prior distribution on the success probability p is a Beta distribution, then the posterior distribution for p will also be a Beta distribution.

These conjugate families simplify the calculations required to obtain the posterior distribution, as the posterior distribution can be derived analytically and often has a closed-form expression. This property makes Bayesian inference more tractable in many practical situations.

However, it is important to note that conjugacy is not always necessary for Bayesian inference. In cases where the prior distribution and the likelihood function do not belong to a conjugate family, numerical methods or approximations may be required to obtain the posterior distribution.

Chapter 4

Clustering

Sometimes, when estimating the underlying probability distribution of events distributed over spatial and time coordinates, it might not be easy to estimate future events with higher accuracy by employing a single probability distribution function for the entire dataset, especially when dealing with large datasets. Therefore, for ease of modelling purposes, we try to individually group the spatially distributed observations into different clusters using some algorithm and then estimate the underlying probability function confined to that particular cluster.

This is particularly helpful in our case, where we can group the mines into various clusters using their latitude and longitudes and then estimate the probability distribution individually in those clusters. We employ the K-means algorithm for our clustering purposes, which is described in detail in this chapter.

4.1 K-means Clustering Algorithm

In the K-means approach, we classify n spatial points representing mines, denoted by $\{x_i\}_{i=1}^{n}$ in $\mathbb{R}^2$. The objective is to identify k central points, or centroids denoted by $\{c_j\}_{j=1}^{k}$, that minimize the cumulative distance from each mine to its closest centroid. (6) These centroids, alongside their corresponding clusters $\{\Gamma_j\}_{j=1}^{k}$, are chosen so that each cluster $\{\Gamma_j\}_{j=1}^{k}$ are mutually disjoint and $\bigcup_{j=1}^{k} \Gamma_j = [n]$. The total distances is expressed as :

$$f(c_j, \Gamma_j) = \sum_{j=1}^{k} \sum_{i \in \Gamma_j} d(x_i, c_j). \tag{4.1}$$

This function is known as the K-means cost function, and it depends on the centroids and their clusters. The term $d(x, y)$ represents a distance metric, quantifying the similarity between two points x and y. For distance measurement, we use the Euclidean distance between two points in our case. For any two points x_i and x_j, with coordinates (x_{i1}, x_{i2}), and (x_{j1}, x_{j2}) ,the Euclidean distance between them is given by:

$$d(x_i, x_j) = \left(\sum_{l=1}^{d} (x_{il} - x_{jl})^2 \right)^{\frac{1}{2}}. \tag{4.2}$$

The K-Means algorithm aims to minimize the within-cluster sum of squares (WCSS), which is the sum of squared Euclidean distances between each mine and its nearest centroid. The objective function J to be minimized is:

$$J = \sum_{j=1}^{k} \sum_{i \in \Gamma_j} \|x_i - c_j\|^2. \tag{4.3}$$

where: k denotes the number of clusters, Γ_j represents the set of points, specifically mines, that are assigned to the jth cluster, x_i signifies the position of the ith mine, c_j indicates the centroid of the jth cluster, and $\| \cdot \|$ corresponds to the Euclidean norm.

The algorithm proceeds through iterative steps, denoted by t, as follows:

1. **Initialization**: k initial centroids, $\{c_j^{(0)}\}_{j=1}^{k}$, are chosen, typically at random from the set of mine locations.

2. **Assignment**: The clusters at any iteration t are defined by a collection of sets $\Gamma_j^{(t)}$, each corresponding to a cluster j, with j ranging from 1 to k. A cluster $\Gamma_j^{(t)}$ is essentially a subset of the mine locations which are closer to the centroid $c_j^{(t)}$ of cluster j than to any other centroid $c_k^{(t)}$ where k is different from j. This is formalized by:

$$\Gamma_j^{(t)} := \{i : \|x_i - c_j^{(t)}\| \leq \|x_i - c_k^{(t)}\|, \forall k \neq j\}. \tag{4.4}$$

where $c_l^{(t)}$ is the centroid of cluster l at iteration t.

3. **Update**: The centroids are dynamic, recalculated at each iteration to reflect the current assignment of mine locations to clusters. The recalculation aims to minimize the within-cluster sum of squared Euclidean distances, and by definition for cluster j at iteration t, the new centroid $c_j^{(t+1)}$ is updated as:

$$c_j^{(t+1)} := \arg\min_{c_j \in \mathbb{R}^d} \sum_{i \in \Gamma_j^{(t)}} \|x_i - c_j\|^2. \tag{4.5}$$

where $\Gamma_j^{(t)}$ is the set of mines assigned to centroid c_j during iteration t. This minimizer has a closed-form solution as proved below, which is exactly the sample mean of data points in $\Gamma_j^{(t)}$, i.e.,

$$c_j^{(t+1)} = \frac{1}{|\Gamma_j^{(t)}|} \sum_{i \in \Gamma_j^{(t)}} x_i. \tag{4.6}$$

Each $\Gamma_j^{(t)}$ can be visualized as a Voronoi cell in a Voronoi diagram, representing all the mine locations that are closer to the centroid $c_j^{(t)}$ than to any other centroid.

The K-means algorithm proceeds through several such iterations of assignment and updates until the centroids $\{c_j^{(t)}\}$ stabilize, indicating that further iterations do not yield significant changes in the assignment of mines to clusters. The stability of centroids indicates convergence of the algorithm.

Lemma 1 : The optimal centroid for each cluster in the K-means algorithm is the sample mean of all points within the cluster.

Proof : Let's denote a cluster at iteration t as $\Gamma_j^{(t)}$, and the corresponding centroid as $c_j^{(t)}$.

Given the K-means objective function:

$$J = \sum_{j=1}^{k} \sum_{i \in \Gamma_j} \|x_i - c_j\|^2 \tag{4.7}$$

We seek to find the $c_j^{(t+1)}$ that minimizes J for the set of points $\{x_i\}_{i \in \Gamma_j^{(t)}}$.

Consider the cost function of assigning a set of points to a centroid z:

$$\text{cost}(C; z) = \sum_{x \in C} \|x - z\|^2 \tag{4.8}$$

The task is to find a z that minimizes this cost. The derivative of the cost function with respect to z is:

$$\frac{\partial}{\partial z}\text{cost}(C; z) = \frac{\partial}{\partial z} \sum_{x \in C} \|x - z\|^2 = -2 \sum_{x \in C} (x - z) \tag{4.9}$$

Setting the derivative equal to zero for optimality:

$$-2 \sum_{x \in C} (x - z) = 0 \Rightarrow \sum_{x \in C} x = |C| \cdot z \tag{4.10}$$

Solving for z:

$$z = \frac{1}{|C|} \sum_{x \in C} x \tag{4.11}$$

This indicates that the cost is minimized when z is the mean of all points in C. Translating this result to our cluster $\Gamma_j^{(t)}$, we get that the new centroid $c_j^{(t+1)}$, which minimizes the within-cluster sum of squares, is the mean of the points in the cluster:

$$c_j^{(t+1)} = \frac{1}{|\Gamma_j^{(t)}|} \sum_{i \in \Gamma_j^{(t)}} x_i. \tag{4.12}$$

4.1.1 The Elbow Method

As we have discussed in the K-means algorithm for the initialisation step, we need to mention the number of clusters beforehand, according to which the algorithm will divide the data points into the specified number of clusters. The Elbow Method is a widely used heuristic in k-means clustering to determine the optimal number of clusters k for a given dataset.(7)

The method involves running the clustering algorithm for a range of k values and computing the within-cluster sum of squares (WCSS) for each value of k. A plot of WCSS versus the number of clusters typically reveals a sharp decrease in WCSS as k increases, but this

rate of decrease slows down significantly after a certain point, creating an "elbow" in the graph. The optimal number of clusters is often considered at this elbow point, where the reduction in WCSS starts diminishing. Mathematically, one might seek the point k at which the second derivative of the WCSS with respect to k is maximized, indicating a maximum curvature on the plot, or more informally, where the rate of decrease shifts from being rapid to more gradual.

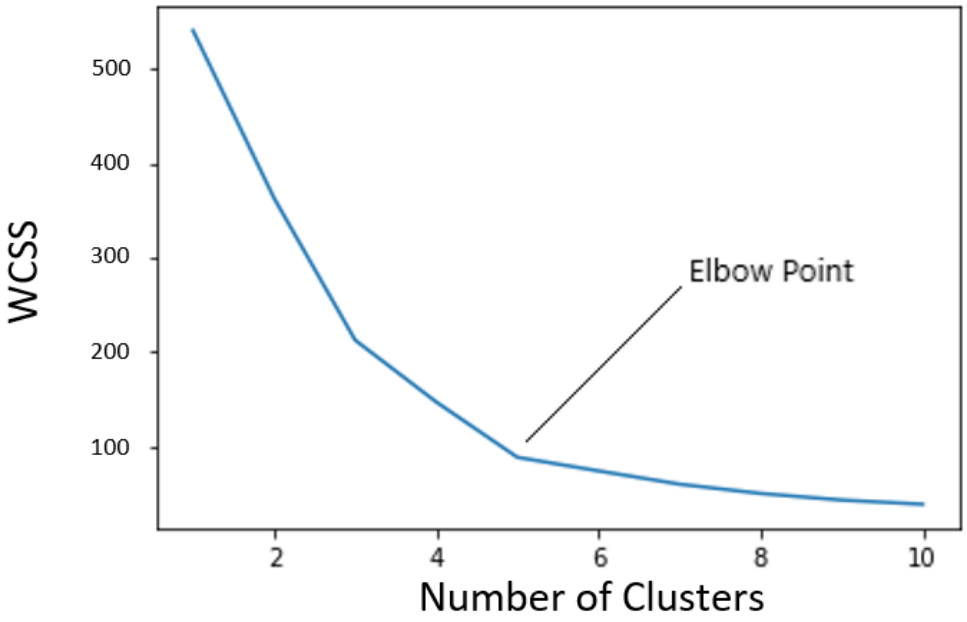

Figure 4.1: The Elbow Method: WCSS vs. Number of Clusters. The "elbow" point is indicated by the number of clusters beyond which the WCSS begins to decrease at a slower rate.

However, as recent studies (Schubert 2023)(4) have shown, the Elbow method is both subjective and unreliable, particularly in cases where there is no such "clear elbow." and alternate heuristics such as the variance-ratio-criterion or the average silhouette width are employed which is beyond the scope of this thesis.

Chapter 5

Maximum Likelihood Estimation

After the observed data points, which are distributed spatially, are grouped into various clusters, we estimate the underlying probability distribution of each cluster separately. This chapter describes our first statistical model to estimate the probability distribution using the Maximum Likelihood Estimation approach.

Our proposed model that will be used for prediction purposes is structured into two main components::
1. The first part of the model will forecast the number of supplies that will be produced for each identified cluster each day based on data available for previous days.
2. The second part of the model will focus on the spatial distribution of the supplies within each cluster.

5.1 Likelihood Fucntion

Definition 1:*Likelihood Function* : The likelihood function, represented as $\mathcal{L}(\theta|x)$, is a function of the parameter(s) θ given the data x (12). It is proportional to the probability of the data given those parameter values. Specifically, for a sample of data points $X = (X_1, \ldots, X_n)$, the likelihood function given an observation x is expressed as:

$$\mathcal{L}(\theta|x) = f(x|\theta). \tag{5.1}$$

where $f(x|\theta)$ is the joint probability distribution function (pdf) for continuous variables or the probability mass function (pmf) for discrete variables.

For a discrete dataset, the likelihood function can be interpreted as the probability of observing the data given specific values of θ, such that $\mathcal{L}(\theta|x) = P(X = x|\theta)$. In the process of comparing likelihoods for different parameter values, if we find that $\mathcal{L}(\theta_1|x) > \mathcal{L}(\theta_2|x)$, it suggests that the observed data is more likely under the parameter values θ_1 than under θ_2. This comparison forms the basis of the likelihood principle, which states that among the considered parameter values, θ_1 is a better estimate for the true parameter value than θ_2.

When $X_1, \ldots, X_n$ constitute a sample drawn independently and identically from a population characterized by a probability density function (pdf) or probability mass function (pmf) $f(x|\theta_1, \ldots, \theta_k)$, the likelihood function is defined by:

$$L(\theta|x) = L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k). \tag{5.2}$$

Definition 2:*Maximum Likelihood Estimator* : For each sample point x, let $\hat{\theta}(x)$ denote the parameter value at which $L(\theta|x)$ attains its maximum as a function of θ, with x remaining constant. A maximum likelihood estimator of the parameter θ based on a sample X is $\hat{\theta}(X)$.

So, suppose we try to find the maximum likelihood estimates of the number of supplies given historical data from each identified cluster. In that case, we will be able to identify the parameter values of the probability distribution from which the number of supplies is coming from and hence the probability distribution of each cluster. Note that the number of supplies from each cluster is a discrete dataset. Also, a preliminary analysis of the number of supplies from each cluster reveals its periodic nature.

5.2 Proposed Model

Consider a cluster denoted by C, comprising of N mines. The observed discrete data represented as $X_1, X_2, \ldots, X_k$ captures the daily number of supplies from C, i.e. the number of mines from cluster C that had production over a span of k days. We assume that each

day's daily production is independent of any other day, i.e. $X_1, X_2, \cdots, X_n$ are assumed to be independent and identically distributed random variables.

Since the daily number of supplies from each cluster is a count data, we can assume this follows a Poisson distribution, characterised by a parameter λ_i, which varies according to the cluster. This variation embeds the model's adaptability, expressed as:

$$X_i \sim \text{Poisson}(\lambda_i) \quad \text{with} \quad \lambda_i = \lambda\omega\phi(\omega i), \tag{5.3}$$

where λ is an average rate parameter and the circular probability density function ϕ, evaluated at ωi.

This "circular Poisson model" is designed to handle data with a periodic structure. Traditional Poisson models are more linear and might not be the best fit for data that exhibits cyclic patterns, such as events that occur more frequently at certain times of the day or during specific seasons of the year.In this model, the intensity λ can vary depending on the position within the cycle, allowing it to capture variations in event frequencies as they change over the course of the cycle.

5.2.1 Circular Probability Density Function

The probability density function (pdf) ϕ is defined on the interval $(0, 2\pi)$. For modelling purposes, we employ the cardioid function as the circular pdf $\phi(\theta)$:

$$\phi(\theta, \rho) = \frac{1 + 2\rho\cos(\theta)}{2\pi}, \tag{5.4}$$

where θ is a phase term, and ρ denotes a modulation factor. This formulation allows the model to capture and adapt to cyclical variations in mine production, informed by daily observations. In our study, $\omega = \frac{2\pi}{u}$, where $u \in \{1, 2, 3, 4, \ldots, 50\}$, and this range can be extended up to 365 if necessary.

Cardioid function typically refers to a mathematical function that describes the shape of a cardioid—a heart-shaped curve. The standard cardioid is described by the parametric equations:

$$x(\theta) = r(1 - \sin(\theta))\cos(\theta) \tag{5.5}$$

$$y(\theta) = r(1 - \sin(\theta))\sin(\theta). \tag{5.6}$$

Instead of using standard trigonometric functions (like sine and cosine) alone, the cardioid function combines them in a specific way to offer a more flexible representation of cycles. This flexibility can be especially beneficial when the cyclic patterns in the data aren't perfectly sinusoidal or when they exhibit certain asymmetries.

The expression $\phi(\omega \cdot i)$ indicates that the phase angle ϕ at time i depends on the angular frequency ω and the specific time or position i. This allows the model to account for variations in the phase as time progresses, making it possible to capture cyclic patterns that change over time.

5.2.2 Likelihood Function

Since the number of supplies is iid Poisson random variables,the likelihood function is expressed as a product of the observed data points, leading to the following expression:

$$L(\lambda_i, \omega, \rho | X_1, X_2, \cdots X_n) = \prod_{i=1}^{n} \text{Poisson}(X_1, X_2, \cdots X_n; \lambda_i). \tag{5.7}$$

with the Poisson parameter λ_i described as

$$\lambda_i = \lambda\omega\phi(\omega i).$$

The expression is proportional to:

$$\prod_{i=1}^{n} \frac{\lambda_i^{x_i} e^{-\lambda_i}}{x_i!} \propto \lambda^{\sum x_i} \prod_{i=1}^{n} (\omega\phi(\omega i))^{x_i} \times e^{-\lambda\omega\Sigma\phi(\omega i)} \tag{5.8}$$

where λ is an average rate parameter, and n is the total number of days.

Our main objective is determining the Maximum Likelihood Estimators (MLE) for λ and ρ. We seek these estimators within the set of possible values for ω, specifically $\omega \in \left\{ \frac{2\pi}{u} \,\middle|\, u = (1, 2, \ldots, 100) \right\}$. To determine the MLE of λ, we can fix ω and perform the estimation.

Optimisation using Scipy's Minimize Function

To obtain the optimal values of λ and ρ, we utilise the 'minimise' function from Scipy's 'optimise' module. This function helps minimise the negative log-likelihood, aiding us in finding the parameters' Maximum Likelihood Estimates (MLE). The bounds ensure that λ is non-negative and ρ is between -1 and 1.

By employing the MLE approach and leveraging the Poisson distribution with a periodic component, we can model the production of mines within a cluster and predict their future output. This model assumes that the daily production follows a periodic pattern, which can be captured by the circular probability density function $\phi(\theta, \rho)$. The Maximum Likelihood Estimators for the parameters are obtained through optimisation techniques, ensuring the best fit for the observed data.

It's important to note that this model assumes independence among the daily productions within the cluster, which may or may not hold in practice. Additionally, choosing the cardioid function as the circular pdf is a modelling assumption.

The Maximum Likelihood Estimate will give us the number of mines producing within a particular cluster. Given this number, we are now interested in which specific mines will have production within that cluster by spatial analysis. However, as we can see from the results, this approach gives absurd results for the number of mines. Hence, we must devise an alternative approach to obtain the number of mines before proceeding to spatial analysis.

Chapter 6

Bayesian Model

In the previous chapter, we discussed the Maximum Likelihood Estimation (MLE) approach for estimating the parameters of the circular Poisson model and essentially the number of supplies from a given cluster that did not perform well. In this chapter, we describe our second statistical modelling approach using Bayesian estimation techniques.

Let's consider a cluster C containing N mines, and the observed data from these mines is represented as $X_1, X_2, \ldots, X_n$, where X_n denotes the number of mines supplied by cluster C on day n.

The random variables X_n are assumed to follow a Poisson distribution with a rate parameter μ_n, which is defined as:

$$\mu_n = \mu\omega_j\phi_{\rho_j}(\omega_j n). \tag{6.1}$$

Here, μ is the average rate parameter, ω_j and ρ_j are the circular frequency and modulation factor, respectively, and $\phi(\theta, \rho)$ is the circular probability density function (pdf) given by:

$$\phi(\theta, \rho) = \frac{1 + 2\rho\cos(\theta)}{2\pi}. \tag{6.2}$$

6.1 Prior Distributions

The Bayesian approach requires specifying prior distributions for the model parameters. On Day 1, the prior distributions are defined as:

$$\pi^{(1)}(\mu, \omega, \rho) \propto \pi^{(1)}(\mu) \times \pi^{(1)}(\omega, \rho) \tag{6.3}$$

We assume the prior for μ follows a Gamma distribution.

$$\pi^{(1)}(\mu|\omega, \rho) \sim \text{Gamma}(\nu_0, \mu_0) = \frac{\mu_0^{\nu_0} e^{-\mu_0 \mu} \mu^{\nu_0 - 1}}{\Gamma(\nu_0)}. \tag{6.4}$$

where ν_0 and μ_0 are the shape and rate parameters, respectively.

As mentioned in section 3.5, the gamma distribution is conjugate to the Poisson distribution. Since we have a Poisson likelihood as the number of supplies is Poisson Random Variables, a gamma prior for μ will ensure the posterior distribution for μ will also be gamma. This conjugacy simplifies the computational process and makes Bayesian updates straightforward.The gamma distribution also ensures that estimates for μ remain positive. Given that μ represents a rate; negative values wouldn't make sense, so this property is desirable.

The priors for ω and ρ are discretized over a set of predefined pairs (ω_j, ρ_j), with equal probabilities $\frac{1}{k}$:

$$\pi^{(1)}(\omega, \rho) \sim \left\{ \begin{array}{cccc} (\omega_1, \rho_1), & (\omega_2, \rho_2), & \ldots, & (\omega_k, \rho_k) \\ \frac{1}{k} & \frac{1}{k} & \cdots & \frac{1}{k} \end{array} \right\}. \tag{6.5}$$

In this discretization, a grid of 100 $\times$ 100 values is considered, with ρ varying from -1 to 1 and ω varying from 1 to 100.

Likelihood Function

The likelihood function for Day n is derived from the observed data X_n:

$$L(X_n \mid \mu, \omega, \rho) = \frac{\mu_n^{X_n} e^{-\mu_n}}{X_n!}. \tag{6.6}$$

6.2 Posterior Distributions

Using the prior distributions and the likelihood function, the posterior distributions for Day n are obtained as:

$$\pi(\mu, \omega, \rho \mid X_n) \propto L(X_n \mid \mu, \omega, \rho) \times \pi(\mu \mid \omega, \rho) \times \pi(\omega, \rho) \tag{6.7}$$

Simplifying the posterior distributions separately for μ, ω, and ρ, we get:

$$\pi(\mu \mid X_n, \omega_j, \rho_j) \propto \mu^{X_n} e^{-\mu\omega_j\phi_{\rho_j}(\omega_j n)} e^{-\lambda_{n-1}(\alpha_j)\mu} \mu^{\nu_{n-1}(\alpha_j)-1} \tag{6.8}$$

$$\pi(\mu \mid X_n, \omega_j, \rho_j) \to \text{Gamma}(\nu_n, \lambda_n) \tag{6.9}$$

where $\nu_n(\alpha_j) = \nu_{n-1}(\alpha_j) + X_n$ and $\lambda_n(\alpha_j) = \lambda_{n-1}(\alpha_j) + \omega_j\phi_{\rho_j}(\omega_j n)$.

For ω_j and ρ_j, the posterior distribution is given by:

$$\pi(\omega_j, \rho_j \mid X_n) \propto \left[\frac{\Gamma(X_n + \nu_{n-1}(\alpha_j))}{(\omega_j\phi_{\rho_j}(\omega_j n) + \lambda_{n-1}(\alpha_j))^{X_n + \nu_{n-1}(\alpha_j)}}\right] \frac{(\omega_j\phi_{\rho_j}(\omega_j n))^{X_n}}{X_n!} \frac{\lambda_{n-1}(\alpha_j)^{\nu_{n-1}(\alpha_j)}}{\Gamma(\nu_{n-1}(\alpha_j))} p_j^{(n-1)}. \tag{6.10}$$

Here, $\alpha_j = (\omega_j, \rho_j)$, and $p_j^{(n-1)}$ is the prior probability of the pair (ω_j, ρ_j) on Day $(n-1)$.

6.3 Updating Posterior Distributions

The posterior distributions for Day n become the prior distributions for Day $(n+1)$. The updating process proceeds as follows:

$$\pi^{(n)}(\mu, \omega, \rho) \propto \pi^{(n)}(\mu|\omega, \rho) \times \pi^{(n)}(\omega, \rho). \tag{6.11}$$

where,

$$\pi^{(n)}(\mu|\omega, \rho) \sim \text{Gamma}(\nu_n, \lambda_n). \tag{6.12}$$

$$\pi^{(n)}(\omega, \rho) \sim \left\{ \begin{array}{cccc} (\omega_1, \rho_1), & (\omega_2, \rho_2), & \dots & (\omega_k, \rho_k) \\ p_1^{(n)} & p_2^{(n)} & \dots & p_k^{(n)} \end{array} \right\}. \tag{6.13}$$

The updated probabilities $p_j^{(n)}$ for the pairs (ω_j, ρ_j) are computed as:

$$p_j^{(n)} = \frac{C_j}{\sum_{i=1}^{k} C_i}. \tag{6.14}$$

where C_j is a quantity proportional to the posterior distribution of (ω_j, ρ_j) on Day n.

$$C_j = \left[\frac{\Gamma(X_n + \nu_{n-1}(\alpha_j))}{(\omega_j \phi_{\rho_j}(\omega_j n) + \lambda_{n-1}(\alpha_j))^{X_n + \nu_{n-1}(\alpha_j)}} \right] \frac{(\omega_j \phi_{\rho_j}(\omega_j n))^{X_n}}{X_n!} \frac{\lambda_{n-1}(\alpha_j)^{\nu_{n-1}(\alpha_j)}}{\Gamma(\nu_{n-1}(\alpha_j))} p_j^{(n-1)}. \tag{6.15}$$

After the update, we obtain a single value of ν_n and $\lambda_n, p^{(n)}$Vectors of length k .

6.4 Prediction

After updating the posterior distributions, predictions for future days can be made. The predictive distribution for Day $(n+1)$ is obtained by integrating over the posterior distributions of μ, ω, and ρ from Day n:

$$P(X_{n+1} = l \mid X_n) = \int_0^\infty \frac{[\mu\omega_j\phi_{\rho_j}(\omega_j(n+1))]^{X_{n+1}} e^{-\mu\omega_j\phi_{\rho_j}(\omega_j(n+1))}}{X_{n+1}!} d\pi(\mu, \omega, \rho \mid X_n). \quad (6.16)$$

For a set of simulated values $(\mu_i, \omega_i, \rho_i, i = 1, 2, \ldots, N_{\text{sim}})$ from the posterior distributions, the predictive probability for $X_{n+1} = l$ can be approximated as:

$$\text{mean}\left(f_\mu\left(X_{n+1} = l, [\mu_i\omega_i\phi_{\rho_i}(\omega_i(n+1)), i = 1, 2, \ldots, N_{\text{sim}}]\right)\right). \quad (6.17)$$

where f_μ denotes the Poisson probability mass function.

This Bayesian approach allows for the sequential updating of parameter distributions as new data becomes available, providing a comprehensive probabilistic framework for modelling and predicting the production of mines within clusters.

We verify the predictions obtained from the Bayesian approach with the actual data, which does not give accurate predictions. We have tried the methods in which we first tried to predict the counts from a particular cluster and then look for the spatial distribution within the cluster. Since this idea of modelling is not proving very useful, let us try to change our approach while building our next statistical model.

Chapter 7

KDE Model

So far, we have continued with the approach of estimating the parameters of the number of supplies each day first and then looking at the spatial distribution of supplies. In this chapter, we will consider an alternative approach using the non-parametric approach of Kernel Density Estimation, where we will structure the model in such a way that it will try to predict on a certain day, which mine within a particular cluster will have the production and then predict the number of supplies from that mine.

7.1 Multivariate Kernel Density

While describing the non-parametric approaches towards density estimation in section 2.4.4 we describe the univariate Kernel Density Estimator. This can be extended to a bivariate setting to consider the longitude and latitude of the mines, as described in this section. Bivariate kde at a point (s_x, s_y) based on sample data points $\{(x_i, y_i), i = 1, 2, \ldots, N\}$ can be written as

$$f(s_x, s_y) = \frac{1}{N} \sum_{i=1}^{N} \kappa_{h_1,h_2}(s_x - x_i, s_y - y_i), \tag{7.1}$$

where $\kappa_{h_1,h_2}(\cdot)$ is a suitably chosen bivariate kernel function Taking $\kappa_{h_1,h_2}(\cdot)$ to be the

standard Gaussian kernel, the spatial kde at any location with (longitude, latitude) coordinates (s_x, s_y), based on N locations $\{(x_i, y_i), i = 1, 2, \ldots, N)\}$, as

$$f(s_x, s_y) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{1}{h_1} \kappa \left(\frac{s_x - x_i}{h_1} \right) \frac{1}{h_2} \kappa \left(\frac{s_y - y_i}{h_2} \right) \right\}, \tag{7.2}$$

where

$$\kappa(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \quad . \tag{7.3}$$

We will use a "rolling Kernel Density approach", which considers previous production history to predict the next day's production. The KDE model aims to estimate the production probability for each mine on a given day based on its location (latitude and longitude) and past production history (represented by the assigned weights and bandwidths). The KDE approach involves constructing a smooth, continuous probability distribution function (PDF) from the observed data points (mine locations) by summing up kernel functions centred at each mine. Since we are interested in estimating the probability of production for each mine, rather than the continuous *pdf*, we evaluate the KDE estimator at the location of each mine (s_x, s_y) The kernel density estimator is defined as location with (longitude, latitude) coordinates (s_x, s_y), on M different mines $\{(x_i, y_i), i = 1, 2, \ldots, M)\}$

$$\hat{f}(s_x, s_y) = \sum_{i=1}^{M} \left\{ \kappa \left(\frac{s_x - x_i}{h_x} \right) \kappa \left(\frac{s_y - y_i}{h_y} \right) \right\}. \tag{7.4}$$

Where: $\hat{f}(x)$ is the estimated PDF at location x, $h = (h_{\text{lat}}, h_{\text{lon}})$ is the vector of bandwidths for latitude and longitude, $K(\cdot)$ is the kernel function,a Gaussian kernel

7.1.1 Adaptation for Latitude and Longitude

Given the geographical nature of our data (latitude and longitude), we extend the bandwidth calculation to a bivariate setting, calculating separate bandwidths for each dimension. For a bivariate KDE with variables x (latitude) and y (longitude), the bandwidth matrix H is

defined as:

$$H = \begin{bmatrix} h_x^2 & 0 \\ 0 & h_y^2 \end{bmatrix} \tag{7.5}$$

where h_x and h_y are the bandwidths for latitude and longitude, respectively, calculated using Silverman's rule adapted for each dimension.

To account for the day-to-day variability in production patterns, these bandwidths are recalculated daily using only the data available to that day separately for each cluster., ensuring that the model's sensitivity adapts to the most recent trends.

7.1.2 Numerical Stability and Thresholding

In practice, we impose a lower limit on the bandwidth to avoid numerical issues that may arise from extremely small bandwidth values, which could, in turn, lead to overfitting. If the calculated bandwidth falls below a threshold $h_{\min}$, we set the bandwidth to this threshold:

$$h_x = \max(h_x, h_{\min}), \quad h_y = \max(h_y, h_{\min}). \tag{7.6}$$

where $h_{\min}$ is typically chosen based on empirical observations of the dataset's scale and variance (e.g., $h_{\min} = 0.002$ as mentioned).

7.2 Assigning Weights: First Approach

To incorporate the assigned weights w_i, which represent the production history of each mine, we modify the kernel density estimator by multiplying each kernel function with its corresponding weight. The weights w_i are derived from the normalized production counts of each mine up to the current day. Let n_i be the production count for mine i, which represents the number of days the mine has had at least one production up to the current day. The weight w_i is calculated as

$$w_i = \frac{n_i}{\sum_{k=1}^{M} n_k} \quad . \tag{7.7}$$

We can think of this method of assigning weights to each mine proportioning to the total production. As evident here, the sum of all the weights for each day will add up to 1, i.e. $\sum_{i=1}^{M} w_i = 1$. Therefore, the weighted kernel density estimator will be :

$$\hat{f}(s_x, s_y) = \sum_{i=1}^{M} w_i \left\{ \kappa\left(\frac{s_x - x_i}{h_x}\right) \kappa\left(\frac{s_y - y_i}{h_y}\right) \right\}. \tag{7.8}$$

The values obtained from this estimator are then normalized to ensure that they sum up to 1, representing a valid PMF:

$$f(s_x, s_y) = \frac{\hat{f}(s_x, s_y)}{\sum_{i=1}^{M} \hat{f}(s_x, s_y)} \quad . \tag{7.9}$$

7.3 Assigning Weights: Second Approach

Instead of using the normalized production counts to assign weights, this approach considers the periodic behaviours of production from each cluster. It incorporates the recency and periodicity with which each mine has produced. Let N denote the total observed production duration (in days) and let Δt represent the inter-production intervals for a mine, i.e., intervals between the days on which it has production. So Δt can vary from 1 to $N - 1$ days.These inter-production times are assumed to be independent and identically distributed (i.i.d.).

For example, if we consider 200 days as the total duration of production, then Δt can vary from 1 to 199. Say a mine has production on days 1, 25, 50, 100, and 200. Then Δt for this mine will be $\Delta t_1 = 1$, $\Delta t_2 = 24$, $\Delta t_3 = 25$, $\Delta t_4 = 50$, $\Delta t_5 = 100$. For simplicity, we assume that all the mines have production on day 0, so $\Delta t_1 = 1$.

Now, for a given day n (which is less than N) and a cluster C, we obtain all such Δt for all the mines that belong to that cluster C. We denote this by X and we assign weights to each Δt. Even though the maximum value of Δt in this case will be n, we will assign weights to each Δt for the complete duration, i.e., from 1 to $N - 1$, using this cluster-wise data up to day n by kernel density estimation on this data. The bandwidth is calculated using Silverman's rule of thumb for this data and the expression is as follows, utilizing the dataset X:

$$\omega(\Delta t) = \frac{1}{|X|} \sum_{\Delta t \in X} K\left(\frac{\Delta t - x}{h}\right) \quad . \tag{7.10}$$

where K represents the Gaussian kernel function.

The normalisation ensures that all the weights sum up to 1 .We extend this to all the remaining clusters and perform this analysis each day up to N. Since there is new incoming data each day, the bandwidth h and the weights for Δt will be updated cluster-wise for each day. Each day we are essentially populating a matrix of rows $N-1$ and columns equal to the number of clusters.

Now, we are going to assign weights w_i' to the mine. We first check which cluster the mine belongs to, and then from the Δt weight matrix, we give the weights corresponding to the duration from the last production day to the next day. Let's consider the same example again where the mine has production on days 1, 25, 50, and 100, and on days 101 and 151, we assign weights to the mine. First, we check which cluster this mine belongs to. We put the weight as the value corresponding to $\Delta t = 1$ for day 101 and $\Delta t = 51$ for day 151 for that mine. Therefore, the weight assigned to a mine is the weight $\omega(\Delta t)$ corresponding to its inter-production time Δt.

The kernel density estimator modified according to the new weights will be :

$$\hat{f}(s_x, s_y) = \sum_{i=1}^{M} w_i' \left\{ \kappa\left(\frac{s_x - x_i}{h_x}\right) \kappa\left(\frac{s_y - y_i}{h_y}\right) \right\} \quad . \tag{7.11}$$

The values obtained from this estimator are then normalized to ensure that they sum up to 1, representing a valid PMF:

$$f(s_x, s_y) = \frac{\hat{f}(s_x, s_y)}{\sum_{i=1}^{M} \hat{f}(s_x, s_y)} \quad . \tag{7.12}$$

Integrating $\omega(\Delta t)$ into the predictive model enhances its forecasting accuracy by aligning more closely with the temporal production patterns capturing the periodic behaviours and recency of production of mines.This methodology aligns with the inherent production patterns observed within clusters and provides a robust framework for enhancing the predictive

accuracy of production forecasts.

7.4 Adaptive Bandwidth: Third Approach

The adaptive bandwidth kernel density estimation (KDE) model adjusts the bandwidth for each data point based on its local density value estimated by a preliminary fixed bandwidth KDE. The adaptive process is guided by the principle that areas of higher density require a smaller bandwidth, while sparser areas necessitate a larger bandwidth.

We use the second approach to obtain our preliminary KDE estimate.Given a preliminary fixed bandwidth KDE estimate $\hat{f}_p(x_i, y_i)$ using bandwidths h_x and h_y, the adaptive bandwidth at each point (x_i, y_i) is proportional to:

$$\hat{h}_{x,y}(x_i, y_i) \propto \hat{f}_p(x_i, y_i)^{-1/2} \tag{7.13}$$

This adaptive bandwidth allows for local variation in density, providing an improved representation of the underlying distribution of the data, which is particularly beneficial when dealing with heterogeneous population densities. In practice, the refined bandwidths $h_x(x, y)$ and $h_y(x, y)$ are computed as:

$$h_{x,adaptive}(x, y) = \alpha_x \hat{f}_p(x, y)^{-1/2}, \quad h_{y,adaptive}(x, y) = \alpha_y \hat{f}_p(x, y)^{-1/2} \quad . \tag{7.14}$$

where α_x and α_y are constants, typically chosen as the median of the preliminary bandwidths divided by the square root of the corresponding preliminary kernel density estimates.

Utilizing these newly adjusted bandwidths, the final KDE value $\hat{f}_{\text{Prop}}$ is recalculated to provide an estimator as :

$$\hat{f}(s_x, s_y) = \sum_{i=1}^{M} w_i' \left\{ \kappa \left(\frac{s_x - x_i}{h_{x,adaptive}} \right) \kappa \left(\frac{s_y - y_i}{h_{y,adaptive}} \right) \right\} \quad . \tag{7.15}$$

The values obtained from this estimator are then normalized to ensure that they sum up to

1, representing a valid PMF:

$$f(s_x, s_y) = \frac{\hat{f}(s_x, s_y)}{\sum_{i=1}^{M} \hat{f}(s_x, s_y)} \quad . \tag{7.16}$$

The normalized KDE values obtained represent the estimated pmf for the production of each mine on the given day.

We employ logistic regression to model the probability of a mine having production on the next day. This is treated as a binary classification problem, where the outcome y_i for mine i is 1 if there is production and 0 if there is no production. The logistic regression model relates the log odds of the binary outcome to the predictor variable, which in this case is the kernel density estimated value for each mine. Let p_i denote the KDE value associated with mine i.

$$\log\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = \beta_0 + \beta_1 \cdot p_i \quad . \tag{7.17}$$

Here, $P(y_i = 1)$ represents the probability that mine i will have production on the next day, β_0 and β_1 are the model coefficients to be estimated,

The probability of production can be derived from the log odds as:

$$P(y_i = 1) = \frac{\exp(\beta_0 + \beta_1 \cdot p_i)}{1 + \exp(\beta_0 + \beta_1 \cdot p_i)} \quad . \tag{7.18}$$

To obtain a binary prediction from the model's probability output, we introduce a cutoff value c:

$$y_{\text{pred}}(i) = \begin{cases} 1 & \text{if } P(y_i = 1) \geq c \\ 0 & \text{if } P(y_i = 1) < c \quad . \end{cases} \tag{7.19}$$

A common choice for the cutoff is 0.5, known as the logistic cutoff. At this cutoff value, the model classifies a mine as having production if the predicted probability exceeds 0.5, and as having no production otherwise.

The logistic cutoff can be derived by solving the logistic regression equation for the KDE value at which $P(y_i = 1) = 0.5$:

$$\log\left(\frac{0.5}{1-0.5}\right) = \beta_0 + \beta_1 \cdot \text{logistic cutoff} \tag{7.20}$$

$$0 = \beta_0 + \beta_1 \cdot \text{logistic cutoff} \tag{7.21}$$

$$\text{logistic cutoff} = -\frac{\beta_0}{\beta_1}. \tag{7.22}$$

This logistic cutoff value represents the KDE threshold at which the model predicts a 50% probability of having production on the next day for a given mine.

Multiple cutoff values (e.g., $\frac{1}{No.Of Mines}$, $\frac{2}{No.Of Mines}$, $\frac{3}{No.Of Mines}$, mean(KDE_{norm}), Logistic cutoff) are explored to find the optimal threshold for classification. The model's performance is evaluated using accuracy, sensitivity (actual positive rate), and specificity (negative rate) and F score, which are calculated based on the predicted and actual production values for each mine.

Sensitivity (True Positive Rate)

Sensitivity, or the true positive rate, measures the proportion of actual positives (mines with production) that are correctly identified by the model. It is defined as:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad . \tag{7.23}$$

where True Positives are mines correctly predicted to have production, and False Negatives are mines that produced but were not identified by the model.

Specificity (True Negative Rate)

Specificity, or the true negative rate, quantifies the proportion of actual negatives (mines without production) correctly identified as such. It is calculated as:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \quad . \tag{7.24}$$

where True Negatives are mines correctly predicted to have no production, and False Positives are mines predicted to produce but did not.

F-Score

The F-score is a harmonic mean of precision and sensitivity, offering a balance between them. It is particularly useful when the class distribution is imbalanced. The F-score is defined as:

$$F\text{-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \quad . \tag{7.25}$$

Where Precision (Positive Predictive Value) is the proportion of true positive predictions in all positive predictions made by the model:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad . \tag{7.26}$$

Chapter 8

Optimisation

We have discussed three statistical models to sample the probabilities of production from each mine. Once we obtain the probabilities from the models, the next step involves optimizing the allocation of resources. In our problem, this means allocating the amount of coal produced from each mine to an electricity-generating power plant, considering the presence of railway routes, the cost of transporting the coal between these two locations and other constraints. This chapter describes the process of optimally allocating resources and our proposed model for optimization.

Optimization is focused on finding the most favourable solution from a set of possible alternatives. Specifically, a mathematical optimization challenge involves determining a function's maximum or minimum value, known as the *objective function*, within the confines of potential solutions, termed the *feasible region*. The feasible region is typically a subset of $\mathbb{R}^n$, the n-dimensional space of real numbers, and the objective function is a mapping from $\mathbb{R}^n$ to $\mathbb{R}$.

In this chapter, we concentrate on a subclass of optimization problems known as linear programming problems (LPs). An LP is an optimization problem in $\mathbb{R}^n$ where the objective function and constraints are linear. The objective function is expressed as:

$$c_1x_1 + c_2x_2 + \cdots + c_nx_n. \tag{8.1}$$

Here, c_i represents a real constant for dimension i (where i ranges from 1 to n).The objective is to find this function's maximum (in case of maximization problems) or minimum (for minimization problems) value over the feasible region. The objective is to find the vector $\mathbf{x}$ (composed of decision variables $x_1, x_2, \ldots, x_n$) within the feasible region that either minimizes or maximizes the value of the objective function. The set of solutions defines the feasible region for an LP to a finite number of linear constraints, which take the form of inequalities and equalities:

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{inx_n} \leq b_i \quad \text{for } i = 1, \ldots, s. \tag{8.2}$$

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{inx_n} = b_i \quad \text{for } i = s+1, \ldots, m. \tag{8.3}$$

In these expressions, each a_{ij} is a known coefficient, and each b_i is a constant that defines the bounds of the feasible region for the respective constraint. The feasible set of the LP is the vector $x = (x_1, x_2, \ldots, x_n)$ that satisfies all the constraints.

Decision variables are the variables that will determine the solution to the linear programming problem. They represent the unknowns the LP will solve for and are usually denoted as $x_1, x_2, \ldots, x_n$. These variables must satisfy the constraints and are used to evaluate the objective function. In the context of linear programs, decision variables are typically required to be non-negative, represented as:

$$x_j \geq 0 \text{ for } j = 1, \ldots, n. \tag{8.4}$$

The optimal solution of an LP is the set of values for the decision variables that yield the best (maximum or minimum) objective function value while simultaneously satisfying all of the constraints.

8.1 Multiobjective Optimization

In many real-world optimization problems, multiple, often conflicting, objective functions need to be optimized simultaneously. This scenario is known as multiobjective optimization.

Instead of a single objective function, we have a vector-valued objective function $\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_k(\mathbf{x}))$, where $\mathbf{x} \in \mathbb{R}^n$ is the vector of decision variables, and each $f_i(\mathbf{x})$ is an individual objective function to be minimized or maximized.

Unlike in single-objective optimization, where the solution is clearly defined by the optimal value of the objective function, multiobjective optimization solutions are typically not completely ordered. This leads to the concept of Pareto optimality, which is central to multiobjective problems. The goal in multiobjective optimization is to find the set of Pareto optimal solutions, which are not dominated by any other feasible solution. A solution $\mathbf{x}^*$ is said to be Pareto optimal if there does not exist another feasible solution $\mathbf{x}$ such that $f_i(\mathbf{x}) \leq f_i(\mathbf{x}^*)$ for all $i = 1, 2, \ldots, k$, and $f_j(\mathbf{x}) < f_j(\mathbf{x}^*)$ for at least one j.

There are various methods to solve multiobjective optimization problems, including the weighted approach and the epsilon constraint method. (Deb. K, 2014) (3)

8.1.1 Weighted Approach

In the weighted approach, the multiple objective functions are combined into a single objective function by assigning weights to each objective function and summing them up. The resulting single-objective optimization problem is:

$$\begin{aligned} \text{Minimize (or Maximize)} \quad & \sum_{i=1}^{k} w_i f_i(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{S}. \end{aligned}$$

where $w_i \geq 0$ are the weights for each objective function $f_i(\mathbf{x})$, satisfying $\sum_{i=1}^{k} w_i = 1$ and $\mathcal{S}$ is the feasible region defined by the constraints. . This approach assumes the objectives are commensurate and can be meaningfully combined.

The weights w_i represent the relative importance of each objective function. By varying the weights, different Pareto optimal solutions can be obtained. However, this approach has limitations, as it may not capture all Pareto optimal solutions, especially if the Pareto frontier is non-convex, and the choice of weights can be subjective.

8.1.2 Epsilon Constraint Method

The epsilon constraint method is another approach to solving multiobjective optimization problems. In this method, one objective function is chosen as the primary objective, and the other objective functions are converted into constraints by setting an upper bound (epsilon level) for each function.

The problem formulation for the epsilon constraint method is:

$$\begin{aligned} \text{Minimize (or Maximize)} \quad & f_p(\mathbf{x}) \\ \text{subject to} \quad & f_i(\mathbf{x}) \leq \epsilon_i, \quad i = 1, 2, \ldots, k, \quad i \neq p \\ & \mathbf{x} \in \mathcal{S}. \end{aligned}$$

where $f_p(\mathbf{x})$ is the primary objective function, ϵ_i are the upper bounds (epsilon levels) for the other objective functions $f_i(\mathbf{x})$, and $\mathcal{S}$ is the feasible region defined by the constraints.

This method involves solving multiple optimization problems by varying the bounds, ϵ_i, thus tracing out the trade-off surface, known as the Pareto frontier, which represents non-dominated solutions. It is especially effective when the objectives are incommensurate, or the Pareto frontier is non-convex. The epsilon constraint method has the advantage of generating Pareto optimal solutions in a systematic way, but it may require solving multiple single-objective optimization problems to explore the Pareto frontier.

Each iteration yields a Pareto optimal solution, with different ϵ_i values producing different points on the Pareto frontier. We should ensure that the values of ϵ_i are set such that the feasible region $\mathcal{S}$ remains non-empty.

This method is beneficial when the objectives are incommensurable or when finding solutions across the entire range of trade-offs among objectives is important. It is a versatile approach that can reveal the nature of the trade-offs between objectives and is effective even when the Pareto frontier is non-convex.The choice of method from either these two depends on the specific situation.

8.2 Proposed Model

After the initial sampling, we obtained the probability of coal production at each mine by using the proposed KDE model with the modified weights and adaptive bandwidth approach. We define "Supply Capacity" for each mine as the number of power-generating plants each mine supplies on the day the mine is in production. We assume that if a mine has production on a particular day, the supply capacity will equal the average of the last seven production days of that mine. These supply capacities are actual true recorded values that our model did not predict. The *kde* values multiplied by the average of the previous seven supply capacities give the total amount of coal produced by the mine on production day.

Following the quantification of production, the subsequent step involves the distribution of coal to power plants. For this model, it is assumed that each power plant possesses an unlimited inventory capacity; essentially, they can receive an infinite supply of coal. The infrastructure connecting the mines to the plants comprises predefined railway routes, with each mine having established connections to a set number of power plants. Furthermore, the transportation cost of shipping coal from a particular mine to a designated power plant is predetermined and known.

Our objective is to facilitate equitable distribution of coal, ensuring that each power plant maintains a minimum inventory equivalent to a supply of seven days(for instance). However, this duration is adjustable based on specific requirements. The allocation of coal from each mine to the power plants is thus optimized to minimize transportation costs while adhering to the constraint of maintaining the requisite inventory levels.

The mathematical formulation of our model is articulated below, encompassing supply and demand dynamics, transportation logistics, and inventory management principles.

Supplies	Mines	Probabilities		Plants	Existing Inventory
S_1	M_1	p_1	–	N_1	a_1
S_2	M_2	p_2	(c_{ij})	N_2	a_2
S_3	M_3	p_3	–	N_3	a_3
				N_4	a_4

The model utilizes the following notations for its formulation:

- M_i: Mine i , N_j: Power plant j.
- S_i: Supply Capacities from mine i.This discrete quantity specifies the average number of plants each mine has supplied to in the last seven production days.
- p_i: Probability of coal production at mine i affecting the route to power plant j.
- a_j: Existing inventory at power plant j.This is also a discrete quantity defined as the number of days the plant has enough coal for electricity generation.
- c_{ij}: Cost of transporting coal from mine i to power plant j.
- v_{ij}: Binary indicator of the existence of a railway route from mine i to power plant j (1 if a route exists, 0 otherwise).
- x_{ij}: Decision variable representing the quantity of coal transported from mine i to power plant j.
- M: A sufficiently large number representing an upper bound in constraints where needed.
- l and k: Variables representing the upper and lower bounds on the inventory levels across all power plants.
- w_1 and w_2: Weights indicating the relative importance of equitable distribution and transportation cost minimization in the objective function.

The formulation of our optimization model is as follows, with the dual objectives of minimizing expected transportation costs and ensuring equitable coal distribution:

$$\text{Minimize: } w_1(l-k) + w_2 \left(\sum_i \sum_j c_{ij} x_{ij} \right). \tag{8.5}$$

Subject to:

$$a_j + \sum_i x_{ij} \geq k \quad \forall j, \tag{8.6}$$

$$a_j + \sum_i x_{ij} \leq l \quad \forall j, \tag{8.7}$$

$$\sum_j x_{ij} \leq S_i \cdot p_i \quad \forall i, \tag{8.8}$$

$$x_{ij} \leq M v_{ij} \quad \forall i, j. \tag{8.9}$$

The objective function aims to balance transportation costs with the goal of maintaining equitable inventory levels, represented by minimizing the range $(l - k)$. Constraints (1) and (2) ensure that each power plant's inventory remains within specified bounds, promoting reliability in supply without overstocking. Constraint (3) provides that the total coal supplied by each mine does not exceed its production capacity, adjusted by the probability of production. This ensures a realistic limitation on the amount of coal distributed from each mine. Constraint (4) guarantees that coal shipments only occur along existing railway routes, encapsulating logistical feasibility.

This comprehensive formulation allows for an optimal coal distribution strategy that minimizes costs and supports reliable energy production, considering the realistic production capacities and probabilities of each mine.

One may argue that since for the total amount of coal to be transported for each mine, we are using the probabilities, and it may also happen that even if the probablity is high for a particular mine on a day, production still might not happen. In those cases, the optimization model will work poorly. Our optimization model only prescribes readiness for allocating coal to organizations, but the event may or might not happen. This uncertainty can be reduced by introducing more constraints to the model, such as chance constraints, which will make the model more robust. This uncertainty can also be handled by stochastic optimization, which is not discussed in the thesis.

Chapter 9

Results

In this chapter, we describe the results obtained from applying three statistical approaches we have discussed so far on the dataset.

9.1 Results of MLE

Using our modelling approach, we obtain the following maximum likelihood estimates for the parameters.

cluster	ω	λ	ρ	likelihood
0	0.055116	351.678573	-0.000880	123.823276
1	0.116355	611.589073	0.014100	892.010463
2	3.141593	3.415844	0.001448	282.298101
3	0.047600	384.561020	0.004485	532.378045
4	1.047198	6.374997	-0.028522	16.648253

Table 9.1: Optimization results of the maximum of likelihood and the corresponding values of ω, λ, ρ

The optimizer tries to vary ω, and in the neighbourhood of the λ and ρ, it tries to maximize the log-likelihood. Sometimes minimizing the negative of log-likelihood is impossible

for certain values of ω as the optimizer might encounter issues like getting stuck in local minima, numerical instability, or reaching the maximum number of iterations.

The maximum likelihood estimates of the clusters are nowhere close to the actual supplies from the clusters each day. For each cluster except cluster 4, the number of supplies daily varies from 0-20, whereas for cluster 4, it varies from 0 to 1. We know that the maximum likelihood estimate for a Poisson-independent and identically distributed sample converges to the sample mean. So even if we consider the maximum sample means for each cluster it is nowhere close to the MLE estimates obtained for clusters 0,1,3,4. It is very unlikely that this model captures the underlying probability distribution of the dataset.

9.2 Bayesian Model Results

This snapshot of the output from the R code presents a comprehensive simulation of predicted daily supplies to different clusters over an extended period, reflecting changes in supply dynamics. This detailed distribution shows the predicted number of supplies for each day simulated, spanning across 1000000 iterations.

We notice a decreasing trend of zero supplies. Initially, the number of days with zero supplies was high, but there was a noticeable decrease over time however, the no of supplies being 0 is still high. Considering the simulations of all the 320 days(not displayed here), we conclude that the distribution of supplies has become more widespread over time. Initially, the majority of predictions are concentrated around 0 to 3 supplies per day. As the model progresses, we observe more days with higher supplies (4 supplies and above), indicating an increase in the variability of supply distribution. The appearance of days with unusually high supplies (e.g., 10 or more supplies) towards the later iterations might indicate special events, anomalies, or perhaps errors in the prediction model. Their rare occurrence suggests they are not the norm but rather exceptions. Comparing this with our true data for the number of supplies, we see that the model doesn't perform very well. Although it performs better than MLE, it can still not accurately predict the number of supplies. Simulations suggest that on most days the number of supplies will be zero, which is not the case. Cluster 0 has only three days in a span of 320 days, and no mine from it has production. The results are similar for other clusters as well.

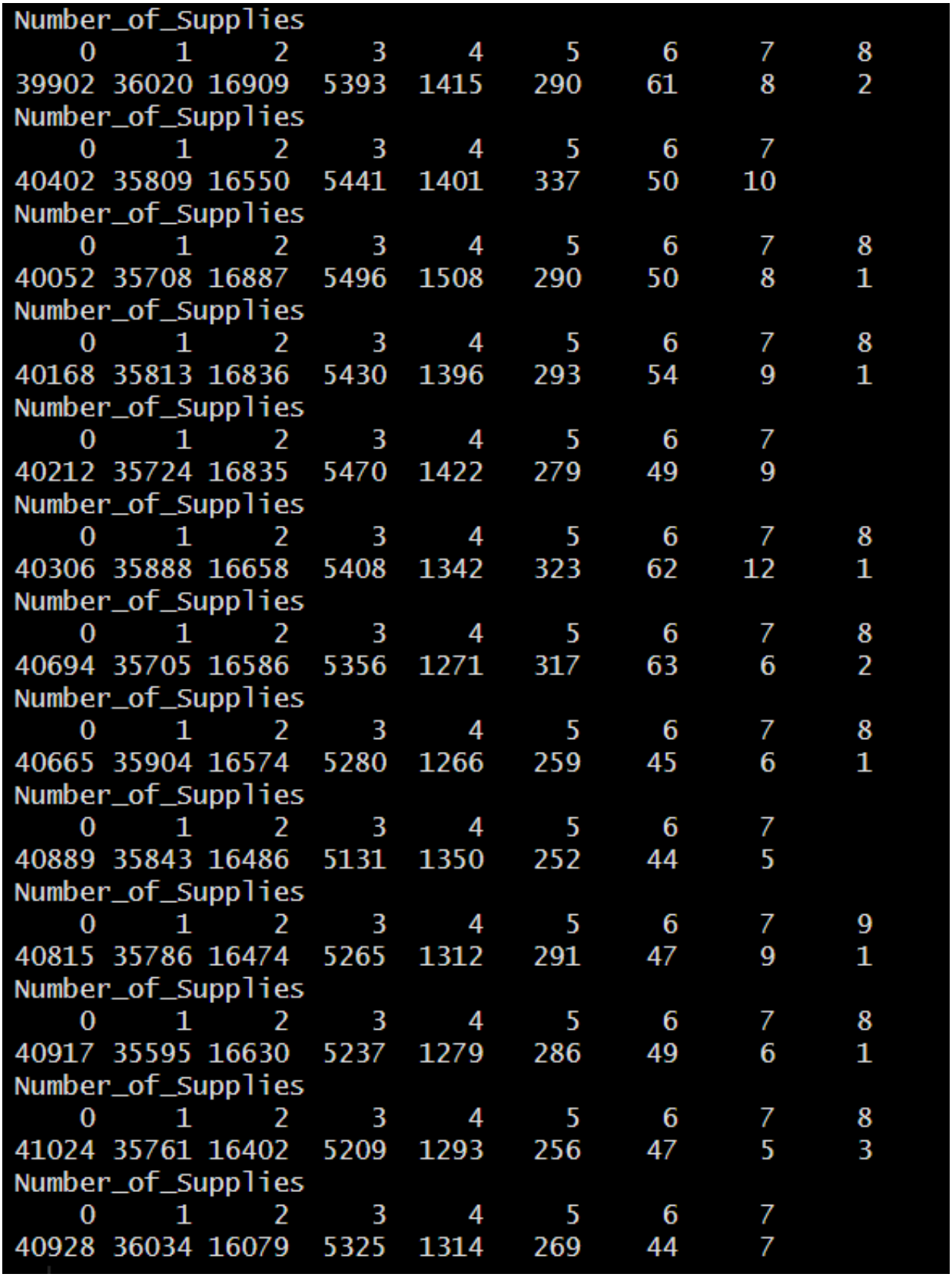

Figure 9.1: predicted supplies for Cluster 0 .

9.3 KDE results

After constructing the model based on the three approaches in the KDE model, We verify the result obtained and check for accuracy, sensitivity, and specificity metrics. We also compare them with the existing Machine learning classification algorithms for mine production. For the Machine learning classifier, we treat this problem of production as a classification problem, with 0 indicating no production and 1 indicating production. The algorithm works by predicting the supplies of each mine for next day considering previous 10-day supplies. We calculate the accuracy, sensitivity, and specificity for all the models on each day, which are displayed in the table.

Mean Metrics:	**Accuracy**	**Sensitivity**	**Specificity**	**Accuracy (Last 120 days)**	**Sensitivity (Last 120 days)**	**Specificity (Last 120 days)**
KDE Model 1 (Mean)	0.83	0.71	0.85	0.82	0.81	0.82
KDE new weight(Mean)	0.88	0.67	0.91	0.89	0.78	0.90
KDE new weight adaptive mean	0.87	0.76	0.88	0.88	0.82	0.89

Table 9.2: Mean of performance metrics over the whole duration and last 120 days.

We also compute the F-score of the three modelling approaches from these, which are 0.71,0.71,0.75 for KDE model 1,2,3, respectively, when they are calculated using the prediction over the whole duration. When we consider the most recent 120 days, the F-scores are 0.81, 0.80, and 0.85, respectively. This shows that the KDE models are adapting to new incoming data.

Results suggest that this model is significantly better than the MLE and Bayesian Approaches. While comparing with the Machine Learning Models, we notice that although this model cannot beat the accuracy of those models, it can outperform them when considering the sensitivity and specificity metrics.

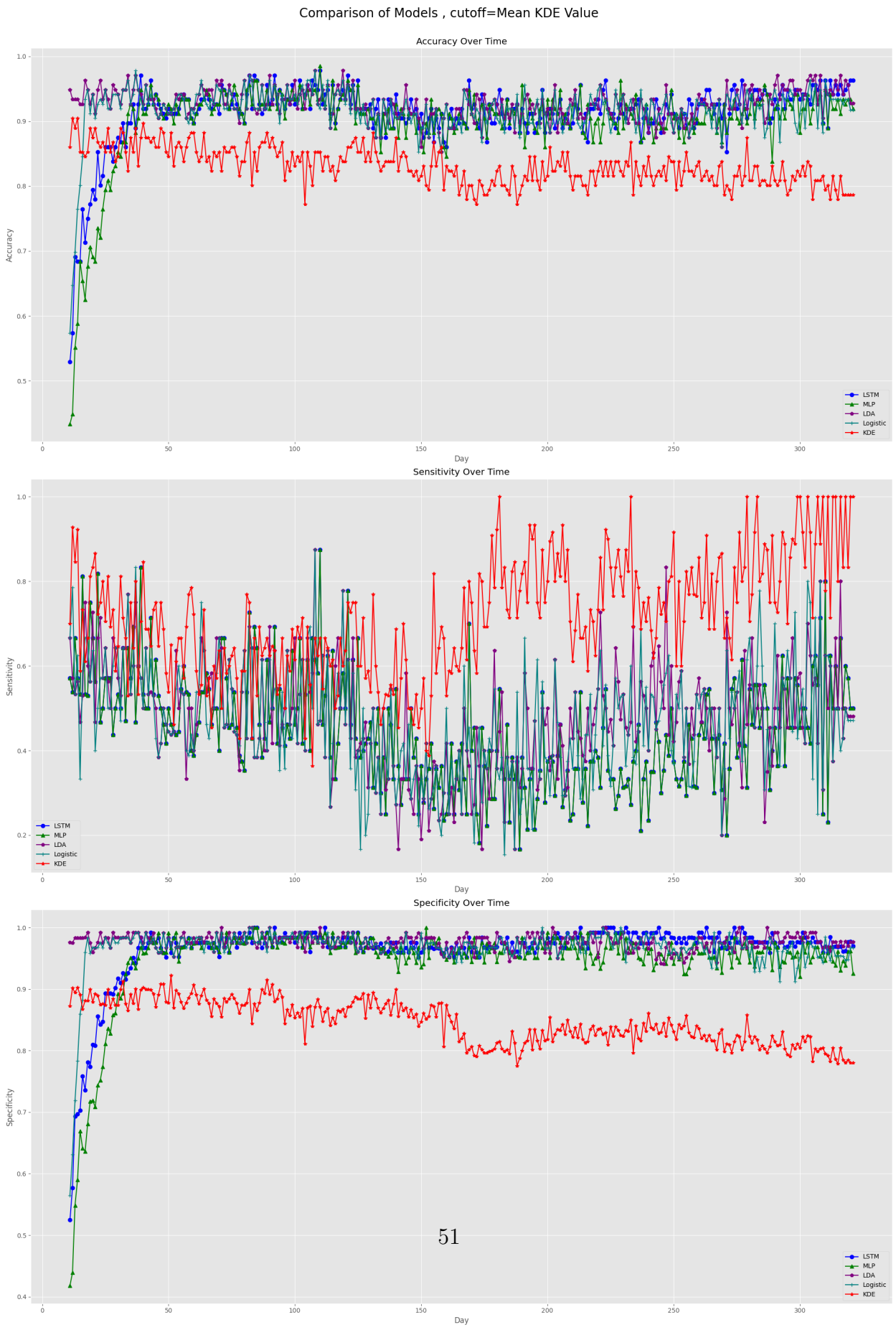

Figure 9.2: Comparison of accuracy, sensitivity, and specificity over time for Model 1

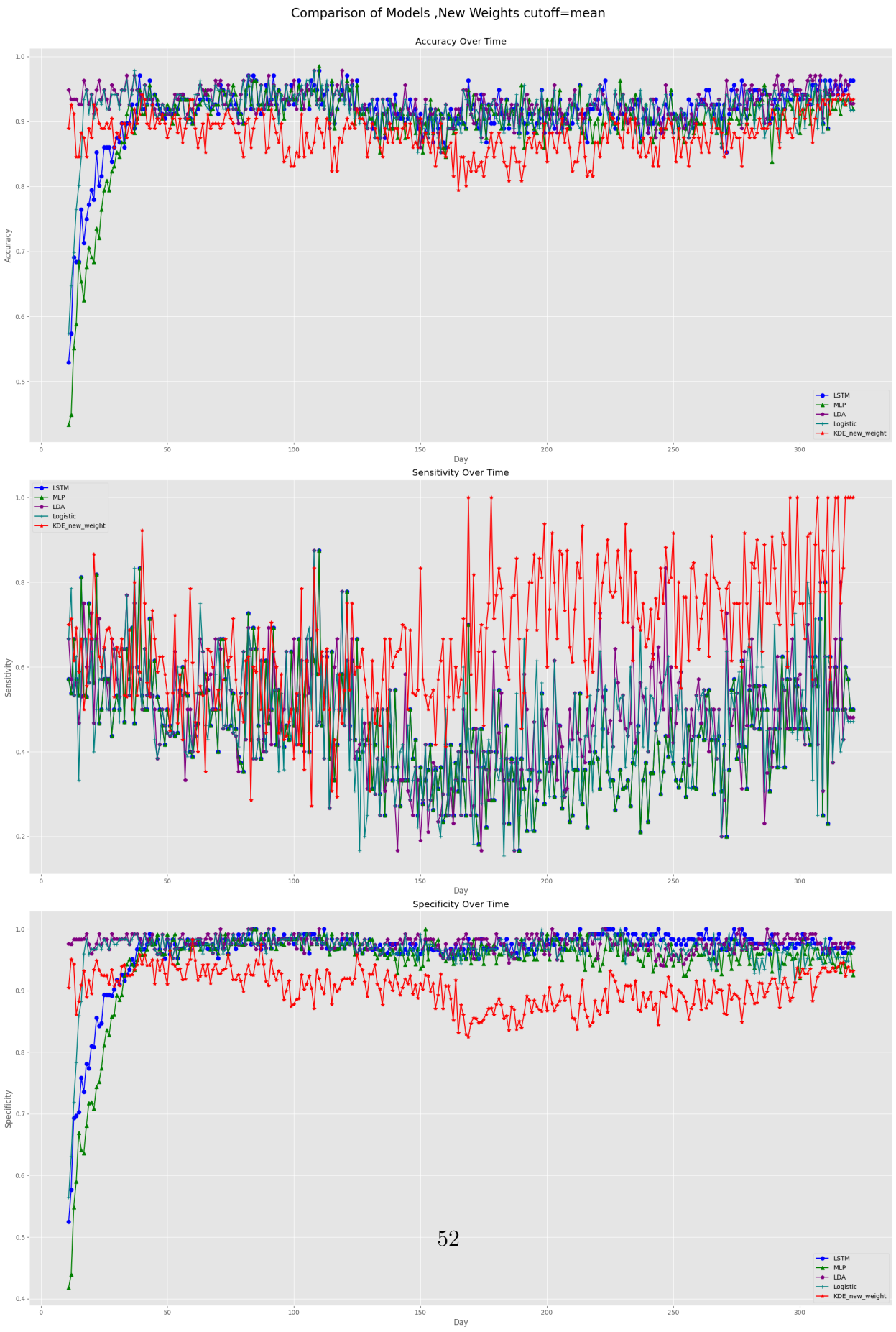

Figure 9.3: Comparison of accuracy, sensitivity, and specificity over time for Model 2

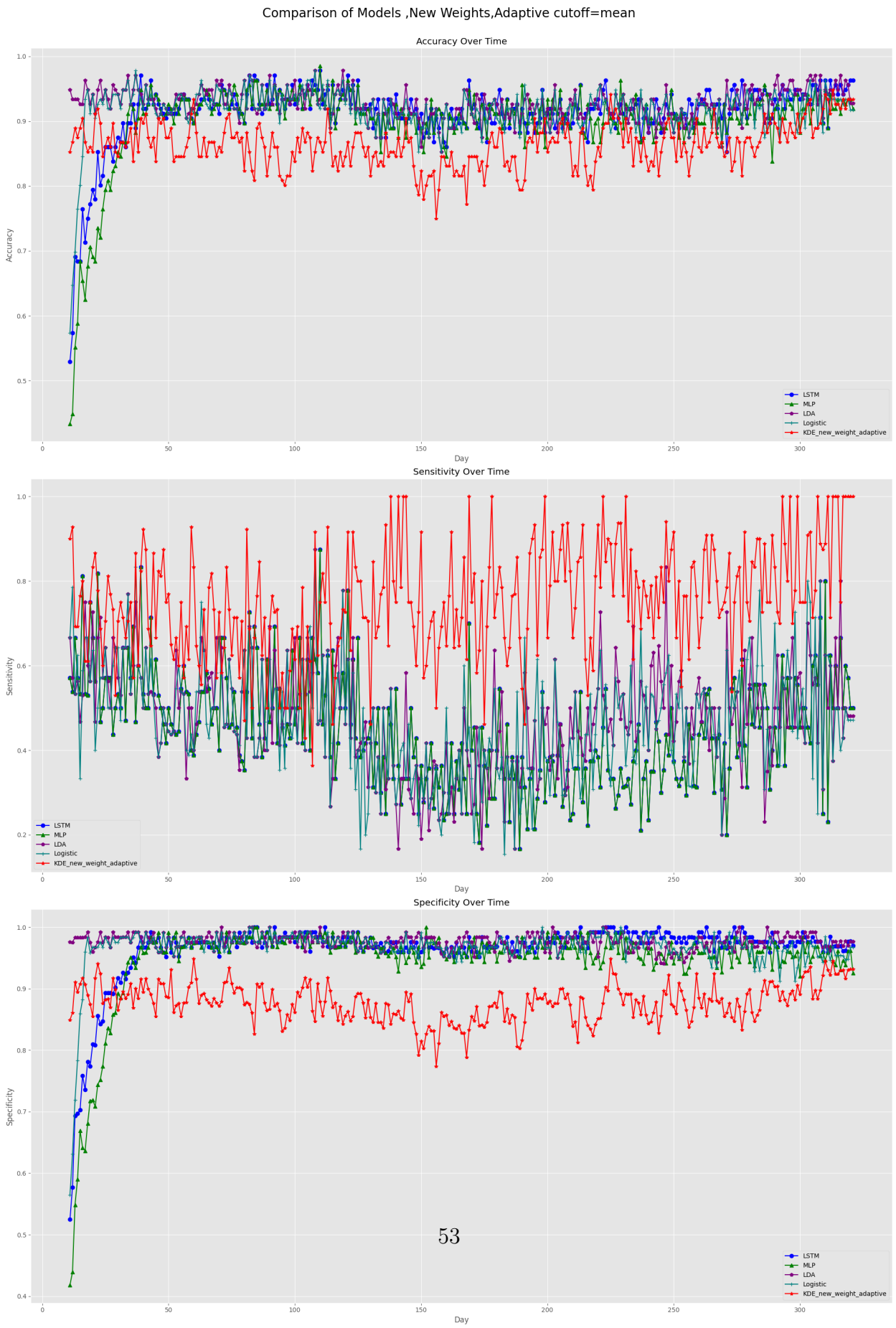

Figure 9.4: Comparison of accuracy,sensitivity and specificity over time for Model 3

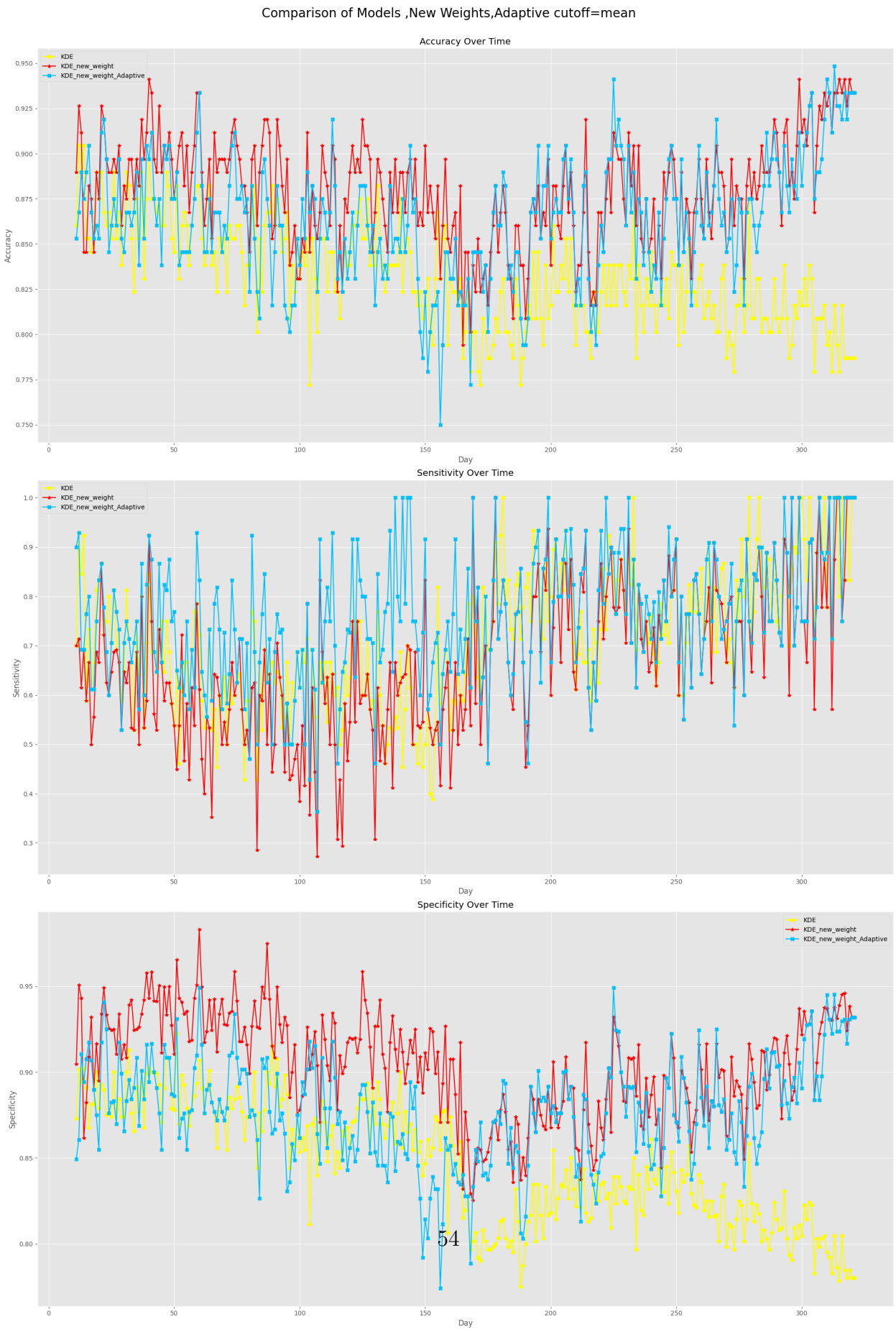

Figure 9.5: Comparison of accuracy, sensitivity, and specificity over time for KDE models

Chapter 10

Conclusion and Future Work

The MLE approach yielded maximum likelihood estimates for cluster parameters that were significantly divergent from actual data, suggesting a poor fit for capturing the underlying probability distribution of the dataset. The expected number of supplies from clusters each day varied widely from the MLE estimates, indicating the model's limitations in accurately predicting the true distribution of supplies.

In Bayesian modelling, despite noticing a decreasing trend in zero supplies over time, the model did not accurately predict the distribution of supplies. The predicted days with zero supplies remained high, and the distribution became more widespread, which, although showing an increase in variability, still did not align well with actual data. The model performed better than MLE but fell short in accurately forecasting supplies.

The KDE approach presented significant improvements over the MLE and Bayesian models regarding accuracy, sensitivity, and specificity metrics. This model was compared against machine learning classification algorithms, showing superior sensitivity and specificity metrics, thus offering a robust alternative. The introduction of an adaptive weighting mechanism in the KDE model, based on the production history and periodic behaviours of mines, represents a methodological innovation.

By adapting the bandwidth based on local data density, the model offers a more nuanced and accurate density estimation, significantly improving predictive accuracy over traditional fixed-bandwidth KDE models. The adaptive nature of the KDE model makes it particularly

valuable in scenarios where data patterns and distributions are subject to change. This adaptability ensures that predictive models remain relevant and accurate over time despite fluctuations in underlying data trends.

The integration of these predictive models with optimization strategies offers a comprehensive framework for managing coal distribution effectively, with potential implications for similar predictive tasks in discrete spatial and temporal settings and can be adapted to various other sectors where forecasting discrete events is crucial, such as healthcare (e.g., predicting patient admissions), agriculture (e.g., crop yield forecasting), and urban planning (e.g., traffic flow prediction). Future efforts could focus on applying the models to a broader range of datasets to validate and potentially improve model robustness.

The current models could be further refined to better handle uncertainties inherent in predictive modelling. Techniques such as Bayesian hierarchical models, stochastic modelling, or ensemble methods could provide more nuanced approaches to uncertainty quantification.

We could focus on developing stochastic optimization models and chance-constrained optimization frameworks tailored to the specifics of discrete event predictive modelling. This includes selecting appropriate probability distributions for uncertain parameters, determining risk tolerance levels for chance constraints, and validating these models against real-world data. By addressing the uncertainties inherent in forecasting and decision-making for discrete events, these approaches offer a path toward more robust, reliable, and effective models for future research and application.

Bibliography

[1] Silverman, Bernard W. Density estimation for statistics and data analysis. Routledge, 2018

[2] Wand, Matt P., and M. Chris Jones. Kernel smoothing. CRC press, 1994

[3] Deb, Kalyanmoy. "Multi-objective Optimization." Search Methodologies (2014): 403.

[4] Schubert, Erich. "Stop using the elbow criterion for k-means and how to choose the number of clusters instead." ACM SIGKDD Explorations Newsletter 25, no. 1 (2023): 36-42.

[5] Wikipedia contributors, "Kernel density estimation," Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=Kerneldensityestimation&oldid=1210998018 [accessed March 27, 2024)].

[6] Lin, Sheldon. "MATH-SHU-236 Lecture 6: k-means." Last modified Spring 2020. Accessed [accessed March 27, 2024]. https://cims.nyu.edu/ sling/MATH-SHU-236-2020-SPRING/MATH-SHU-236-Lecture-6-kmeans.pdf.

[7] "Determining the Optimal Number of Clusters: 3 Must-Know Methods." Datanovia. Accessed [accessed March 27, 2024]. https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/.

[8] Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. "Consumer credit-risk models via machine-learning algorithms." Journal of Banking & Finance 34, no. 11 (2010): 2767-2787.

[9] Ijaz, Muhammad Fazal, Muhammad Attique, and Youngdoo Son. "Data-driven cervical cancer prediction model with outlier detection and over-sampling methods." Sensors 20, no. 10 (2020): 2809.

[10] Satpathi, Anurag, Parul Setiya, Bappa Das, Ajeet Singh Nain, Prakash Kumar Jha, Surendra Singh, and Shikha Singh. "Comparative analysis of statistical and machine learning techniques for rice yield forecasting for Chhattisgarh, India." Sustainability 15, no. 3 (2023): 2786

[11] Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. Bayesian data analysis. Chapman and Hall/CRC, 1995.

[12] Casella, George, and Roger L. Berger. "Statistical lnference." Duxbury press (2002).