SCHOOL OF MATHEMATICS AND PHYSICAL SCIENCES DEPARTMENT
OF MATHEMATICS

EXPLORATORY DATA ANALYSIS AND MATHEMATICAL MODELLING OF AN ACCIDENT AND EMERGENCY DEPARTMENT DATASET

By

Faosiyat Tiamiyu-Tijani

Submitted in partial fulfilment of the requirement for the MSc degree in Data Science at the University of Sussex

August 2024

Supervised by Dr. James Van Yperen

ACKNOWLEDGEMENT

I sincerely appreciate my supervisor, whose unwavering support, invaluable guidance, and remarkable patience have been essential throughout this research. His expertise and insight have profoundly shaped this dissertation.

This dissertation was inspired by my Instructor, with whom I had interacted on my desire to pursue a project with real-world relevance and impact. Who supervised this project, and his mentorship and expertise were instrumental in guiding this research.

I extend my most profound appreciation to my partner, whose unwavering support and faith in my ability strengthened me during this academic endeavour. Your unwavering belief and role have been crucial to this research and career.

ABSTRACT

The National Health Service (NHS) in the United Kingdom (UK) is facing significant challenges in its Accident and Emergency (A&E) department, where patient waiting times are on the rise, and traditional solutions, such as increasing the number of staff, have not solved the issues effectively. Mathematical modelling, like the queueing model, has proven to be effective in operational management for tackling similar problems related to efficiency, resource allocation, and customer satisfaction-while these models provide a framework for understanding queue dynamics, calibrating them with data poses a challenge, and this is why this project aims to use data and mathematical analysis to lay the groundwork for the development of the queueing model.

Previous research by (Armony et al. 2015) [1] has observed by eye three distinct queue regimes by just a graphical representation: increasing queue length during early morning hours, decreasing length in the afternoon, and stable length during intermediate hours. Therefore, this study will build upon this by developing an algorithm to automate a system that detects those regimes' presence in the A&E data and generates interesting statistics and parameters to develop a queueing model tailored for the A&E setting in further research; this tailored queue model, once developed, holds the promise of creating a digital twin of the A&E queueing dynamics current situation, enabling resource managers to make informed decisions and implement effective strategies. This research offers a novel opportunity to address NHS challenges around A&E and lay the groundwork for the future development of a tailored queue model that could potentially revolutionize the healthcare system.

Table of Contents

Chapter 1

1.1 Introduction

In the healthcare sector, particularly in hospitals, efficient resources (doctors, equipment, nurses, etc.) and patient flow (queue) management are paramount. The Accident and Emergency (A&E) department serves as a frontline service, addressing various medical needs ranging from minor injuries to life-threatening emergencies. Due to this, it faces varying levels of demand and complexity, making managing waiting times a multifaceted challenge.

Among these challenges, winter pressures have been reported to contribute significantly to the strain and increase in demand for A&E services [[2]. Winter pressures, for instance, present some of the most challenging times for health and care services. During winter, demand rises sharply due to the increased prevalence of influenza-like illnesses, respiratory diseases associated with colder weather, such as asthma and pneumonia, and infectious winter vomiting bugs like norovirus. Adding to the burden is the rise of dementia. High-income countries typically report dementia diagnosis rates between 20-50%. England, however, stands out with one of the highest dementia diagnosis rates globally. Recent data published in July 2024, confirms this alarming trend, and highlights that 487,432 people in England were diagnosed with dementia in June, marking an increase to a current diagnosis rate of 65%—the highest since the pandemic began [3]. This diagnosis surge may significantly impact A&E services as the healthcare system copes with a growing number of elderly patients requiring emergency care. The UK A&E department has already and is currently experiencing pressure, and this new surge in dementia diagnoses could put even more strain on the system.

In April 2011, a new set of clinical quality indicators was introduced, replacing the previous four-hour waiting time standard, stating that at least 95% of attendees should be admitted, transferred, or discharged within four hours of arrival at any A&E department. 'This report' (NHS England 2011) sets out data coverage, data quality, and performance information for five A&E indicators: (1) Left department before being seen for treatment rate, (2) Re-attendance rate, (3) Time to initial assessment, (4) Time to treatment, and (5) Total time in A&E, became the yardstick for measuring the quality of care in the A&E department [4] .

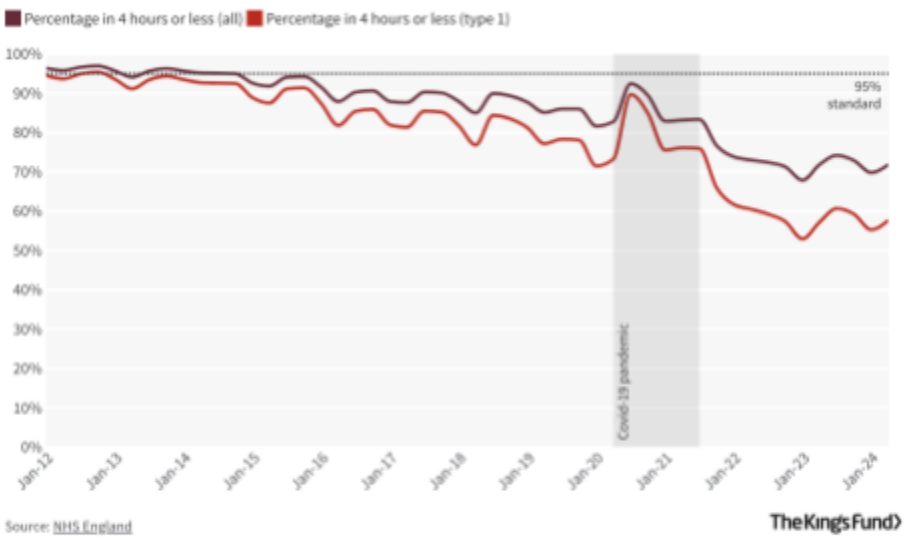


Figure 1: A&E Quarterly Performance (www.kingsfund.org.uk 2023)

The results were promising initially. In England, on average, about two million people visit A&E monthly. Five years after implementing the clinical quality indicators standard, the proportion of patients who typically departed the department within four hours after arrival improved from 3.4% in 2011 to 2.9% in 2016. Ten years later (April 2020), the number of patients leaving without treatment dropped impressively to just 0.7%. Following a decade of remarkable and consistent performance, the tide turned in April 2021, with 2.7% of patients leaving the department without obtaining treatment—a 3% increase from the same month the previous year. And by December 2023, statistics show this figure had climbed to 5.4% of patients departed the hospital without treatment, signalling a troubling trend. What was an uncommon event has become far more regular; ***Figure 1*** shows the rise in patients who wait more than 12 hours following an admission decision. In Q3 2022–2023, 36,131 patients waited more than 12 hours to be admitted to a ward and the number of people waiting more than four hours for an admission (often referred to as 'trolley waits' or 'corridor waits') has increased from 29,929 in Q3 2010/11 to 465,154 in Q3 2022/23.

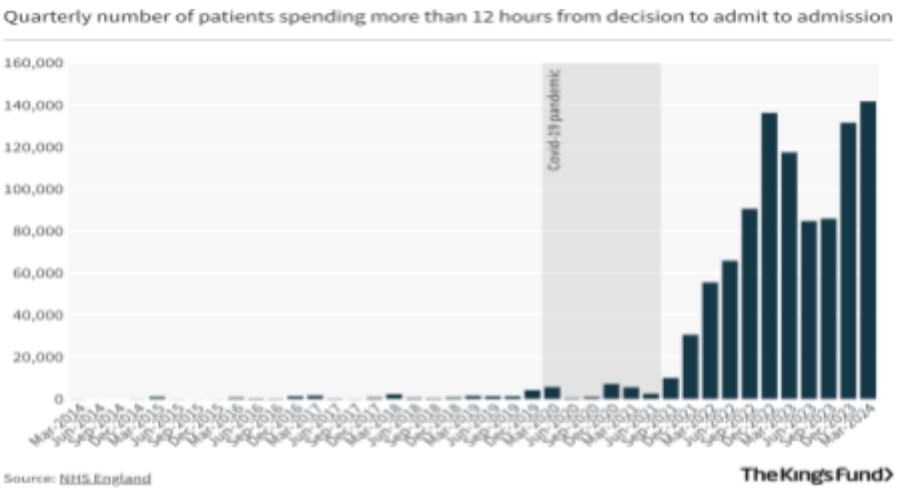


Figure 2: Quarterly Record of Patient Spending More 12 hours in A&E

Despite medical advances that have reduced the average length of hospital stays in recent years, rising emergency admissions continue to strain the UK's healthcare system. For example, the waiting list increased from 150 in Q1 2014 to nearly 150,000 in Q1 2024 (see Figure 2). As these problems become endemic to the NHS, long waiting times in A&E departments will become a common and pressing issue. The prolonged waiting times not only compromise patient satisfaction but also have potential implications on patients' health and the overall functioning of the healthcare system. Understanding the factors influencing waiting times is crucial for optimising resource allocation, staffing levels, and operational processes within A&E departments. One key metric in assessing the performance of A&E departments is waiting time, which refers to the duration from a patient's arrival at the department to the commencement of medical treatment.

Given these challenges, the A&E department requires solutions beyond just administrative fixes, such as addressing staff shortages, bed occupancy, resource limitations, etc. Leveraging data, science, and technology will be essential to modernise and optimise the system. The Lord's Public Services Committee investigated the issues with healthcare and pointed out that emergency care services are operating on an outdated model that does not reflect current demand or clinical practices. Efforts to address this issue have stalled, and continuing with this outdated model will hinder the delivery of necessary care [5]. The committee's report also outlined an action plan to tackle the ongoing challenges of A&E departments–one of the recommendations that stood out and resonated with our study is the development of a new model for emergency care that can recognise the current crisis, understand the type of demand services face, and incorporate clinical best practices, ensuring that emergency health services are fit for purpose in the long run.

One powerful tool that could be deployed in A&E departments to replicate the current situation and support resource managers is the queueing model. This model is a mathematical framework used to analyse and predict the behaviour of systems where entities (such as patients, tasks, or jobs) arrive, wait in line (queue), and are processed by one or more servers according to specific rules. Queueing theory, which originated with the work of Danish engineer Agner Krarup Erlang in 1909, focuses on managing wait times and queue lengths to improve system-level performance measures like throughput and average wait time and has been widely used in various fields, such as telecommunications, computer networks, manufacturing, transportation, and customer service [6]. However, due to the unique nature of A&E departments, a novel approach is required to adapt these models effectively. Much research has been conducted, and numerous studies have explored using data science, mathematics, and technology to enhance the health sector [7].

For instance, a study by (Armony et al. 2015) demonstrated the feasibility of using data and simulation-based research to address challenges in emergency departments. They visually identified three distinct regimes in the A&E dataset: an increase in queue length during early morning hours (3 am to 9 am), a decrease in queue length during PM hours (midday to midnight), and a stable queue length during intermediate hours (3 am to 9 am); this study paves the way for further research, which we are building upon using mathematical techniques and data analysis to develop an algorithm that will automate of those distinct regimes [1].

In this study, we generated and analysed the queue length of patients in the A&E dataset. Due to the skewness and curved nature of the queue length distribution, we fitted a gamma distribution to generate parameters the shape parameter (α) and scale parameter (β) for our clustering analysis. We then deployed two clustering methods: K-means and DBSCAN algorithms. This system will produce actionable statistics insight, enabling NHS resource managers to make data-driven decisions to optimise A&E operations. By integrating these advanced analytical tools, the A&E department can transition from reactive to proactive management, ultimately enhancing patient care and operational efficiency.

1.2 Research Objectives

Expanding on the existing work of (Armony et al. 2015) to automate the finding of the distinct regimes in the A&E data, we have set two objects, which is as follows:

- Obtaining the Queue Length: This project aims to analyse the A&E dataset to uncover significant trends and visualise the main characteristics of the queue length of the patients to identify distinct regimes throughout the day. This will enable us to assess the effectiveness of the A&E operating system and gauge patient satisfaction.

- Determine the different regimes automatically given queuing data: The project aims to improve Armony's research by creating an algorithm that automates hospital queue data to identify distinct regimes in the Accident and Emergency dataset using mathematical modelling such as KMEAN, DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

In summary, this project will transform the management of A&E departments within the NHS by offering evidence-based insights and actionable recommendations derived from advanced data analysis and modelling techniques that will be performed in this study. By establishing the groundwork and generating essential output parameters for developing a queue model that can create a digital twin of A&E queue dynamics and automate the identification of the queue regimes.

1.3 Impact and Significance

The NHS in the UK faces operational challenges due to the high volume of patients with acute or unknown illnesses, which causes excessive workloads for the emergency department. Previous research, by (Mehandiratta, n.d.) and (Haghighinejad et al. 2016), has demonstrated the effectiveness of mathematical models such as queue and data science in optimising patient flow, reducing waiting times, and improving resource allocation in healthcare [7] This research will

Laying the Groundwork for Queue Model Development: This project will calibrate waiting data to generate the parameters necessary for developing a tailored queue model. Doing so will establish a foundation for creating a locally adapted model that enables resource managers to simulate and evaluate the impact of strategies, such as staffing adjustments, on queue length before implementation. This process will pave the way towards achieving the ultimate objective: developing a model for local use that allows the NHS to create a digital twin of the current queue. This digital twin will provide a mathematical representation that can enhance operational efficiency and improve patient care outcomes.

Evidence-Based Decision-Making for Resource Managers: Through data analysis of generating insightful statistics, this project will enable resource managers to predict the arrival and departure rates based on the time corresponding to a decline or increase in patient numbers to make informed operational decisions such as adjusting queue structures or staff numbers based on data-driven.

Efficient automation of queue systems: This project will speed up the process of gaining insights and generating rapid reports by developing an automated queue system to identify regimes in A&E data. This automation relieves resource managers of manual work and provides real-time information on queue dynamics, facilitating proactive patient flow management and resource allocation.

In summary, this project will transform the management of A&E departments within the NHS by offering evidence-based insights and actionable recommendations derived from advanced data analysis and modelling techniques that will be performed in this study. By establishing the groundwork and generating essential output parameters for developing a queue model that can create a digital twin of A&E queue dynamics and automate the identification of the queue regimes, this project has the potential to significantly improve operational efficiency and patient care outcomes in A&E settings.

1.4 Dissertation Structure

Chapter 2: Literature Review: This chapter explores the literature on operational research in the healthcare sector, focusing on the impact of technology, modelling, data analysis and science in Accident and Emergency (A&E) settings. It reviews the wealth of studies investigating how these tools and methods influence healthcare operations, efficiency, and patient outcomes.

Chapter 3: Experimental Method: This chapter details the experimental methodology and describes the dataset selected for the study. It provides an in-depth explanation of the K-Means and DBSCAN models, focusing on their application in grouping the regimes identified through analysis of the A&E dataset.

Chapter 4: Results and Discussion: This chapter interprets the calculated queue length and clustering results using the two mathematical methods. The finding is discussed in the context of the existing literature and the research questions.

Chapter 5: Conclusions: This chapter synthesises the main findings of the study and provides an overview. It discusses opportunities for future research, reflects on important insights observed, and offers suggestions for both operational researchers and practitioners.

References and Appendices: The dissertation concludes with references to scholarly works and appendices for further exploration. The appendices include supplementary material and visuals that provide additional insight into the exploratory analysis.

Chapter 2

2.1 Literature Review

This chapter reviews existing research on improving healthcare systems using data science, mathematics, and technology. It explores how these tools enhance efficiency, reduce wait times, and optimise overall performance, with a particular focus on Accident and Emergency (A&E) settings. Extensive research conducted by renowned experts is summarised as follows:

Impact of Modelling and Exploratory Data Analysis in Healthcare

(Armony et al. 2015) explored patient flow at Rambam Hospital using EDA, revealing key features such as distinct pattern regimes and questioning the effectiveness of simple queueing models in capturing the complexity of Emergency Department (ED) and Internal Ward (IW) operations. Their study underscored the importance of an integrative view of hospital units, linking ED bottlenecks to IW physician protocols. The researchers examined queueing models' applicability to capture the ED's complex operational reality, the relevant time scales and operational regimes for modelling patient length in IWs, and the influence of patient transfer protocols on patient delay, workload division, and fairness. However, the potential limitations of this study were the generalizability of their findings to other hospitals or the time-consuming nature of identifying regimes by eye. They emphasised the need for novel queueing models and theories to improve ED operations [[1].

Clustering and Data Mining Techniques in Healthcare

Data science and clustering methods have significantly impacted healthcare. (Kushwaha & Das, 2020) demonstrated the use of clustering algorithms, such as K-means and Hierarchical Agglomerative Clustering (HAC), for disease detection, patient segmentation, and attribute classification. (Bateja, Dubey, and Bhatt 2021) evaluated the performance of standard clustering algorithms like K-means and DBSCAN on healthcare datasets, proposing a cloud-based solution for implementing clustering on healthcare data [8]. These clustering techniques uncover hidden patterns, predict diseases, and enable timely treatments, leading to cost-effective and quicker results.

Ali and Buti highlighted the increasing use of data mining techniques for disease detection in healthcare, noting that clustering medical data into small chunks helps in pattern discovery and data point retrieval. Traditional data mining techniques extract features from datasets, while clustering reveals overall correlations between data attributes [10,11]. The K-means clustering algorithm is particularly useful for reliable and effective disease detection in massive datasets, underscoring the importance of efficient data mining methods for extracting valuable insights and improving decision-making [9]. Also, paper presents an efficient K-Means clustering algorithm using Self Organizing Map (SOM) to overcome the problem of finding centroids in traditional K-Means. The two-staged algorithm uses SOM to create prototypes and create clusters and uses two healthcare datasets for the experiments. The proposed method is accurate, scalable, and applicable to various domains, demonstrating its unsupervised learning and topology preservation properties [10].

Telemedicine and Technology in Emergency Departments

The application of telemedicine has proven effective in reducing patient transfers in A&E departments. Studies by (Keane 2009) and (Benger 2000) highlighted the efficiency of telemedicine for medical advice and follow-up in minor injury units. Telemedicine offers reliable services that reduce the need for referring clinicians and ensure efficient patient care [11,12].

Regime Detection and Optimization Algorithms

Regime detection algorithms in healthcare, particularly in EDs (emergency departments), have been shown to optimise operational efficiency and resource allocation. Thorwarth et al. (2016) developed an algorithm using hidden Markov models to identify operational states within EDs, helping optimise staffing levels, reduce waiting times, and improve patient care [16]. These algorithms provide valuable insights into healthcare system dynamics, enabling data-driven decision-making and operational improvements [13].

Enhancing Emergency Department Operations

Recent research has focused on optimising ED operations to reduce waiting times and improve patient care. (Sinreich, Jabali, and Dellaert 2012) Developed algorithms that combine simulation and optimisation to address ED overcrowding and enhance operational efficiency by creating effective work shift schedules [[14]. Their study introduced two iterative heuristic algorithms integrating simulation and optimisation models to schedule the shifts of physicians, nurses, and technicians; considering the treatment of patients by multiple care providers, the algorithms reduced patient waiting time by 20-64%.

Similarly, (Arisha and Abo-Hamad 2013) employed an integrated approach using stochastic modelling and evolutionary algorithms to optimise staff schedules while ensuring continuity of care—their method balanced workload distribution and adapted to patient flow variability, leading to more efficient staffing [15].

Additionally, (Ganguly, Lawrence, and Prather 2014) developed an analytic model to create effective ED staffing plans by accounting for patient demand patterns and provider skill profiles. This approach aligned staff availability with patient needs, improving operational efficiency [16]. (Lee et al. 2015) implemented a decision support system combining machine learning, simulation, and optimisation, significantly enhancing ED efficiency by reducing length of stay by 33% and readmissions by 28%. Their work demonstrated the effectiveness of leveraging advanced data-driven techniques to streamline ED operations [17].

Data Science and Big Data in Healthcare

Data science has revolutionised healthcare by leveraging big data, machine learning, and advanced analytics to improve patient care and decision-making (Subrahmanya et al. 2021). The healthcare sector generates vast amounts of data from various sources, including electronic medical records, IoT devices, and genomic databases [[18,19]. This data enables personalised medicine, prescriptive analytics, and automated health reporting (Bhavnani, Muñoz, and Bagai). Data science techniques, such as data mining and artificial intelligence, help process and analyse structured and unstructured data, improving healthcare quality and strategic decision-making [24,25]. However, challenges remain, including data quality, standardisation, and integrating new data sources into healthcare saystems. Despite these challenges, the application of data science in healthcare is driving a paradigm shift from disease-based to patient-based diagnostics, revolutionising the industry [[20]

In summary, these studies collectively highlight the potential of data-driven approaches to enhance health systems, especially ED operations, reduce waiting times, and improve overall patient care. By leveraging the combination of simulation, optimisation, machine learning, and stochastic modelling, these methodologies offer substantial improvements in the management and efficiency of emergency departments.

Chapter 3

3.1 Data Description

This study sources its dataset from the publicly accessible Israeli Hospital (Rambam Medical Centre), a facility with about 1000 beds and 45 medical units. The data includes monthly patient flow and hospital operations from 2004 to 2007, which contains a database that warehouses four different tables in CSV format: (1) database_physical_details (2) database_visit_details (3) database_ward_first_procedure (4) database_xrays_visits.

- Visit Details Table: contains information on each patient entering the hospital daily.
- Physical Details Table: includes information on the hospital unit and ward that the patient arrived at or was admitted to.
- Ward's First Procedure Table: includes information on the patient's admission and procedure time.
- X-ray Visits Table: contains information about patients related to the X-ray department.

There is a common feature between these four tables which gives general information about the patients: The patient's patient and medical case ID, gender, age, entry and exit dates and times, duration, hospital units, treatment outcome, and entry group—which indicates whether the patient arrived at the emergency room or straight to the hospital. Also, In all of our data files, many columns, such as department/units, entry/exist group, and outcome_id, are codenamed from 1 to 200 for each of the hospital's wards because the hospital setting has several wards and departments, Codes 1 through 11 are assigned to all ward related to emergency care, with code 1 assigned to emergency internal medicine units, which is our primary interest for this study.

The visit table offers additional features such as “ED_hours and ED_Duration”, which allow us to calculate the exact duration, including days and hours, a patient spends before exiting the hospital. These are the core variables needed to obtain our queue length. In addition, we require entry and exit dates and times, along with information confirming that patients entered the hospital via the emergency department and are in the emergency internal medicine unit. These details are essential for our analysis and critical for replicating the insight from the Armony's publication, as shown in ***Figure 3***. The visit table file includes all this information, and its richness and comprehensiveness make it the most suitable dataset choice for accurately capturing and analysing patient flow for our study.

3.2 Data Wrangling

We performed a data wrangle on the chosen visit dataset to create new features by separating the date and hour components from datetime columns; we decomposed the time series into daily and hourly components and calculated the exact exit time from the emergency ward. This data wrangling was essential to speed up the modelling process, create an hourly rate, and calculate queue length. This enabled us to reproduce the same visual from the Rambam Hospital research paper (see Figure 3), which our study is building on.

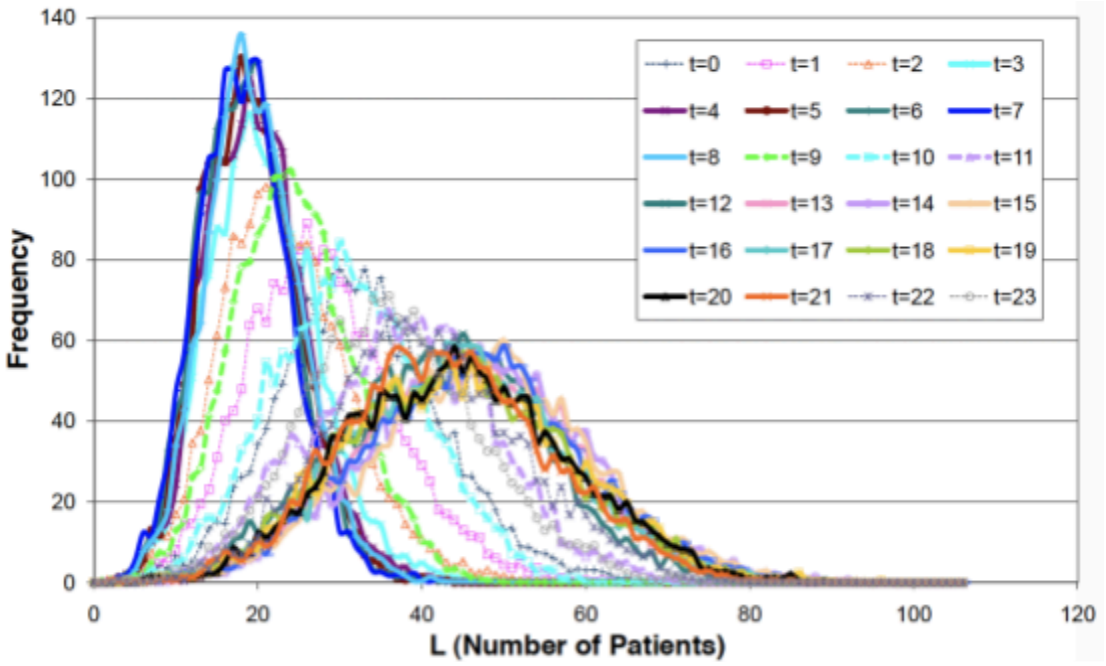


Figure 3: Distribution of the number of patients in the ED per hour of the day (Armony et al. 2015)

3.3 Data limitation

This research study is being conducted with the hope that the UK NHS will benefit from its findings. However, there are several limitations related to the dataset we used, and we have outlined them as follows:

- Due to the lack of publicly accessible NHS data and limited time to seek a partnership or source data in-house, we decided to use a non-UK publicly available dataset with the assumption that all hospitals will possess a unique and common type of information that is applicable across different hospital settings. Using any hospital data other than the Rambam dataset should still allow for applying any methodology and approach in this study. However, extensive data preprocessing and wrangling may be required to ensure it conforms to the same format as the dataset used in this study.

- The lack of case severity and demographic data in the available data might also hinder the assessment of the emergency department's performance, resource allocation, and identification of care access disparities.

- Finally, ambulance arrivals significantly affect crowding in emergency departments (EDs). Effective management of dispatch and arrival schedules can help alleviate peak-time congestion. However, our dataset lacks detailed information on ambulance arrivals, transport times, and specific ambulance data, which could impact patient flow modelling and ED performance. Including this data would enable a more comprehensive analysis.

Chapter 4

Experiment Methods

This chapter discusses the methods used to analyse emergency department datasets, including the tools used for data analysis and visualisation and mathematical modelling for exploring queue dynamics and patterns.

4.1 Environment and Library

Python version 3.10 is the primary programming language used for this analysis in the Google Collaboratory environment. This is a platform's cloud-based Jupyter Notebook environment. It eliminates the hassle of setting up Python environments and installing libraries, as many popular libraries come pre-installed and are easy to use. This study uses several Python libraries, including (1) NumPy, pandas, datetime and SciPy for data analysis techniques, (2) matplotlib for data visualisation, (3) Scikit-learn for model building and evaluation metric to automate and cluster the identified regimes in the A&E dataset.

4.2 Obtaining the Queue Length

Queue Length is the crucial first step in our analysis process, which involves summarising the main characteristics of a dataset with visuals such as lines and histograms to uncover underlying structures, extract important variables, and detect outliers and anomalies.

After data wrangling, we conducted mathematical computations and data analysis to establish queue formation and examine the distribution of A&E department data. This involved generating density charts to visualise patient data throughout the day and identifying patterns in patient flow to pinpoint periods of high, low, and moderate activity. These insights are crucial for addressing our first research objective: identifying distinct regimes in patient flow.

We started by identifying the time frame of the data set and determining the minimum and maximum entry dates of patients in the ED. This allowed us to establish the total duration over which the data was collected. The calculation of these dates, denoted as data0(earliest entry date) and dataT (latest entry date), set the boundaries for our analysis. This can be represented mathematically as:

$$date0 = min\,(entry\_date)$$

$$dateT = max\,(entry\_date)$$

Next, we formed the queue based on two approaches which are:

Using this date range, we created a Data Frame where each row represented a minute within the specified period. This Data Frame was crucial for tracking the number of patients in the ED at any given minute. For each minute, we calculated the cumulative number of patients who had arrived and left the ED by applying lambda functions that summed up the relevant counts from the original data set. The cumulative counts are represented as:

$$A(t) = \sum\left(entry\_date \leq t\right)$$

$$D(t) = \sum\left(exit\_ED \leq t\right)$$

The number of patients in the ED at each minute, denoted $L\,(t)$was then calculated using the formula:

$$L\,(t) = A(t) - D(t)$$

where:

- $L\,(t)$ represents the queue length or the number of patients in the ED at a time t.
- $A\,(t)$ is the cumulative number of patients who have arrived on time t.
- $D\,(t)$ is the cumulative number of patients who have left in time t.

- **Hourly Approach:** This method calculates the queue length at hourly intervals. It assumes that the queue length within each hour is stable and does not change significantly. However, we observed that this approach led to a loss of information about fluctuations within each hour.
- **Minute Approach:** This method calculates the queue length at minute-by-minute intervals. It provides a closer approximation of the continuous-time process, capturing more details about the

fluctuations in queue length. This makes the minute calculation approach more suitable for when precision is crucial, such as in a research study closely analysing queue dynamics.

$$\underline{L}(m) = \frac{1}{N_h} \sum_{t \in h} (L(t))$$

$$HL(m) = \{L(j) : j \bmod 1440 \equiv m,\ j \in \{0, \ldots, num\ minutes\}\}$$

- $L\ (t)$: This represents the average number of patients in the ED during the hour t.
- N_h: This is the number of minutes in the hour h. Since there are 60 minutes in an hour, N_h it is typically 60.
- $\sum_{t \in h} L(t)$: This is the sum of the number of patients in the ED at each minute t within the hour h.
- Here, $L\ (t)$ denotes the number of patients in the ED at the minute or hour t .

For our study, we adopted the minute interval approach to form the queue by calculating the total number of minutes between these two dates. This was done by converting the time difference into minutes and adding one to ensure inclusivity. This step provided the foundation for creating a comprehensive date range for the analysis.

We counted the number of patients in the queue, which represents the number of patients in the emergency department at specific times of day, to use as input for our time series analysis. Finally, we visualised the patterns in A&E data to identify high, moderate and low activity periods.

After identifying distinct regimes and replicating the distribution of patients per hour, we automated the detection of these regimes, avoiding the manual visual assumptions used in (Armony et al. 2015) in the next section.

4.3 Mathematical Modelling Method

In this section, we address our second research question, which aims to automatically identify the distinct regimes seen in the ***Figure 4*** chart from Armony's publication. We discuss the approach and mathematical methods we used to achieve the clustering.

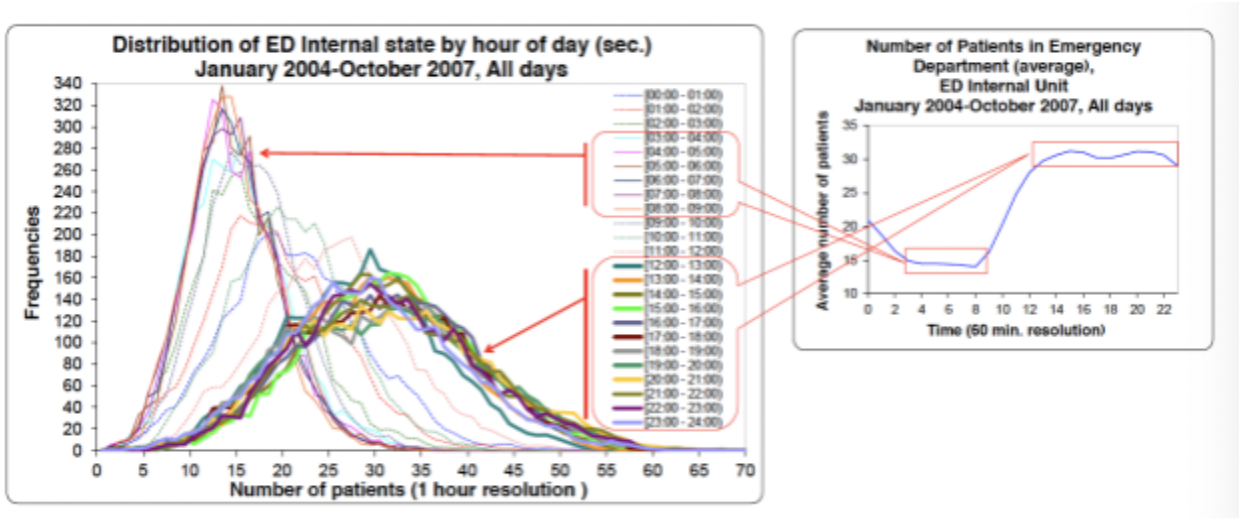


Figure 4; Distribution and average of patients in the ED Internal
Source (Armony et al. 2015)

In ***Figure 4,*** the left plot displays the frequency of patients in the Emergency Department (ED), revealing distinct regimes that exhibit clustering behaviour, with clear groupings and transitional periods between them; these clusters were viewed by eyes from previous research work. To take a step further, we automatically cluster the distinct regime by beginning with preparing the data for clustering, which is fitting the gamma distribution and obtaining the alpha (shape) and beta (scale) parameters, which were extracted from the queue length of patient data. To automate the detection of these regimes, we deployed two models: K-means and DBSCAN.

Our methodology assumes that patient arrival patterns in the Emergency Department (ED) are consistent and stationary over time, with stable Gamma distribution parameters across days. It also assumes that hourly patterns are independent and homogeneous across days and that the Gamma distribution effectively captures the arrival patterns despite potential complexities or outliers.

Gamma Distribution

The Gamma distribution is often chosen for fitting curves because it effectively models skewed distributions commonly found in real-world datasets, even though it is continuous, and our A&E queue data is discrete [21]. While the Negative Binomial distribution might seem like a natural fit for discrete data for our A&E data as it is also discrete but, the Gamma distribution was chosen due to its flexibility and ease in capturing the skewness present in the data. The Gamma distribution's continuous nature does not pose a significant issue in the context of our data, as the primary objective is to accurately capture the underlying skewness of the data rather than strictly adhering to its discrete nature. This approach provides valuable insights into the underlying patterns.

We fit the gamma distribution to the queue length of our data using the scipy.stats library in Python, which employs Maximum Likelihood Estimation. This process transforms our raw data into a parametric form and generates the shape parameter (α) and scale parameter (β). By fitting the curves into a parametric family using the gamma distribution, we extracted the shape (α) and scale (β) parameters, which were used as input for our model, facilitating further analysis and comparison of curve features using clustering algorithms like DBSCAN and K-Means.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely used technique for reducing dimensionality and visualising high-dimensional data, especially in cluster analysis. PCA-based visualisation is used for cluster identification and determining optimal cluster numbers and has been applied in melanoma classification, gene expression data, and video processing applications. Clustered blockwise PCA, which partitions data into blocks, applies PCA to each block, and merges subspaces, offers improved efficiency and scalability for large datasets.

Though the nature of our data already exists in a two-dimensional, we might think PCA isn't necessary as the usual purpose of PCA is dimensional reduction [22], but in our study, we still implement PCA in our modelling algorithm; this application of Principal Component Analysis (PCA) was necessary and beneficial for several reasons. PCA helps us by reducing noise and focusing on the most significant directions of variability, which might not align with the original axes. It also reveals underlying structures, improves visualisation by rotating data to align with the principal components, and provides a more interpretable basis for our data. Additionally, PCA enhances our cluster separation and simplifies the analysis by centring and normalising our A&E data, making the patterns or groups within the data more apparent and more straightforward to interpret [23].

K-Means Clustering

K-means clustering is an unsupervised machine learning algorithm originating from signal processing that divides a set of n observations into k clusters, assigning each observation to the cluster with the closest mean (cluster centre or centroid) representing that cluster [30]. The steps we took to implement the K-Means cluster for our study include data standardisation, clustering of the generated parameter from gamma distribution fitting using the K-Means implementation from the scikit-learn library and visualising the clusters by employing Principal Component Analysis (PCA).

Density-Based Spatial Clustering

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a nonlinear, unsupervised machine learning algorithm that uses spatial location and distance to identify clusters in data. It is primarily used when the data has irregular shapes or when there is no prior knowledge about the number of clusters. It is a powerful algorithm tool for identifying diverse clusters and handling noise effectively. Its performance can be sensitive to the choice of parameters, requiring careful tuning for optimal results.

However, datasets with regions of varying densities may pose challenges to DBSCAN. To explore an alternative clustering method, we applied the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to automate the identified regimes in the A&E data by using the generated parameter from the gamma distribution output as the input. It operates on the principle of density connectivity, defining clusters as areas where points are densely packed together, separated by regions of lower density. Each point in the dataset is classified as either a core point, a border point, or noise based on specific parameters such as epsilon (ε) and minPts.

DBSCAN iteratively explores the dataset, expanding neighbourhoods around core points to encompass other nearby points and thus forming and identifying distinct clusters of distribution parameters, facilitating further analysis and interpretation of patterns. This method does not require specifying the number of

clusters in advance and can identify noise points like the K-Mean algorithm. The steps involved in achieving the DBSCAN cluster for our study include data standardisation, clustering using the DBSCAN implementation from the scikit-learn library and visualising the clusters by employing Principal Component Analysis (PCA) to reduce the dimensionality of the data to two principal components.

This study's two mathematical methods–K-means and the DBSCAN approach—have pros and cons. K-means is straightforward and provides clear, easily interpretable clusters with consistent sizes, making comparison and analysis easier. DBSCAN effectively identifies outliers as noise and can find clusters of various shapes and densities, offering a more detailed understanding of the data. K-means minimise variance within clusters, leading to evenly sized and shaped clusters.

While DBSCAN relies on data density, resulting in varied cluster sizes and shapes and the identification of noise points. K-means requires the number of clusters to be predefined, influencing the results. In contrast, DBSCAN does not require a predefined number of clusters but needs parameters like epsilon and minimum points, which influence cluster formation.

The DBSCAN's parameters significantly influence cluster formation but can introduce subjectivity. To mitigate this, hyperparameters can be carefully chosen using methods like the "elbow method" or domain knowledge. Automated approaches such as grid search or randomised search can further reduce subjectivity and enhance the robustness of the clustering process. In our model, we used domain knowledge to define the hyperparameters (epsilon = 0.5 and min_samples = 2), tailoring them to fit our dataset's unique patterns and needs and ensuring that the clustering process was relevant and practical. However, exploring other methods could be beneficial in evaluating different outcomes [24].

The mathematical modelling method, which includes Gamma fitting and clustering of the A&E dataset, aims to understand and manage the variability in patient arrivals systematically. By analysing and clustering hourly patient data, this approach provides crucial insights for optimising Emergency Department (ED) operations. Fitting Gamma distributions helps model arrival patterns, while clustering identifies similar hours, enabling data-driven staffing decisions and better resource allocation during peak and quiet times. This method improves queue management by anticipating patient surges, leading to more efficient flow and reduced wait times. Additionally, it supports predictive analytics, allowing the ED to forecast future patient loads based on historical data, thereby enhancing patient care and operational efficiency.

Model Evaluation

Evaluating a clustering model can be challenging, especially when dealing with unsupervised classification [8], where there are no predefined labels or automated approaches to validate the results. In our study, the clustering model's evaluation is particularly complex due to two key factors:

- **Unsupervised Classification**: Since our problem is unsupervised, we don't have ground truth labels for comparison, making traditional evaluation metrics like accuracy, precision, or recall unsuitable.
- **Lack of Automated Benchmark**: The identification of distinct regimes in our dataset has historically been done manually using Armony's approach, which relies on visual inspection rather than an automated process. This further complicates our ability to evaluate the model using standard methods.

Given these challenges, we used the **Silhouette Score** as our evaluation metric. The Silhouette Score, however, measures how well-separated the clusters are by comparing the mean distance between each point and others within its cluster to the mean distance to points in the nearest neighbouring cluster. While not a definitive measure of the model's overall performance, it does provide insight into how well the clusters are formed.

Since there is no ideal way to evaluate our model, we focused on determining whether our approaches could classify or group similar items more effectively than manual identification by eye. Therefore, we calculated the Silhouette Score for our K-Means and DBSCAN models, as well as for Armony's approach, which was manually calculated by inputting their classifications as labels into the Silhouette Score algorithm. By comparing the Silhouette Scores, we could assess which method—K-Means or DBSCAN—better groups similar hours together, thereby reflecting a more accurate and objective classification.

Chapter 5

Result and Discussion

5.1 Obtaining the Queue Length

The primary goal of exploring the queue length in this study is to discover the pattern of patient flows in our A&E data and generate insights that ensure a comprehensive understanding of the data. Additionally, we aim to replicate the output and confirm the insights from the (Armony et al. 2015) paper, as shown in ***Figure 3***.

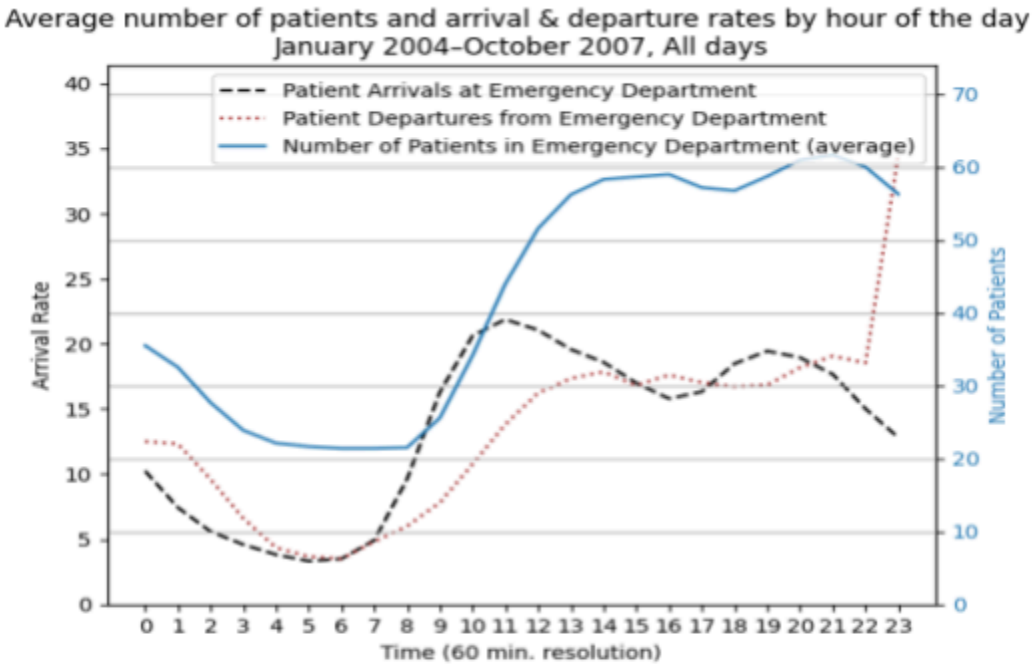


Figure 5; Hourly Trends in ED Arrivals, Departures, and Occupancy

Figure 5 shows the hourly patient flow trends in the emergency department (ED) from January 2004 to October 2007. The arrival rate line shows the highest rate after midnight, decreases until early morning, and then rises again, peaking in late afternoon and evening. The departure rate line follows a different pattern, remaining low during early morning hours and increasing steadily throughout the day, with the highest rates occurring in late evening. The number of patients in the ED is lowest in the early morning hours and increases steadily throughout the day, reaching its peak during late afternoon and early evening. These findings underscore the importance of strategic staffing in managing patient flow and ensuring efficient ED operations.

Using the queue length data, we plotted a line chart distribution to gain insights into patient flow in the A&E department. The results in ***Figure 6*** below illustrate the distribution of the number of patients in the Emergency Department (ED) per hour of the day. The graph showcases distinct patterns of patient arrivals throughout the 24 hours, with significant variability in frequency across different hours. These distribution curves align with the patterns observed in the Armony et al. study, confirming that we are on the right track. The analysis reveals how patient flow fluctuates, identifying peak, moderate and lower activity periods, which is crucial for optimising resource allocation and managing operations efficiently. The left chart highlights the variability and distribution of patients within each hour, while the right chart provides a smooth, averaged view of patients across the day.

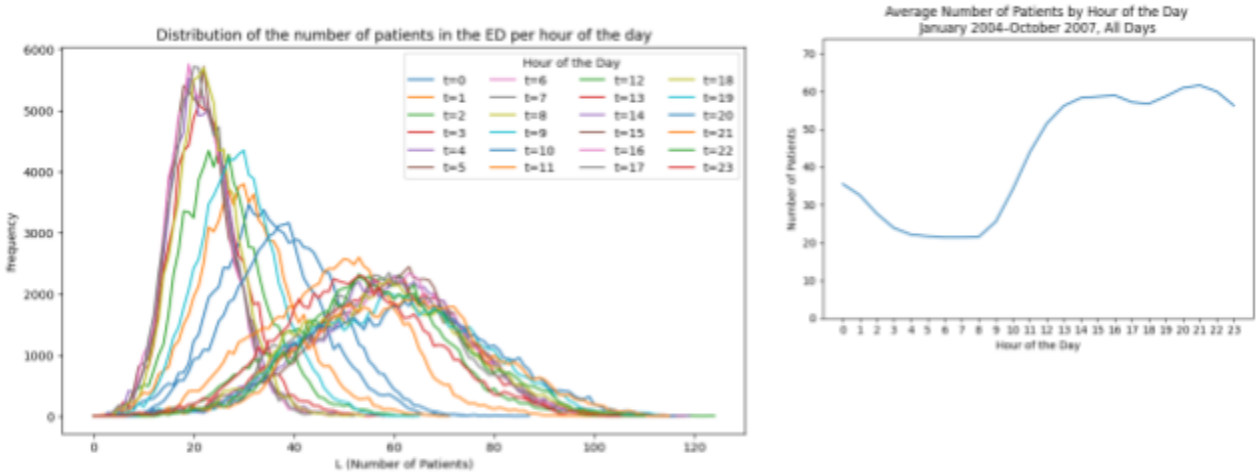


Figure 4; Distribution and average of patients in the ED Internal Source (Armony et al. 2015)

As seen in ***Figure 6***, the left chart illustrates the distribution of the total number of patients in the Emergency Department (ED) at different hours of the day over the period from 2004 to 2007. Each curve corresponds to a specific hour and shows the variation in patient numbers during that time. The curves are primarily unimodal, with each peaking at a certain number of patients, representing the most frequently observed patient count during that hour across the entire analysed period. During the early morning hours, particularly around 0:00 to 1:00 AM (t=0 and t=1), the peak occurs at about 20 patients. This peak is quite sharp, meaning the ED frequently sees around 20 patients during these hours, with less variation. This consistent pattern allows the ED to better predict and prepare for patient load during these hours.

However, the distribution is much broader during the afternoon, mainly around 2:00 to 3:00 PM (t=14 and t=15), with patients ranging from 20 to 60. This indicates a higher degree of variability, meaning that the ED may see anywhere from 20 to 60 patients during these hours.

The right chart, which depicts the average number of patients in the ED by hour, provides an overall trend across the day. It shows that the lowest average patient numbers occur between 4:00 and 6:00 AM. aligning with the narrow distributions seen in the left chart during these hours. As the day progresses, the number of patients increases, peaking in the late afternoon, specifically between 4:00 and 5:00 PM. This peak corresponds with the wider variability observed in the left chart during the afternoon.

A notable difference arises when comparing our results with the existing paper(see Figure 4). Our results show that the lowest recorded number of patients is 20 during the early morning hours, and the peak is 60. In contrast, existing publications report a lower minimum of 15 patients and a peak period of 30. This discrepancy may be attributed to differences in the datasets used, which aggregate data from the Emergency Department and the Internal Wards (IW). Our analysis focuses exclusively on the ED dataset, and including IW data in previous studies likely resulted in a broader range of patient numbers, especially during periods of lower and peak activity, thus accounting for the observed differences.

The charts provide complementary insights into the patient flow within the Emergency Department (ED), and by analysing these charts side by side, one can identify how specific hours on the left chart correlate with broader daily trends shown on the right. This connection allows for a better understanding of patient flow patterns and the variability in patient numbers throughout the day. These insights recognise the stable periods during early morning hours and the more variable, busier times in the afternoon.

In summary, while all charts reveal similar trends—fewer patients in the morning and more in the afternoon—we present this information differently. Together, they provide invaluable insights into ED operations, showing when the ED is busiest and how predictable or variable the patient numbers are at different times. These complementary views are crucial for the resource managers to strategically plan to ensure the ED operates efficiently and is well-prepared to meet patient demands and align with previous studies (Armony et al. 2015), as shown in Figure 4 in the experiment method chapter. Notably, while our study and existing publication identify peaks in patient numbers during the early morning and evening hours, our analysis shows a peak at 6 AM, which occurs slightly earlier than the peaks observed in the existing study.

5.2 Mathematical Modelling

Our second research objective was to automate the identification of distinct regimes in the data, which Armony et al. had manually identified by eye. Using K-means and DBSCAN clustering on the gamma distribution parameters (alpha and beta) allowed us to categorise the hours into distinct clusters based on their statistical properties (shape α and scale β) to identify patterns and similarities across different hours rather than subjective interpretation. We employed two clustering techniques. Here, we present and compare the results of these clustering methods to provide deeper insights into the patient distributions /queue length.

Gamma Distribution Fitting

The gamma distribution is then fitted to the queue length and plotted in a 6x4 grid of subplots where each subplot represents one hour of the day. The primary y-axis shows the frequency of patient counts. In contrast, the secondary y-axis fits a gamma distribution α (shape) and β (scale) parameters, which help identify specific patterns or anomalies at different times. The gamma distribution fit provides a statistical perspective on the patient count distribution and parameters for our clustering.

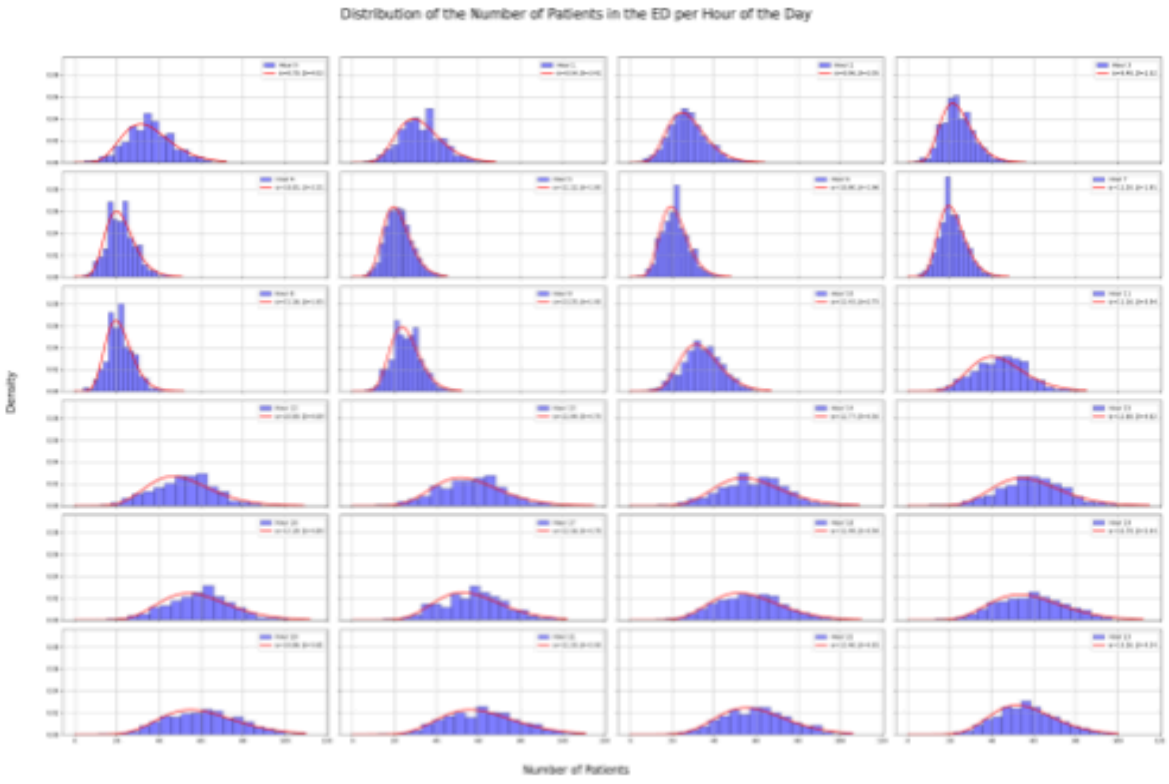
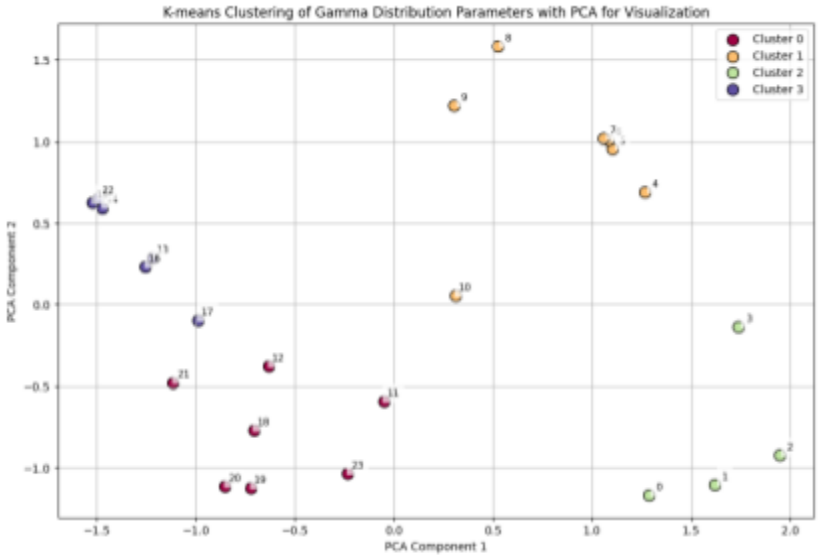


Figure: Hourly Patient Distribution with Gamma Fits

The result obtained from the fit, as shown in ***Figure 7 (see appendix for a larger image for each hour)***, tells us more about the patient arrival patterns, such as Hourly Patterns, which give distribution per hour (1) Early Morning (Hours 0-5): lower patient counts, but with notable variability (2) Midday and Afternoon (Hours 12-18): Higher patient counts, suggesting busier periods in the ED (3) Evening (Hours 19-23): Patterns of decline or stabilisation in patient numbers.

K-means Clustering

The results of a K-means clustering analysis on Gamma distribution parameters, with PCA (Principal Component Analysis) employed for visualisation is showed in ***Figure 8*** below. The modelling successfully identified four distinct clusters of hourly data, each corresponding to different day periods. The clusters align with the hourly variations in the characteristics of the data, revealing patterns consistent with known temporal behaviours.



Comparison Table Between Our Approach and Armony

Hour	Alpha	Beta	Cluster by K-Mean	Existing study identified by eye	Insight by eye (Armony et al. 2015)
0	9.30	3.73			intermediate hours (0:00-2:00)
1	9.09	3.37			intermediate hours (0:00-2:00)
2	8.99	2.60			intermediate hours (0:00-2:00)
3	9.73	2.37			early morning (3:00-8:59)
4	10.70	2.06			early morning (3:00-8:59)
5	11.02	1.96			early morning (3:00-8:59)
6	11.08	1.93			early morning (3:00-8:59)
7	11.10	1.94			early morning (3:00-8:59)
8	11.93	1.91			early morning (3:00-8:59)
9	11.82	2.44			early morning (3:00-8:59)
10	10.94	3.50			intermediate hours (9:00-11:59)
11	10.73	4.42			intermediate hours (9:00-11:59)
12	11.32	4.78			PM hours (12:00-23:59)
13	12.26	4.69			PM hours (12:00-23:59)
14	12.68	4.64			PM hours (12:00-23:59)
15	12.74	4.66			PM hours (12:00-23:59)
16	12.25	4.77			PM hours (12:00-23:59)
17	11.80	4.89			PM hours (12:00-23:59)
18	11.08	5.19			PM hours (12:00-23:59)
19	10.83	5.52			PM hours (12:00-23:59)
20	10.94	5.83			PM hours (12:00-23:59)
21	11.81	5.29			PM hours (12:00-23:59)
22	12.74	4.80			PM hours (12:00-23:59)
23	10.53	4.99			PM hours (12:00-23:59)

Figure 8: K-Mean Method (4 Clusters)

The ***Figure 8*** chart with the clustering ball displays the hour they represent for each day, and Table 1 displays the clustering of Gamma distribution parameters (α and β) for each hour of the day. These clusters reveal distinct patterns in the patient queue in the Emergency Department (ED). By analysing these clusters, we can draw comparisons to existing literature, which is the basis of our study.

The table highlights distinct clusters—blue, red, orange, and green—that group different day hours based on their α and β values. Each cluster represents a unique pattern in patient flow. Cluster 0, shown in red, contains hours with relatively consistent and moderate patient counts. Cluster 1, depicted in orange, includes hours with high variability, often corresponding to early morning periods. Cluster 2, represented in green, encompasses peak hours with higher patient counts and greater variability. Finally, Cluster 3, marked in blue, signifies hours with lower patient counts and less variability, typically occurring during off-peak times.

K-means Clustering Insights

Cluster 0 (Red): This cluster includes hours primarily in the late evening and early night (from 18:00 to 23:59). The data within this cluster shows a high concentration of activity, characterized by relatively high alpha and beta values, indicating periods with heightened dynamism or intensity. This cluster aligns closely with what previous studies (such as Armony et al., 2015) identified as "PM hours," reflecting the end-of-day bustle.

Cluster 1 (Orange): This cluster captures early morning hours (4:00 to 10:59). These hours are typically quieter, with moderate alpha and lower beta values. The clustering method has successfully identified these hours as separate, confirming the period's distinct temporal characteristics. This agrees with prior research's "early morning" classification, suggesting consistency in the clustering approach.

Cluster 2 (Green): Cluster 2 corresponds to the late morning (00:00 to 2:59), where a unique combination of alpha and beta values reflects a notable shift in activity patterns. This aligns with the "intermediate hours" from existing studies, highlighting a transition phase between the early morning and the more active periods.

Cluster 3 (Blue): This cluster represents the afternoon (12:00 to 17:59) with in between the hours has cluster 0(Red); together this two classification is characterized by the highest beta values, suggesting this is the peak activity period. The existing study similarly classifies this period as "PM hours," marked by maximum productivity and engagement.

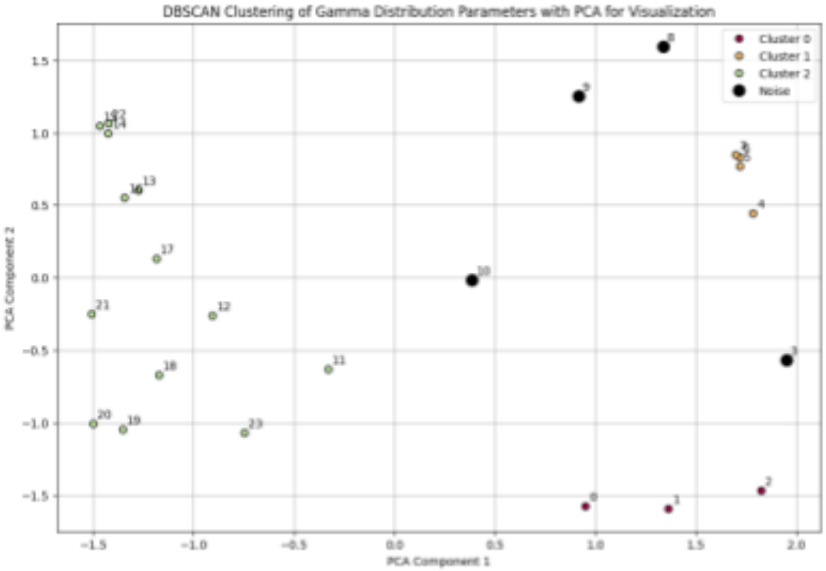
The clustering results and their interpretation provide valuable insights into operational patterns, particularly supporting the notion that patient occupancy in the Emergency Department (ED) follows distinct temporal trends, as observed by Armony et al. (2015). While our approach generally aligns with these established patterns, there are notable discrepancies. Specifically, our method, which utilized K-means clustering, identified two separate clusters within the PM hours, whereas the manual classification by Armony et al. grouped these hours into a single category.

This discrepancy highlights the potential advantage of automated clustering techniques in identifying subtle shifts in data distribution that might not be immediately apparent through visual inspection. The

distinct separation between the clusters, as revealed in our analysis, indicates clear differences in the distribution patterns of patient queues across different hours. K-means clustering proves effective in identifying and categorizing these hours into specific groups, which aids in understanding and managing patient inflow based on predictable patterns.

DBSCAN Clustering

DBSCAN is a density-based clustering algorithm that identifies clusters based on the density of data points, allowing for the identification of noise (outliers).The automated DBSCAN approach is less prone to subjective bias, offering a consistent method to identify regimes across different datasets or time periods. And it ability to identify noise and potentially transitional states provides more meaningful insights than a manual approach.The automated clustering approach adds value by confirming the major regimes identified manually while also revealing subtleties and outliers that could lead to new insights or hypotheses. It is particularly valuable when dealing with large datasets where manual inspection would be impractical.



Comparison Table Between Our Approach and Arnaoy

Hour	Alpha	Beta	Cluster by DBSCAN	Existing study identified by eye	Insight by eye (Arnaoy et al. 2015)
0	9.30	3.73			intermediate hours (3:00-2:00)
1	9.89	3.37			intermediate hours (3:00-2:00)
2	8.89	2.98			intermediate hours (3:00-2:00)
3	9.73	2.37			early morning (3:00-8:59)
4	10.70	2.06			early morning (3:00-8:59)
5	11.02	1.98			early morning (3:00-8:59)
6	11.08	1.90			early morning (3:00-8:59)
7	11.10	1.94			early morning (3:00-8:59)
8	11.03	1.97			early morning (3:00-8:59)
9	11.62	2.44			early morning (3:00-8:59)
10	10.94	3.58			intermediate hours (9:00-11:59)
11	10.73	4.42			intermediate hours (9:00-11:59)
12	11.32	4.76			PM hours (12:00-23:59)
13	12.26	4.69			PM hours (12:00-23:59)
14	12.88	4.64			PM hours (12:00-23:59)
15	12.74	4.66			PM hours (12:00-23:59)
16	12.25	4.77			PM hours (12:00-23:59)
17	11.80	4.80			PM hours (12:00-23:59)
18	11.08	5.19			PM hours (12:00-23:59)
19	10.83	5.52			PM hours (12:00-23:59)
20	10.94	5.60			PM hours (12:00-23:59)
21	11.81	5.28			PM hours (12:00-23:59)
22	12.74	4.66			PM hours (12:00-23:59)
23	10.53	4.98			PM hours (12:00-23:59)

Figure 9: DSCAN Method (4 Clustering)

The chart presented in ***Figure 9*** shows a comprehensive visualisation of the DBSCAN clustering of Gamma distribution parameters of our A&E data, using PCA for enhanced visualisation clarity. This dual representation illustrates the clustering results and provides a comparative analysis between our approach and the existing findings of Armony et al. (2015). The DBSCAN method identified three distinct clusters (Cluster 0, Cluster 1, Cluster 2) or (intermediate hours, early morning, and PM hours) along with a set of points classified as noise differentiated by colour. These clusters show distinct grouping patterns, indicating varied patient visit behaviours across hours.

Cluster Insight:

- **Cluster 0 (Red):** Clusters in this bucket predominantly aligned with AM intermediate hours (0:00 - 2:00) and but not for the PM intermediate hours (9:00-11:59) of existing study. These points have low Alpha and Beta values, suggesting a trend of decreased patient visits during these hours.
- **Cluster 1 (Orange):** This cluster is primarily associated with early morning hours (3:00-8:59), with relatively higher Alpha and low beta values. It captures a significant pattern of patient visits during the early morning, possibly indicating critical or emergency cases.
- **Cluster 2 (Green):** Mostly represents PM hours (11:00 - 23:59), with higher Beta values, indicating increased variability in patient visits during these hours.
- **Noise Points (Cluster 3 Black):** The black points represent noise which however we consider to be our cluster 4, signifying outliers or hours that do not conform to the main clustering patterns. These include hours 8,9,10 and 3, indicating unusual patient visit behaviours during these times. We could assume that these noise points suggest the possibility of transitional periods during which patient visit patterns do not neatly fit into the identified clusters. This could be due to various factors, such as shift changes in hospital staff, variability in patient needs, or other operational dynamics.

Comparison with Armony et al. (2015)

The comparative table in the ***Figure 9*** highlights a significant alignment between our DBSCAN-identified clusters and the existing study's categorisation by eye.PM hours identified by Armony et al. largely correspond with our Cluster 2, affirming the consistency of patterns observed across different studies. Early morning hours show a broader dispersion in our analysis, reflected in both Cluster 0, 1 and Cluster 3, suggesting a better understanding of patient visit patterns during these times.

The DBSCAN method has successfully identified distinct clusters in patient visit patterns, and these findings were validated against the insights from Armony et al. (2015). This cross-verification strengthens the reliability of our clustering approach. Overall, our analysis supports existing findings and adds depth to the understanding of temporal patterns in patient visits.

A key difference between the manual method and DBSCAN is how they handle noise. Armony et al.'s manual method does not identify any noise, assuming all points belong to some regime. In contrast, DBSCAN detects noise, indicating that certain hours do not fit neatly into the defined clusters. This noise aligns with the period that Armony et al. classified as transitional or intermediate time. Although both methods accurately capture the primary regimes—early morning, PM hours, and intermediate hours. However, DBSCAN reveals subtler distinctions, particularly in Cluster 3, which may be missed by the manual approach. This cluster's identification of noise points provides additional insights into transitional periods or outliers, which are aggregated into broader categories by the manual method.

These insights are valuable for improving resource allocation, patient flow management, and overall efficiency in emergency departments. The identification of noise points, in particular, offers critical information about outlier visit behaviors, potentially highlighting areas that require further investigation to address underlying causes or operational inefficiencies.

Temporal Insights from DBSCAN and K-Means in 3D

The visualizations in Figure 10 present a comparative analysis of two clustering algorithms, DBSCAN (left) and K-means (right), applied to a 3D dataset comprising the Alpha, Beta, and Time (Hour of the day) parameters. This approach was intended to enable a deeper understanding of how these clustering techniques segment the data when introducing an additional dimension.

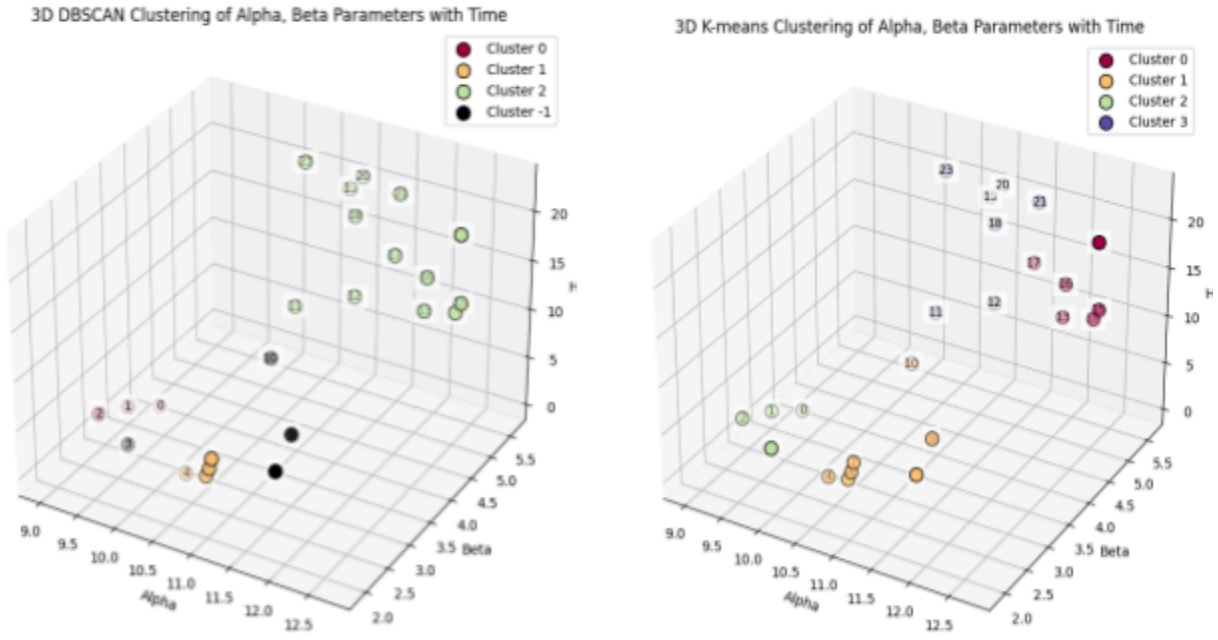


Figure 10: 3D Clustering L(t)

In our further investigation, we introduced a third dimension (Time) with the expectation that it might provide additional insight into the temporal dynamics of our data. We anticipated enhancing the ability to interpret how data points transition over time, potentially revealing insights that a 2D approach might miss.

However, the results presented in ***Figure 10*** show no significant difference compared to our 2D approach. Both approach yield similar insights regarding patient flow changes throughout the day.

Limitation

K-means and DBSCAN clustering methods provide valuable insights into the distribution patterns of patient flow in the Emergency Department (ED). K-means offers clear and interpretable clusters, while DBSCAN adds depth by identifying outliers and capturing clusters of varying shapes and sizes. Together, these methods offer a comprehensive understanding that can enhance management and operational strategies in the ED.

However, there are limitations to consider. While automated methods like K-means and DBSCAN can effectively identify distinct regimes, it's important to acknowledge that the results produced by these algorithms may not always be immediately intuitive or easily interpretable without domain knowledge. Automation can streamline the process and uncover patterns that might not be visible through manual inspection, but interpreting these results often still requires an understanding of the underlying domain.

This is especially true when the clusters identified by the algorithms are based on statistical similarities, which might only sometimes correspond to patterns that are easily explained or recognised with deep domain expertise. Additionally, the sensitivity of methods like K-means to initial parameters highlights the need for careful validation, ensuring that the automated results align with practical, real-world insights.

5.4 Model Evaluation

The Silhouette Scores in ***Table 1*** for our K-Means and DBSCAN models is approximately 57% and 58% respectively, indicating reasonably well-defined clusters. In comparison, the Silhouette Score for Armony's approach, which was done manually by eye, stands at 49%. This lower score suggests that the clusters identified through visual inspection are less distinct.

These scores, along with insights from visualising the clusters that detect subtle variations within the A&E data, lead to a more nuanced understanding of the regimes. For example, the subdivision of the PM hours into different clusters suggests that there might be varying dynamics or patterns within this period that were not recognised by the eye in the existing study**.**

Therefore, we can conclude that the mathematical approaches—K-Means and DBSCAN—are more effective at automatically classifying or grouping hours with similar patient flow patterns in our study compared to manual identification. The higher Silhouette Scores for the algorithmic methods indicate that they group data points in a way that better reflects the underlying structure of the data, leading to more accurate and objective classifications than those achieved by eye.

Clustering Algorithm	**Silhouette Scores**
K-Means Clustering	57%
DBSCAN Clustering	58%
Armony's approach Clustering	49%

Table 1: Model Evaluation Outcome

6 Conclusion

The automated approach using K-means and DBSCAN clustering has demonstrated significant value in objectively and ability to segment the data into clusters that reflect distinct periods of the day and correlate well with existing temporal classifications from prior studies Armony et al. (2015).. This alignment with established patterns enhances the credibility of our clustering approach and provides deeper insights into the temporal dynamics of patient flow in the ED. The results suggest that this method is robust in identifying and categorizing periods with similar characteristics, offering a valuable tool for understanding temporal variations in data-driven contexts. Consequently, this analysis could be applied to optimize resource allocation, improve scheduling, or enhance temporal pattern recognition in various fields.

In conclusion, the automated approach provides a more granular and statistically robust identification of regimes, which is particularly useful in complex datasets where manual inspection might overlook subtle but important distinctions. Our study also opens opportunities for further research, such as exploring the underlying causes of these patterns and extending the research to develop a queueing model replicating the actual situation, aiming to improve operational efficiency.

Reference

[1] *M. Armony, S. Israelit, A. Mandelbaum, Y.N. Marmor, Y. Tseytlin, G.B. Yom-Tov, On Patient Flow in Hospitals: A Data-Based Queueing-Science Perspective, Stochastic Systems 5 (2015) 146–194. https://doi.org/10.1287/14-ssy153.*

[2] Ann. Dunkin, ACM Digital Library., ACM Special Interest Group on Simulation and Modeling., Winter Simulation Conference., Winter Simulation Conference, 2009.

[3] NHS England, Dementia diagnoses in England at record high, (2024).

[4] NHS England, Provisional Accident and Emergency Quality Indicators for England, Provisional Accident and Emergency Quality Indicators for England (2011) 2024.

[5] Lords report on emergency care-Response, n.d. https://www.england.nhs.uk/long-read/next-steps-in-increasing-capacity-and-operational-resilience-.

[6] M. Zukerman, Introduction to Queueing Theory and Stochastic Teletraffic Models, n.d.

[7] R. Mehandiratta, APPLICATIONS OF QUEUING THEORY IN HEALTH CARE, International Journal of Computing and Business Research (n.d.).

[8] R. Bateja, S.K. Dubey, A.K. Bhatt, Evaluation and Application of Clustering Algorithms in Healthcare Domain Using Cloud Services, Second International Conference on Sustainable Technologies for Computational Intelligence (2021). https://api.semanticscholar.org/CorpusID:239101317.

[9] S.I.M. Ali, R.H. Buti, DATA MINING IN HEALTHCARE SECTOR, in: 2021. https://api.semanticscholar.org/CorpusID:236353256.

[10] Z.J. Pan, Principal Component Analysis Based Visualization and Human Melanoma Classification, in: 2001. https://api.semanticscholar.org/CorpusID:17354305.

[11] M.G. Keane, A review of the role of telemedicine in the accident and emergency department, J Telemed Telecare 15 (2009) 132–134. https://doi.org/10.1258/jtt.2009.003008.

[12] J. Benger, A review of telemedicine in accident and emergency: the story so far, Journal of Accident & Emergency Medicine 17 (2000) 157. https://doi.org/10.1136/emj.17.3.157.

[13] A.R. Andersen, B.F. Nielsen, L.B. Reinhardt, T.R. Stidsen, Staff optimization for time-dependent acute patient flow, Eur J Oper Res 272 (2019) 94–105. https://doi.org/10.1016/j.ejor.2018.06.015.

[14] *D. Sinreich, O. Jabali, N.P. Dellaert, Reducing emergency department waiting times by adjusting work shifts considering patient visits to multiple care providers, IIE Transactions 44 (2012) 163–180. https://api.semanticscholar.org/CorpusID:120783597.*

[15] A. Arisha, W. Abo-Hamad, Towards Operations Excellence: Optimising Staff Scheduling For New Emergency Department, in: 2013. https://api.semanticscholar.org/CorpusID:73592461.

[16] S. Ganguly, S.R. Lawrence, M. Prather, Emergency Department Staff Planning to Improve Patient Care and Reduce Costs, Decis. Sci. 45 (2014) 115–145. https://api.semanticscholar.org/CorpusID:22416060.

[17] E.K. Lee, H.Y. Atallah, M.D. Wright, E.T. Post, I. V CalvinThomas, D.T. Wu, L.L. Haley, Transforming Hospital Emergency Department Workflow and Patient Care, Interfaces (Providence) 45 (2015) 58–82. https://api.semanticscholar.org/CorpusID:2885390.

[18] S. V Subrahmanya, D.K. Shetty, V. Patil, B.M.Z. Hameed, R. Paul, K. Smriti, N. Naik, B.K. Somani, The role of data science in healthcare advancements: applications, benefits, and future prospects, Ir J Med Sci 191 (2021) 1473–1483. https://api.semanticscholar.org/CorpusID:237056184.

[19] S.R. Veeranki, M. Varshney, Application of data science and bioinformatics in healthcare technologies, Int J Health Sci (Qassim) (2022). https://api.semanticscholar.org/CorpusID:250653007.

[20] S.P. Bhavnani, D. Muñoz, A. Bagai, Data Science in Healthcare Implications for Early Career Investigators, in: 2016. https://api.semanticscholar.org/CorpusID:2186785.

[21] S.A. Metwalli, What Is Gamma Distribution, Https://Builtin.Com/Data-Science/Gamma-Distribution (2023).

[22] Elitsa Kaloyanova, How to Combine PCA and K-means Clustering in Python?, Https://365datascience.Com/Tutorials/Python-Tutorials/Pca-k-Means/ (2024).

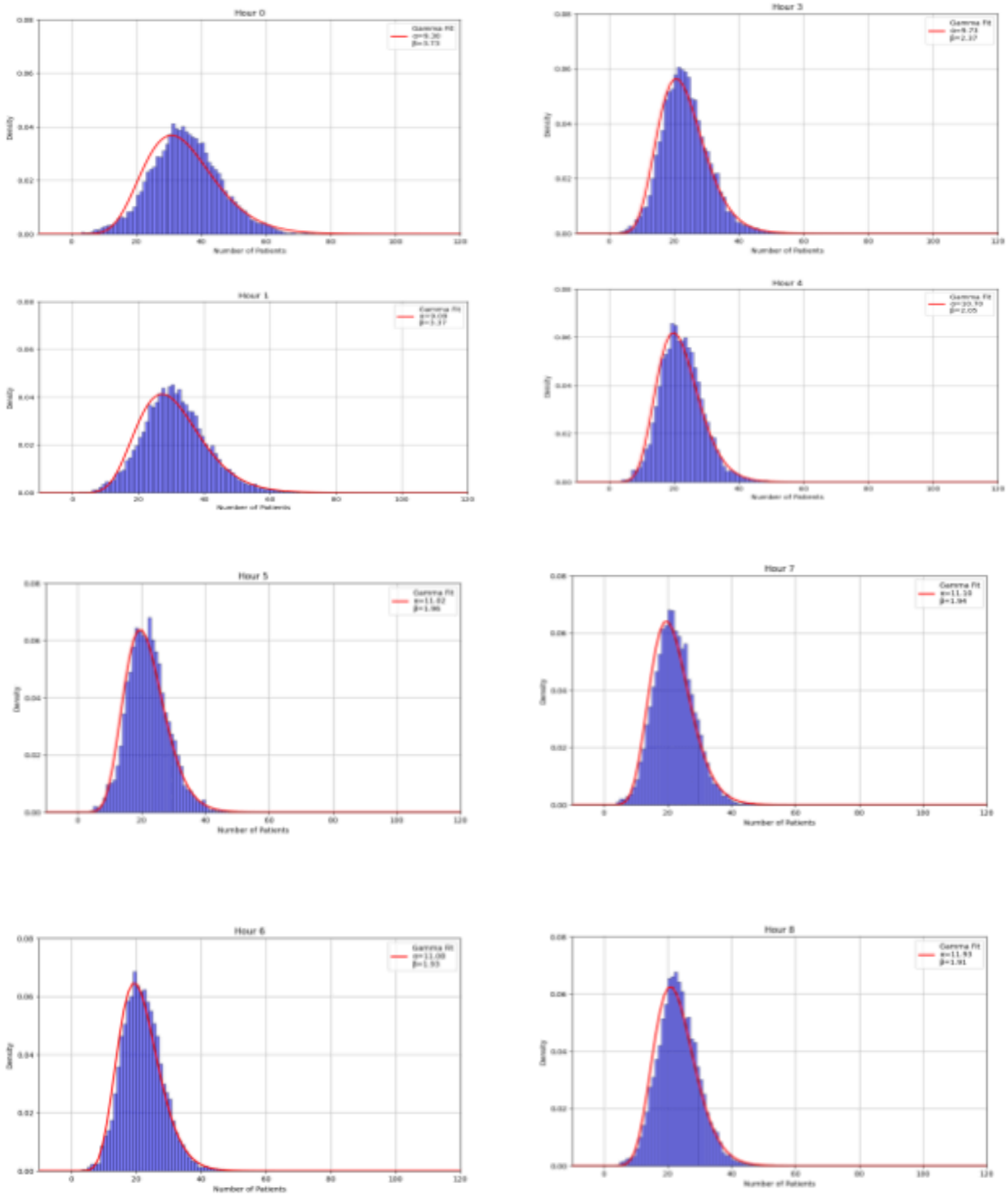
[23] T. Metsalu, J. Vilo, ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap, Nucleic Acids Res 43 (2015) W566–W570. https://api.semanticscholar.org/CorpusID:3848540.

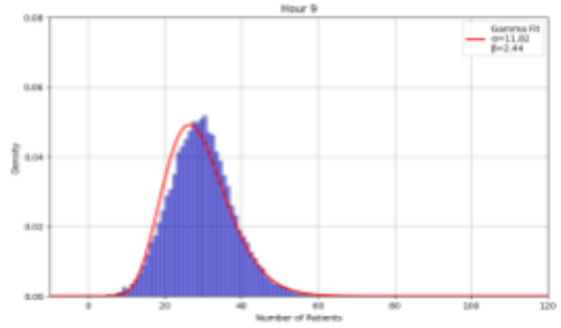
[24] Scikit-learn developers, DBSCAN clustering algorithm, Https://Scikit-Learn.Org/Stable/Auto_examples/Cluster/Plot_dbscan.Html (2023).

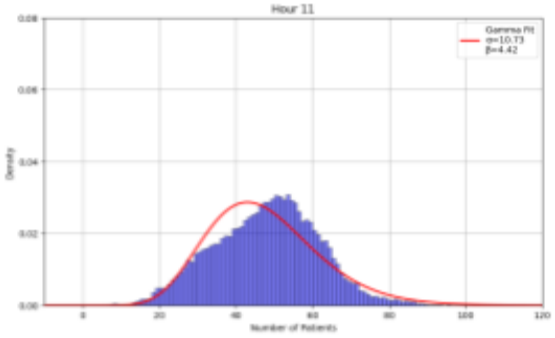
Appendix

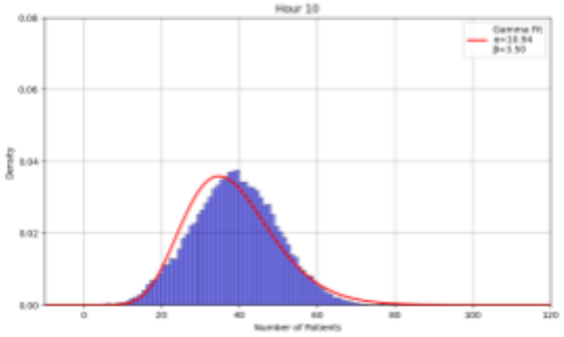
See Code: ***https://github.com/Faosiya/A-E-Data-Modelling/tree/main***
https://github.com/Faosiya/A-E-Data-Modelling/blob/main/A%26E_WAITING_MODELLING.ipynb

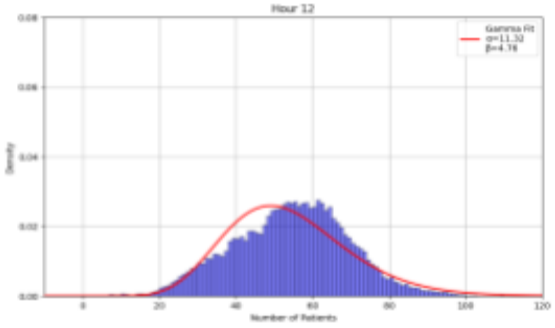
Distribution of the Number of Patients in the ED per Hour of the Day

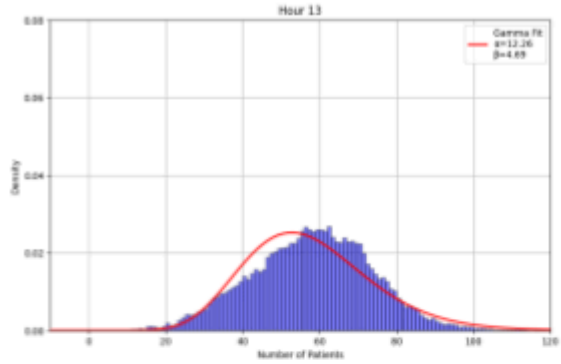


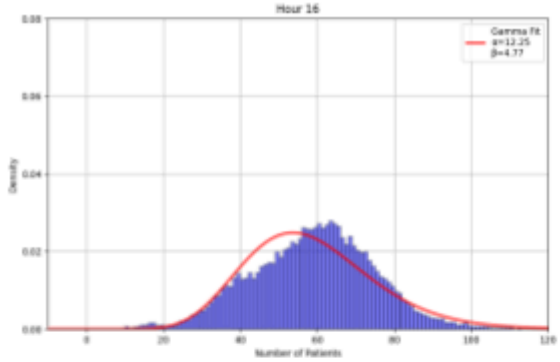

Hour 9
Gamma Fit
α=11.02
β=2.44
Density
Number of Patients

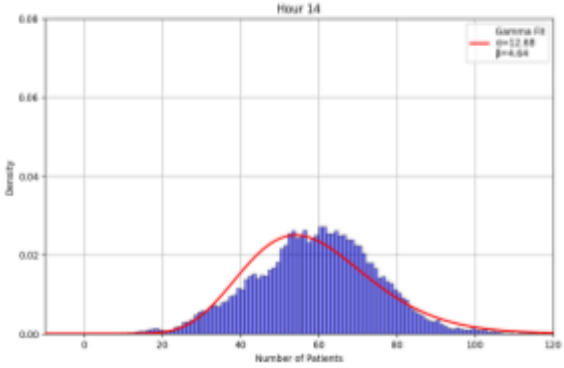

Hour 11
Gamma Fit
α=10.73
β=4.42
Density
Number of Patients

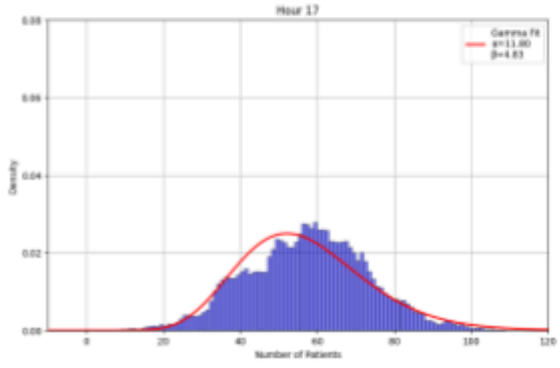

Hour 10
Gamma Fit
α=10.94
β=3.50
Density
Number of Patients

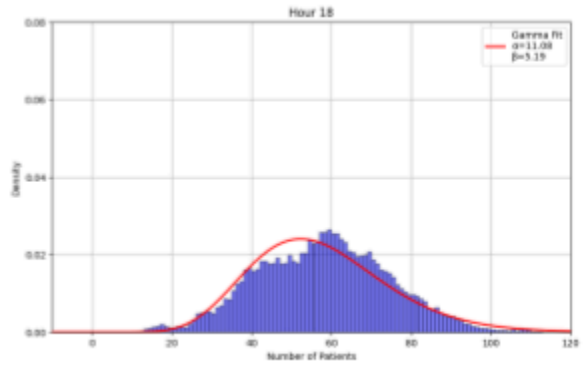

Hour 12
Gamma Fit
α=11.32
β=4.78
Density
Number of Patients

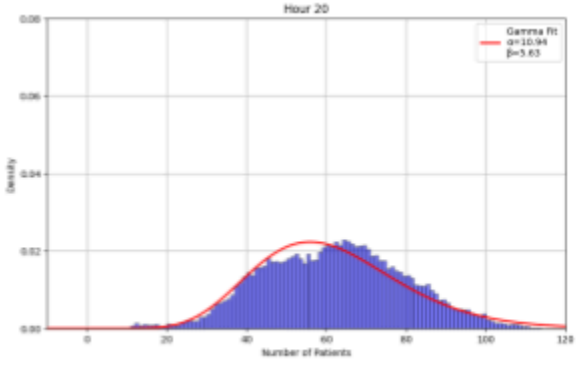

Hour 13
Gamma Fit
α=12.26
β=4.69
Density
Number of Patients

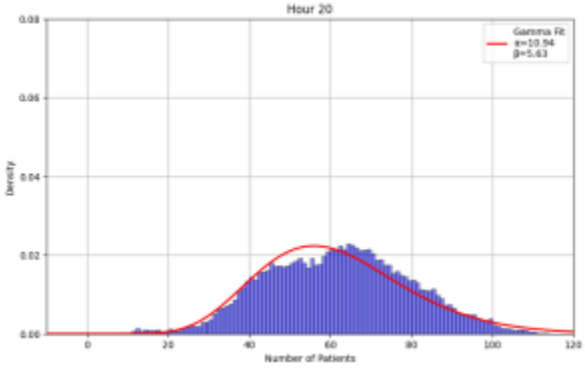

Hour 16
Gamma Fit
α=12.25
β=4.77
Density
Number of Patients

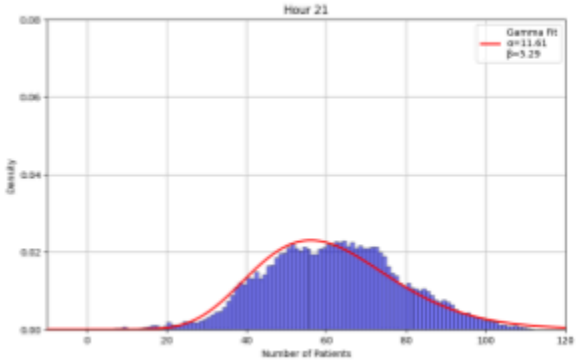

Hour 14
Gamma Fit
α=12.68
β=4.64
Density
Number of Patients

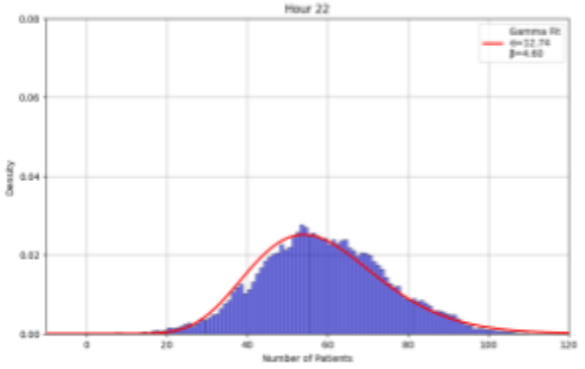

Hour 17
Gamma Fit
α=11.80
β=4.83
Density
Number of Patients

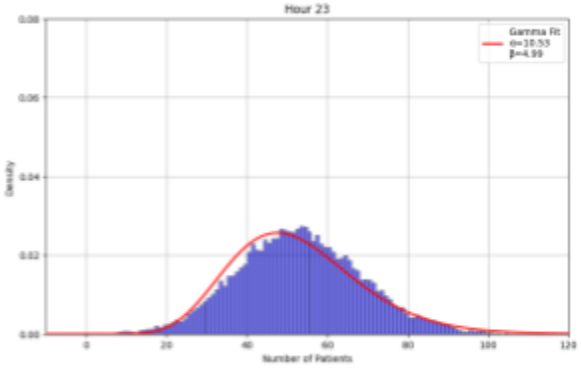

Hour 18
Gamma fit
α=11.08
β=5.19
Density
Number of Patients

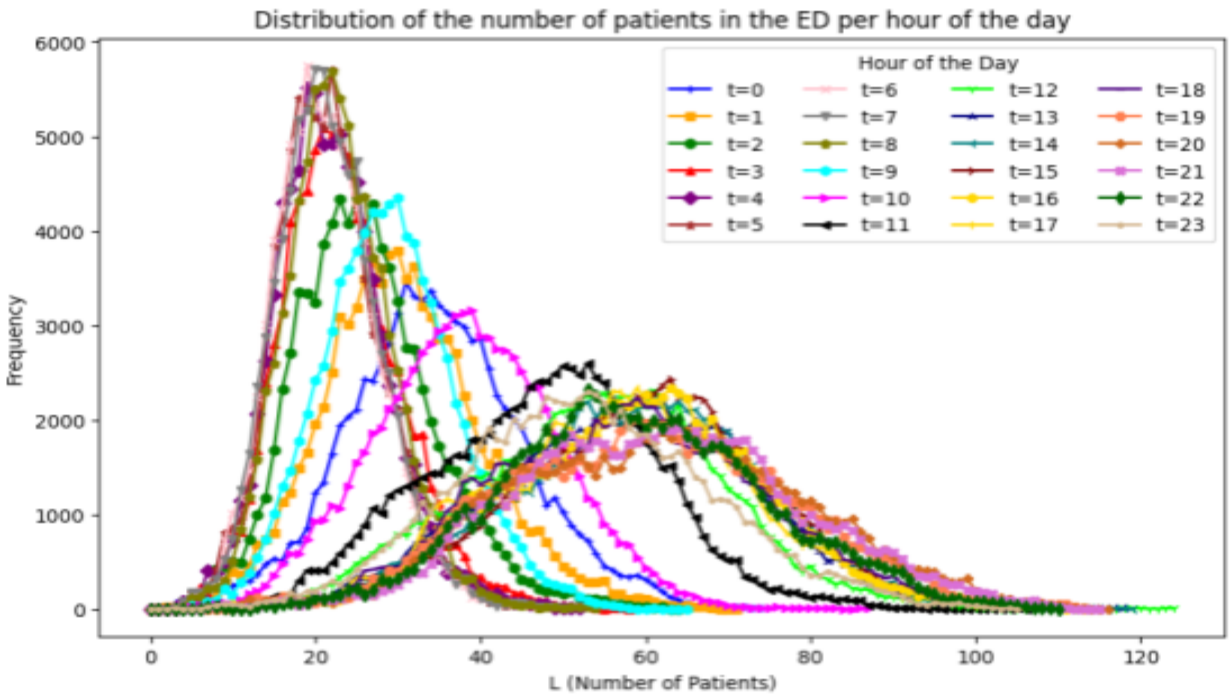

Hour 20
Gamma fit
α=10.94
β=5.63
Density
Number of Patients

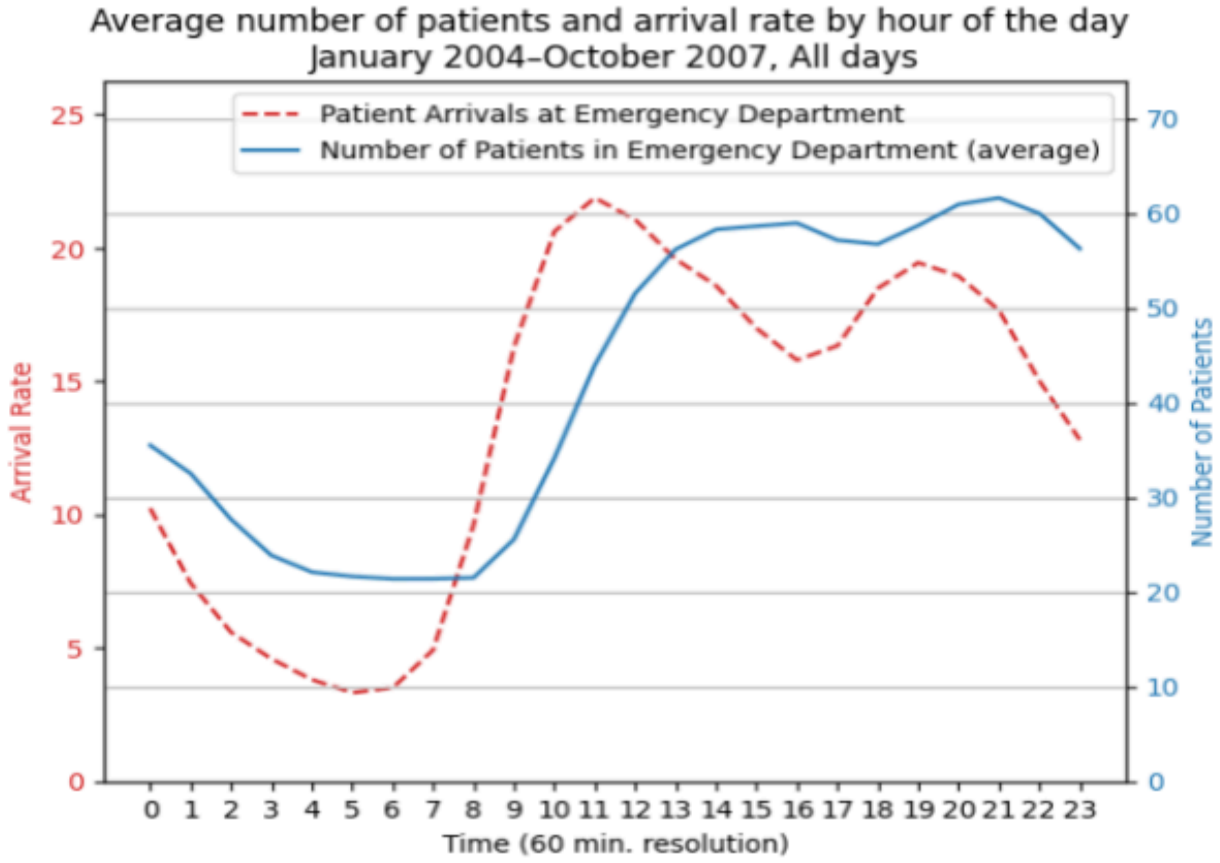

Hour 20
Gamma fit
α=10.94
β=5.63
Density
Number of Patients

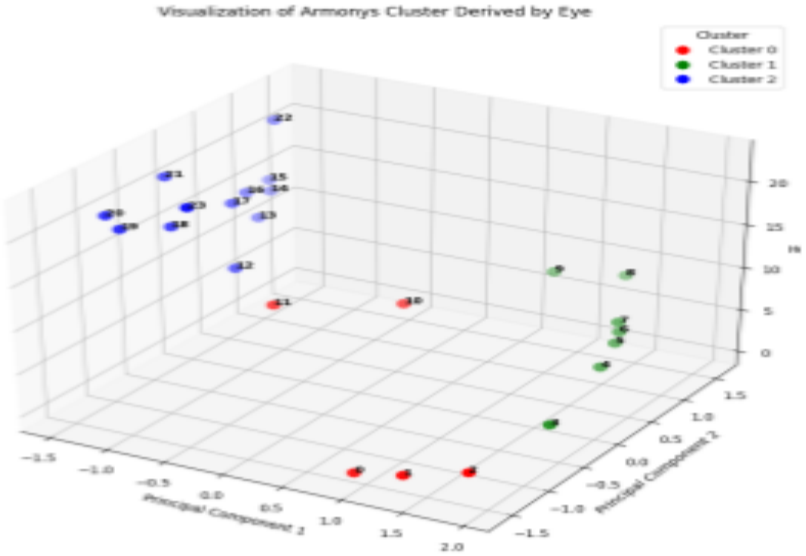

Hour 21
Gamma fit
α=11.61
β=5.29
Density
Number of Patients

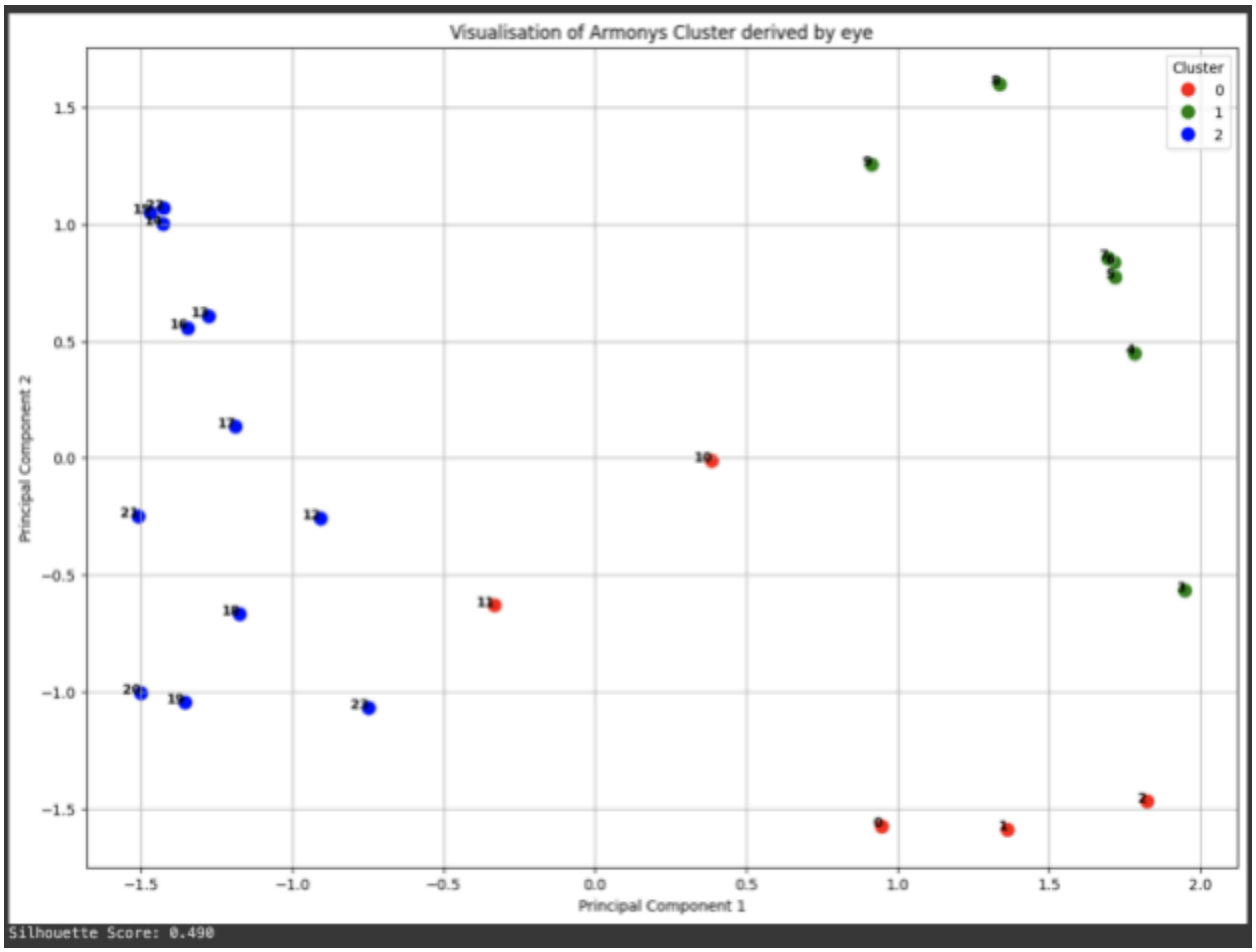

Hour 22
Gamma fit
α=12.74
β=4.60
Density
Number of Patients


Hour 23
Gamma fit
α=10.53
β=4.99
Density
Number of Patients


Distribution of the number of patients in the ED per hour of the day
Hour of the Day
t=0
t=1
t=2
t=3
t=4
t=5
t=6
t=7
t=8
t=9
t=10
t=11
t=12
t=13
t=14
t=15
t=16
t=17
t=18
t=19
t=20
t=21
t=22
t=23
Frequency
L (Number of Patients)


Average number of patients and arrival rate by hour of the day
January 2004–October 2007, All days
Patient Arrivals at Emergency Department
Number of Patients in Emergency Department (average)
Arrival Rate
Number of Patients
Time (60 min. resolution)


Visualization of Armonys Cluster Derived by Eye
Cluster
Cluster 0
Cluster 1
Cluster 2
Principal Component 1
Principal Component 2


Visualisation of Armonys Cluster derived by eye
Cluster
0
1
2
Principal Component 1
Principal Component 2
Silhouette Score: 0.490