Imperial College London

DEPARTMENT OF MATHEMATICS

# Modelling Santander Cycle Hires Using Self-Exciting and Mutually Exciting Point Processes

## M2R Second Year Project


| Students | CID |
|---|---|
| Carlos Cardoso Correia Perello | 01727546 |
| Mathieu Deschenes | 01731887 |
| Patrick Pham | 01850033 |
| Liyang Wang | 01872512 |
| Soumil Singla | 01855286 |

Supervisor

Dr. Francesco Sanna Passino

**Abstract**

In this report, we introduce the theory of Self-Exciting (SE) and Mutually Exciting (ME) point processes, which are powerful mathematical concepts with a wide range of applications across areas such a earthquake modelling and finance. We apply the theory to formulate 5 models describing Santander cycle hires from stations across London, with the model making use of both SE and ME behaviour demonstrating excellent performance. For each model, we derive a recursive form for the log-likelihood which allows for fast computation of optimal parameters via maximum likelihood estimation, using nonlinear optimisation and the Nelder-Mead algorithm. Finally, improvements to these models and other ways to measure their goodness of fit are discussed. Applications for the successful models include optimal bike allocation and re-balancing. Additionally, the theoretical results are general enough to be applicable to any system which might demonstrate SE or ME behaviour. The repository for the code written for this project is included in the Appendix.

***Keywords:*** point processes, Hawkes processes, self-exciting processes, mutually exciting processes.

# Contents

# 1 Introduction

Point processes are a universal class of random variables describing random occurrences of events through time. Many naturally occurring phenomena, such as arrival times at a queue or radioactive decay of an isotope, can be modelled by a class of point processes known as Poisson point processes (Pawitan, 2001). These satisfy the properties that the number of occurrences over a given time interval follows a Poisson distribution with mean proportional to the length of the time interval in which they are observed, and that the occurrence of one of these events does not affect the occurrence of subsequent events. Poisson point processes provide a very simple, yet powerful model that can be used for many applications before considering more sophisticated models.

Whilst Poisson point processes are modelled with a constant rate of occurrence per unit time (or *intensity*), some processes are better modelled by a rate of occurrence that changes over time, such as a queue that becomes more active during the afternoon. This discussion will give rise to the so-called *intensity function* of a point process, $\lambda^*(t)$. For some models, the intensity of the point process can also be increased, or *excited*, by past occurrences. Indeed, an earthquake at a particular location is likely to cause aftershocks which might affect that same location again (Ogata, 1988). This is an example of a so-called *self-exciting* processes, in which the probability of an event occurring is increased by previous events, leading to clustering behaviour. Self-exciting processes can also be observed in a wide range of other systems, from computer network traffic (Price-Williams, Heard, 2020) to stock mid-price changes over time (Laub, Taimre, Pollett, 2015). One type of self-exciting processes we are interested in is called *Hawkes Processes (HPs)*, in which the probability of an event occurring is influenced by *all* prior events. Multiple point processes can also be *Mutually Exciting (ME)* if the occurrence of an event at one of the processes excites the intensity function of other events. A common example of this is financial contagion (Laub, Taimre, Pollett, 2015).

The aim of this project is to study the theory of Poisson processes, Hawkes processes and mutually exciting processes, applying these concepts to model departure times of Santander cycles from bike stations in the London area, where each station is modelled as a point process with random departure times. The report is structured as follows: Section 2 introduces important concepts for our discussion, followed by an exploration of the dataset we are using in Section 3. In Section 4, we define the models we use to model the data. The procedure to fit these models and evaluate their goodness of fit is presented in Section 5, before being applied in Section 6. Finally, Section 7 discusses possible improvements to the models as well as their applications.

# 2 Theoretical Background

Before introducing the models we use on our data, it is important to familiarise oneself with some key mathematical concepts. In this section, we give a quick review of the Poisson and Exponential distributions, define counting processes, and introduce the main ideas we use for our models: Hawkes processes and mutually exciting processes.

## 2.1 Poisson and Exponential distributions

**Definition 1** (Poisson Distribution)**.** *A discrete random variable $X$ is said to follow a* Poisson($\lambda$) *distribution if its probability mass function is given by:*

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

*for $x \geq 0$ and 0 otherwise.*

**Definition 2** (Exponential Distribution)**.** *A continuous random variable $X$ is said to follow an* Exponential($\lambda$) *distribution if its cumulative density function is given by:*

$$P(X \leq x) = 1 - e^{-\lambda x}$$

*where $X$ has support $[0, \infty)$.*

## 2.2 Counting Processes and Point Processes

**Definition 3** (Point Process)**.** *A point process on the non-negative real axis (representing time) is a sequence of random variables $\{T_1, T_2, ...\}$ taking values on $[0, \infty)$ that satisfies $P(0 \leq T_1 \leq T_2 \leq ...) = 1$. That is, the $T_i$ are almost surely ordered. The $T_i$ are commonly referred to as the event times (Rizoiu et al., 2017).*

**Definition 4** (Counting Process)**.** *A counting process $N(t)$ is a stochastic process defined on time $t \geq 0$ and taking values in $\mathbb{N}_0$. Here, $N(t)$ represents the number of events in the interval $[0, t]$, and satisfies $N(0) = 0$. Therefore, $N(t)$ is uniquely determined by the corresponding event sequence, $T_i$.*

The same underlying process can therefore be characterised by either the event times $\{T_1, T_2, ...\}$ or the counting process $N(t)$. We can encode this relationship explicitly using the following representation of $N(t)$ (Rizoiu et al., 2017):

$$N(t) = \sum_{i \geq 1} \mathbb{1}_{T_i \leq t}$$

where $\mathbb{1}_{T_i \leq t}$ is the indicator variable of the event $T_i \leq t$.

## 2.3 Poisson Point Processes

**Definition 5** (Poisson Point Process)**.** *A Poisson point process is a point process in which the number of events occurring during any time interval with length $t$ follows a* Poisson($\lambda t$) *distribution, for some constant rate of occurrence $\lambda > 0$.*

Some characteristics of Poisson point processes that are key to our use of them are given in the lemmas below:

**Lemma 1** (Independence of Events)**.** *The number of events of a Poisson point process occurring in disjoint time intervals are independent (Pawitan, 2001).*

**Lemma 2** (Inter-arrival times follow an Exponential distribution)**.** *For a Poisson point process with arrival times* $\{t_1, t_2, t_3, ...\}$*, we define the inter-arrival times* $\{\tau_1, \tau_2, \tau_3, ...\} = \{t_1, t_2 - t_1, t_3 - t_2, ...\}$*, which are the time elapsed between consecutive events. One can show that for a Poisson point process with rate* $\lambda$*, the inter-arrival times are independent and all follow an* Exponential($\lambda$) *distribution (A).*

This useful property is later used to assess the goodness of fit of our models. We can estimate the rate $\lambda$ of a Poisson point process from data over an observation time $[0, T]$ by using the fact that for this process, $N(T)$ follows a Poisson($\lambda T$) distribution (Pawitan, 2001). We therefore have the formula:

$$\hat{\lambda} = \frac{N(T)}{T}$$

which is the usual unbiased estimator for $\lambda$ for a Poisson distribution.

## 2.4 Non-homogeneous Point Processes

As discussed above, some point processes are modelled better with a rate of occurrence $\lambda^*(t)$ which changes over time. Such processes are known as *non-homogenous* point process. This changing rate of occurrence is formally defined as the *conditional intensity function*:

**Definition 6** (Conditional Intensity Function)**.** *Consider a counting process* $N(t)$ *with associated 'history' of past events* $\mathcal{H}(t)$ *up to time* $t$*. If there exists a non-negative function* $\lambda^*(t)$ *such that*

$$\lambda^*(t) = \lim_{dt \to 0} \frac{E(N(t + dt) - N(t) \mid \mathcal{H}(t))}{dt}$$

*then* $\lambda^*(t)$ *is called the conditional intensity function of the point process (Laub, Taimre, Pollett, 2015).*

Note that the conditional intensity function describes the instantaneous rate at which events occur. Thus, assuming $N(t)$ is a *simple* point process (one where two events cannot occur at precisely the same time) we have that as $dt$ goes to 0, $N(t+dt) - N(t)$ approximately follows a Bernoulli($\lambda^*(t)dt$) distribution (Pawitan, 2001).

## 2.5 Hawkes Processes

One particularly useful type of non-homogenous point process is the so-called *Hawkes* process, which is a type of self-exciting process for which the intensity function depends on *all* the previous events. This statistical model has a wide range of applications, and offers a theory that is easy to work with.

**Definition 7** (Hawkes Process)**.** *A Hawkes process is a self-exciting point process with a conditional intensity function* $\lambda^*(t)$*, which depends not only on some (constant) background intensity* $\lambda > 0$ *but also on the previous event times,* $t_j$*.*

*We write this as*

$$\lambda^*(t) = \lambda + \sum_{t_j<t} \mu(t - t_j)$$

*Where $\mu(t)$ is the 'excitation' function, capturing how events excite the intensity function after they occur. Popular choices to model this function include the scaled exponential, $\mu(t) = \alpha e^{-\beta t}$ for $\alpha, \beta > 0$ (with explosion avoided when $\alpha < \beta$), and the power law function, $\mu(t) = \frac{k}{(c+t)^p}$ for $c, k, p > 0$ (Laub, Taimre, Pollett, 2015).*

The demonstrative figure below shows the intensity function over time for a Hawkes process with event times $t_1, t_2, ..., t_5$ and excitation function $\mu(t) = \alpha e^{-\beta t}$:

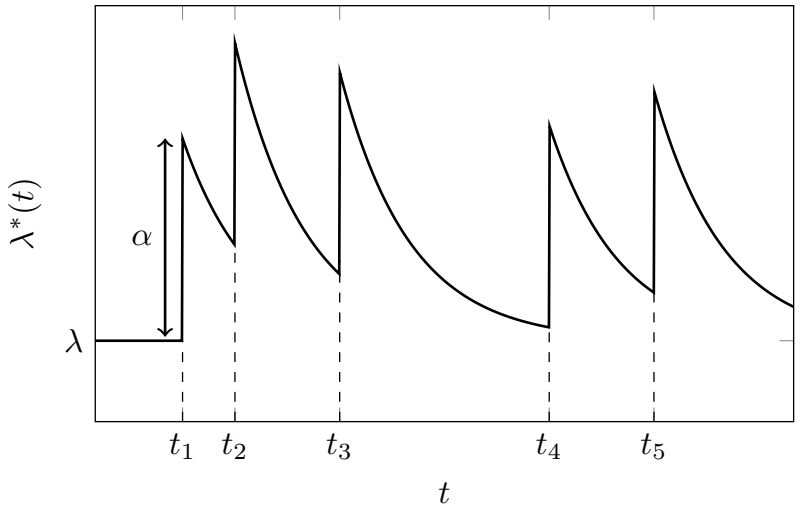


Figure 1: Cartoon of the intensity function of a Hawkes process over time.

We can see the intensity spikes instantaneously after each event, then decaying exponentially. The parameter $\lambda$ represents the constant background intensity, which is the height of the line from time $t = 0$ to $t = t_1$. The parameter $\alpha$ represents the jump in intensity after each event, whilst $\beta$ represents the rate at which the intensity decays after each event.

## 2.6 Mutually Exciting Processes

Another class of non-homogeneous point processes which we are interested in are *mutually exciting* processes. These are groups of point processes for which the intensity function of each process is excited by events occurring at other processes. Some models, such as the mutually exciting Hawkes process (Laub, Taimre, Pollett, 2015), have the property that events excite the intensity function of every process in the system. However, in this report, we only consider a model where events *start* at one station and *end* at another, with the intensity of a given process only being excited by events that end at it. Indeed, Santander cycle departures at a given station might be excited by departures from other stations, but only if those journeys end at the given station. Thus, the intensity for a process for such a model would be:

$$\lambda^*(t) = \lambda + \sum_{t'_k<t} \mu(t - t'_k)$$

where $t'_1, t'_2$... are the *end times* of events which terminate at the process, and $\mu(t)$ is the excitation function as mentioned above.

Figure 2 shows the intensity function for a mutually exciting point process with excitation function $\mu(t) = \alpha' e^{-\beta' t}$ with end times $t'_1$ and $t'_2$:

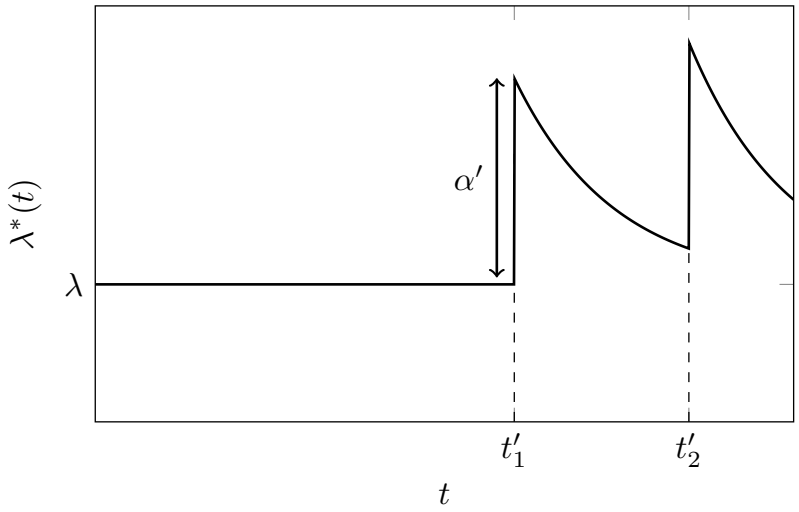


Figure 2: Cartoon of the intensity function of a mutually exciting process over time.

The intensity of this process spikes by $\alpha'$ after each event with end time $t'_j$. It then decays back exponentially to the background intensity according to the rate $\beta'$.

Lastly, Figure 3 shows the intensity function for a point process which is both self-exciting and mutually exciting, with intensity function $\lambda^*(t) = \lambda + \sum_{t_k<t} \alpha e^{-\beta(t-t_k)} + \sum_{t'_k<t} \alpha' e^{-\beta'(t-t'_k)}$:

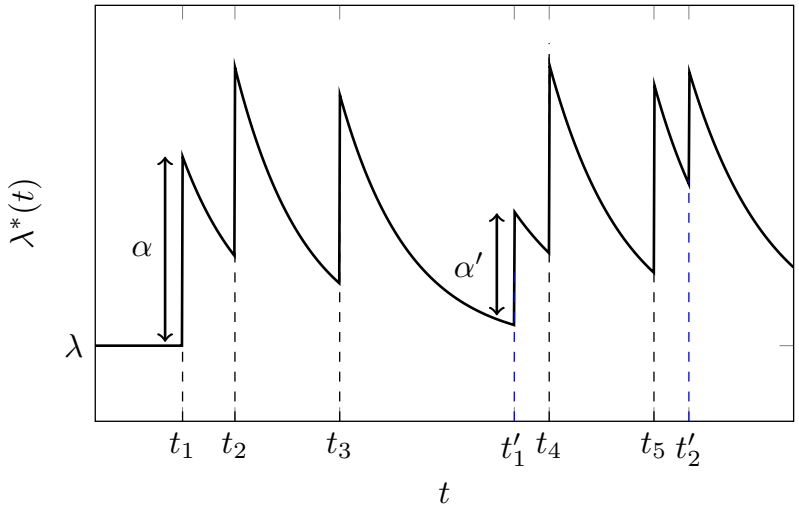


Figure 3: Cartoon of the intensity function of a self and mutually exciting process over time.

This model shows both self-exciting and mutually exciting behaviour: after every event time $t_j$ starting from the process, the intensity function spikes by $\alpha$ and decays back exponentially with rate $\beta$. Furthermore, after every event time $t'_k$ ending at the process, the intensity function spikes by $\alpha'$ and decays back exponentially with rate $\beta'$.

For all the above models, one can use Maximum Likelihood Estimation to compute the parameters that best fit the data given a choice of $\mu(t)$.

One function which is central to studying point processes is defined below:

**Definition 8** (Compensator Function)**.** *The compensator function $\Lambda(t)$ for a general point process with conditional intensity $\lambda^*(t)$ is defined as:*

$$\Lambda(t) \coloneqq \int_0^t \lambda^*(u)du$$

The fundamental role of this function in studying point processes is illustrated by the following elegant theorem. This theorem gives a universal method of assessing fit of point process models for which we have an estimated form of the conditional intensity function.

**Theorem 1** (Random Time Change Theorem)**.** *Suppose $\{t_1, t_2, ..., t_k\}$ is the realisation of a point process with conditional intensity function $\lambda^*(t)$ over an observation time $[0, T]$. If we also have that $\lambda^*(t)$ is positive over $[0, T]$ and $\Lambda(T) < \infty$ almost surely, then the transformed points $\{\Lambda(t_1), \Lambda(t_2), ..., \Lambda(t_k)\}$ form a Poisson process with unit rate (Laub, Taimre, Pollett, 2015).*

# 3 Exploring the data

The dataset we explored in this project describes journeys of Santander cycles between 842 different stations across the London area. In total, our dataset contains $3,961,103$ unique events.
Our aim is to model the departure times of these cycles at each station, which behave as random point processes with intensity depending on various factors.

Each row of our dataset represents one journey, and includes the start and end stations of the journey (represented by their station IDs), the start time, end times and journey durations (in seconds), and the distance (in Km) between the start and end stations. The first five entries are shown below:

| | start_id | end_id | start_time | duration | end_time | dist |
|---|---|---|---|---|---|---|
| 1 | 103 | 37 | 47260920 | 360 | 47261280 | 1.458333 |
| 2 | 39 | 539 | 47260920 | 120 | 47261040 | 0.545517 |
| 3 | 785 | 785 | 47260920 | 300 | 47261220 | 0.000000 |
| 4 | 341 | 159 | 47260980 | 1800 | 47262780 | 1.092775 |
| 5 | 708 | 573 | 47260980 | 1080 | 47262060 | 3.607290 |

Table 1: First 5 rows from the raw dataframe.

The start and end times are represented as the number of seconds since the 1st of January 2015. However, the *resolution* of our time data is only by the minute, therefore all time values are multiples of 60. The distance values were calculated using a separate dataset containing the location of each station, in degrees longitude and latitude. The first 5 rows of it can be found in Table 4 of the Appendix.

## 3.1 Preprocessing

To make the data easier to handle, our first step was to normalise *start_time* and *end_time*. We first defined the origin ($t_{min}$) as midnight on the first day of departure times in our dataset, then divided the time values by 60 to obtain times in minutes.

Another issue in the data was that due to times being rounded by the minute, some time values were equal. This is an issue as our analysis of point processes is in continuous time, and later methods require the time values to be ordered, which requires them to be unique. We therefore introduced small perturbations of $\epsilon_j$ minutes to all start and end time values $t_j$ where $\epsilon_j \sim \text{Uniform}(0, 1)$, giving our final times:

$$\tilde{t_j} = \frac{t_j - t_{min}}{60} + \epsilon_j$$

Note that the perturbations are within measurement error of the data as their purpose is only to make the times different for ordering. The resulting data was the following:

| | start_id | end_id | start_time | duration | end_time | dist |
|---|---|---|---|---|---|---|
| 1 | 103 | 37 | 2.260016 | 6.541469 | 8.801485 | 1.458333 |
| 2 | 39 | 539 | 2.897450 | 1.153272 | 4.050721 | 0.545517 |
| 3 | 785 | 785 | 2.975284 | 4.305295 | 7.280579 | 0.000000 |
| 4 | 341 | 159 | 3.017235 | 30.090668 | 33.107903 | 1.092775 |
| 5 | 708 | 573 | 3.141725 | 18.563206 | 21.704931 | 3.607290 |

Table 2: First 5 rows from the processed dataframe.

After preprocessing the data, we computed some summary statistics for the bike ride durations and distances, which are shown in the table below:

| Statistic | Duration (min) | Distance (Km) |
|---|---|---|
| Mean | 23.723106 | 2.279770 |
| Median | 16.299672 | 1.957944 |
| Mode | - | 0.000000 |
| IQR | 14.528259 | 2.060364 |
| Max | 9433.716466 | 16.748443 |
| Min | 0.009161 | 0.000000 |

Table 3: Table showing the summary statistics for bike ride duration and distance (the mode duration is not shown as all durations only occur once due to the perturbations).

It is worth noting that the mode and minimum distance travelled are both 0Km, as the distance is measured between stations and many bike rides start and end at the same station.

## 3.2 Clustering behaviour

One property present in the data that discourages the hypothesis that the stations behave as simple Poisson point processes is *clustering*. Indeed, cycle departure times are often seen clustered together in the data, such as is shown in the figure below, which shows the first 100 start times for a choice of 3 stations. The first station is in the 90th percentile of station popularity, making it very popular. The second has median popularity, and the last one is in the 10th percentile of station popularity, making it very unpopular. These stations were chosen to show the behaviour of departures from stations depending on their popularity.

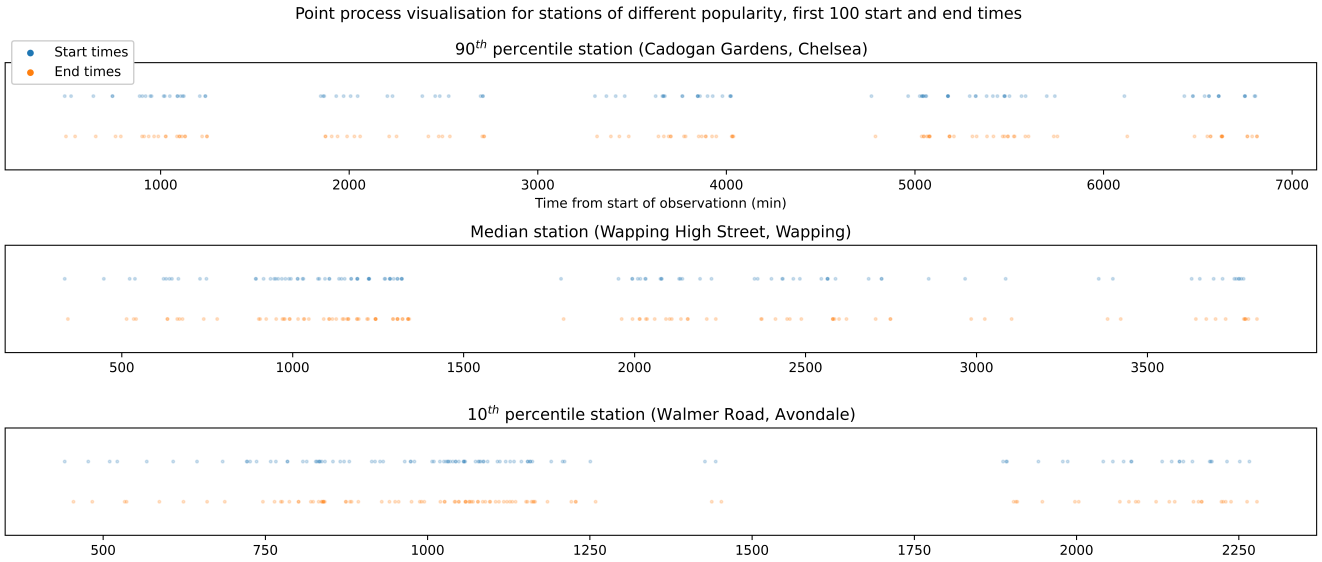


Figure 4: Visualisation of the first 100 start and end times as points in time for different stations.

One can see that the stations generally show clustering behaviour in the departure times, which is not expected under the Poisson process, in which departures would tend to be uniformly distributed over time. However, this behaviour is expected from self-exciting processes.

Another interesting property is that departures from these stations sometimes occur shortly after arrivals at the stations. This indicates that there could be some dependence between arrivals and departures at each station, which is expected under a mutually exciting model. These observations motivate the use of self and mutually exciting processes to model the station departures.

## 3.3 Daily Patterns

One more pattern that arises in the data which discourages the hypothesis that journey start times are Poisson point processes is the presence of daily patterns. Figure 5 below shows a histogram of the first 100,000 start times across the entire data, where the x-axis is in minutes.

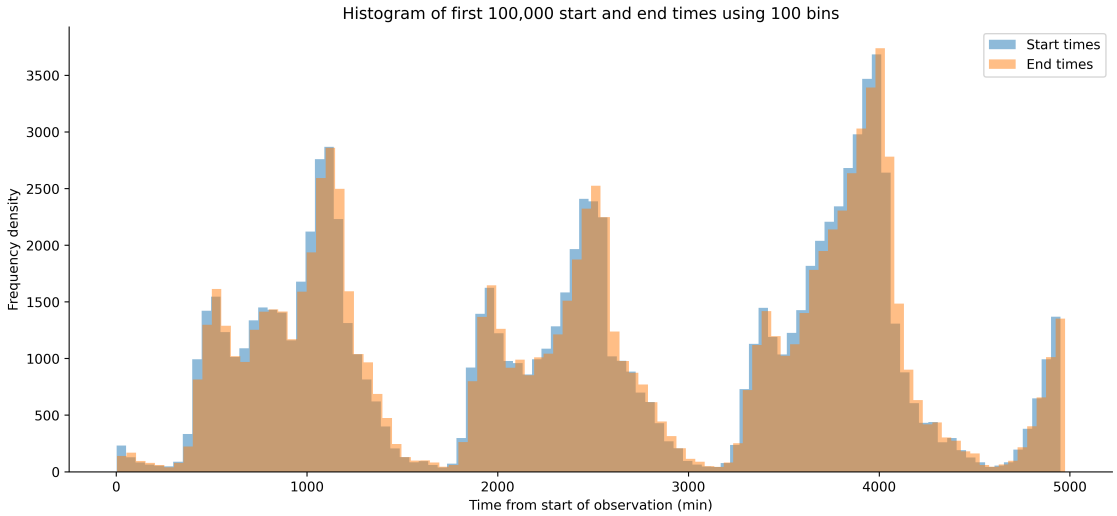


Figure 5: First $100,000$ start and end times visualised as a histogram; the peaks correspond to rush hour, and the valleys coincide with night time.

It is clear that there is a periodic pattern in the departures and arrivals, corresponding to daily patterns of bike usage. Indeed, rentals are expected to be more frequent during the day and drop at night. This is not taken into account in the Poisson model, in which there is a constant rate of occurrence across time. Whilst we do not include daily patterns in our models, we explore it in the Further Work section.

In the next section, we formulate the statistical models that are employed to model our dataset based on the ideas of *self-exciting* and *mutually exciting* point processes.

# 4 Models

In this section, we introduce the models that we fit to the data. Each of them aims to capture the behaviour of the intensity function of the departures from each station. Later models require more sophisticated statistical ideas, but are generally more accurate.

**Definition 9.** *For a station $i$, we define the counting processes $N_i(t)$ and $N_i'(t)$ as the number of departures and arrivals from the station over the time interval $[0, t]$, respectively.*

## 4.1 Simple Poisson Processes (Model 1)

The first model one uses to fit data is usually the simplest possible one to describe the situation. Consequently, our first model simply involves modelling the stations as independent Poisson processes with a fixed rate. We estimate this rate $\lambda_i > 0$ from the data for each station $i$ by computing the unbiased estimator for $\lambda_i$ as seen above:

$$\hat{\lambda_i} = \frac{N_i(T)}{T}$$

For this model, the compensator $\Lambda_i(t) \coloneqq \int_0^t \lambda_i^*(u)du$ is trivially given by:

$$\Lambda_i(t) = \lambda_i t$$

for all $t > 0$. Whilst this model is easy to implement, we will see in later sections that it fails to capture much of the behaviour of our system.

## 4.2 Hawkes Processes (Model 2)

As seen in the previous section, there is good reason to believe that in our system, departures from each station display self-exciting behaviour, or clustering. This is because people often choose to rent cycles as a group, which would result in clustered departure times.

This motivates the formulation of model 2, in which each station is taken to be an independent Hawkes process with conditional intensity function $\lambda_i^*(t)$. As stations do not influence each other under this model, we minimise the likelihood function of each station independently. For each station $i$ we choose our excitation function to be the scaled exponential $\mu_i(t) = \alpha_i e^{-\beta_i t}$ with $\lambda_i, \alpha_i, \beta_i > 0$ and $\alpha_i < \beta_i$. Thus, our intensity function $\lambda_i^*(t)$ given arrival times $t_{i,1}, t_{i,2}, ...$ at station $i$ becomes:

$$\lambda_i^*(t) = \lambda_i + \sum_{t_{i,j}<t} \alpha_i e^{-\beta_i(t-t_{i,j})}$$

The theorem below gives the form of the likelihood function for a general *one-dimensional* point process, which makes it applicable to each of our subsequent models. A one-dimensional point process is one in which there is only one type of event considered (in this case, departure times from stations). It is important to clarify that our models are all one-dimensional, due to the fact that they are only describing the intensity of *departure* times from each station. Indeed, even though Model 3 makes use of *arrival* times at each station, it is still a one-dimensional model, as it only models the intensity of departure times.

**Theorem 2** (Likelihood Function for One-Dimensional Point Processes)**.** *Let $N(t)$ be a regular one-dimensional point process. The likelihood function for a given estimate of the intensity function $\lambda^*(t)$ over an observation period $[0,T]$ is given by:*

$$L = \left[\prod_{j=1}^{N(T)} \lambda^*(t_j)\right] \exp\left(-\int_0^T \lambda^*(u)du\right)$$

*where $t_1, ..., t_{N(T)}$ are the event times observed over the observation time of the data, $T$ (Laub, Taimre, Pollett, 2015).*

Using this result, we can derive the log-likelihood $l_i$ for station $i$ given a conditional intensity function $\lambda_i^*(t)$ for each model by taking the logarithm of the above expression, yielding:

$$l_i = \sum_{j=1}^{N_i(T)} \log(\lambda_i^*(t_{i,j})) - \int_0^T \lambda_i^*(u)du = \sum_{j=1}^{N_i(T)} \log(\lambda_i^*(t_{i,j})) - \Lambda_i(T)$$

where $N_i(T)$ denotes the number of departures from station $i$ over the observation time of the data $[0,T]$, with the departure times $t_{i,1}, t_{i,2}, ..., t_{i,N_i(T)}$. We later compute the optimal parameter values for our data by maximising the log-likelihood for each model using computational methods.

For model 3, our next aim is to evaluate the compensator function $\Lambda_i(t) = \int_0^t \lambda_i^*(u)du$ for station $i$ for a general time $t > 0$. This will permit us to assess the fit of the model via the Random Time Change Theorem, which we discuss in a later section. We can achieve this by splitting the integral of the compensator over the intervals between departure times at station $i$, namely $[0, t_{i,1}], (t_{i,1}, t_{i,2}], ..., (t_{i,N_i(t)}, t]$. This ensures that the sum term inside $\lambda_i^*(t)$ has a fixed number of terms while we integrate over it. This yields:

$$\Lambda_i(t) = \int_0^t \lambda_i^*(u)du = \int_0^{t_{i,1}} \lambda_i^*(u)du + \sum_{j=1}^{N_i(t)} \int_{t_{i,j}}^{t_{i,j+1}} \lambda_i^*(u)du$$

$$= \int_0^{t_{i,1}} \lambda_i du + \alpha_i \sum_{j=1}^{N_i(t)} \int_{t_{i,j}}^{t_{i,j+1}} \left(\lambda_i + \sum_{k=1}^{j} e^{-\beta_i(u-t_{i,k})}\right) du$$

$$= \lambda_i t - \frac{\alpha_i}{\beta_i} \sum_{j=1}^{N_i(t)} \sum_{k=1}^{j} \left[e^{-\beta_i(t_{i,j+1}-t_{i,k})} - e^{-\beta_i(t_{i,j}-t_{i,k})}\right]$$

where we have defined $t_{i,N_i(t)+1} := t$. Note this formula still applies to values of $t$ such that $t_{i,N_i(t)} = t$. Most terms in this double summation cancel out, leaving:

$$\Lambda_i(t) = \lambda_i t - \frac{\alpha_i}{\beta_i} \sum_{j=1}^{N_i(t)} [e^{-\beta_i(t-t_j)} - 1]$$

for $t > 0$ (Laub, Taimre, Pollett, 2015). Using our expression for $\lambda_i^*(t)$ with unknown constants $\alpha_i$, $\beta_i$ and $\lambda_i$, the log-likelihood becomes (Laub, Taimre, Pollett, 2015):

$$l_i = \sum_{j=1}^{N_i(T)} \log\left(\lambda_i + \alpha_i \sum_{l=1}^{j-1} e^{-\beta_i(t_j-t_l)}\right) - \lambda_i T + \frac{\alpha_i}{\beta_i} \sum_{j=1}^{N_i(T)} [e^{-\beta_i(T-t_j)} - 1]$$

While this gives an expression for $l_i$ in closed form, it is computationally intensive to obtain as the double summation term makes this computation $O(N_i(T)^2)$. A solution to this is to re-write this expression in recursive form, re-using quantities previously computed to yield a method of computation that is $O(N_i(T))$. Indeed, defining $A_i(j)$ as follows (Laub, Taimre, Pollett, 2015):

$$A_i(j) \coloneqq \sum_{l=1}^{j-1} e^{-\beta_i(t_j - t_l)} = e^{-\beta_i(t_j - t_{j-1})}(1 + A_i(j-1))$$

for $j = 2, 3, ..., n_i$ and with $A_i(1) = 0$ gives us a recursive relationship for the $A_i(j)$ that can be utilised by seeing that

$$l_i = \sum_{j=1}^{N_i(T)} \log(\lambda_i + \alpha_i A_i(j)) - \lambda_i T + \frac{\alpha_i}{\beta_i} \sum_{j=1}^{N_i(T)} [e^{-\beta_i(T - t_j)} - 1]$$

By first computing each $A_i(j)$ (an $O(N_i(T))$ operation), and substituting into this equation, one has a method for computing $l_i$ which is $O(N_i(T))$. This will later be used to efficiently find parameters that maximise the likelihood for each station with the use of standard python libraries.

## 4.3 Mutually Exciting Processes (Model 3)

Another behaviour which we have good reason to believe influences the departure times from each station is *mutually exciting* behaviour. Indeed, when a rider finishes a journey at an empty station, that bike is likely to be picked up quickly by a nearby person taking advantage of the opportunity, thus creating a new departure. Additionally, some riders choose to check out a bike at a station before their journey has lasted more than 30 minutes before checking it out again to avoid a £2 fee on journeys lasting more than 30 minutes. These behaviours would make our stations mutually exciting point processes, as an arrival at a station would increase the probability that a departure occurs in the near-future.

Model 3 utilises this concept to model departure times at each station. For this model, the intensity function of station $i$ is excited by journeys ending at station $i$. For stations $i = 1, ..., m$, arrival times $t'_{i,1}, ..., t'_{i,N'_i(T)}$ at station $i$ and conditional intensity function $\lambda^*_i(t)$, our model becomes:

$$\lambda^*_i(t) = \lambda_i + \sum_{t'_{i,k} < t} \alpha_i e^{-\beta_i(t - t'_{i,k})}$$

where the excitation function is chosen to be a decaying exponential with $\lambda_i, \alpha_i, \beta_i > 0$ and $\alpha_i < \beta_i$. Note that this model does not display self-exciting behaviour, which is included in model 4.

To fit this model to data, we again want to estimate the parameters $\lambda_i$, $\alpha_i$, and $\beta_i$ for each station $i$ via likelihood maximisation.

The log-likelihood for model 3 is the same as derived above:

$$l_i = \sum_{j=1}^{N_i(T)} \log(\lambda^*_i(t_{i,j})) - \Lambda_i(T)$$

where the event times for each station $i$ are $t_{i,1}, ..., t_{i,N_i(T)}$ and $\lambda_i^*(t)$ is given as above, with unknown constants $\lambda_i$, $\alpha_i$ and $\beta_i$.

Due to the sum inside $\lambda_i^*(t)$, this computation is still $O(N_i(T)^2)$. We therefore use a similar technique to the previous model to compute this efficiently. Indeed, if we define for an integer $h \geq 1$:

$$B_i(h) = \sum_{k=1}^{N_i'(t_{i,h})} \exp\left\{-\beta_i(t_{i,h} - t'_{i,k})\right\}$$

We then use the following argument to derive a recursive form for $B_i(h)$:

$$\begin{aligned}\exp\left\{-\beta_i(t_{i,h} - t_{i,h-1})\right\}(B_i(h-1)) &= \sum_{k=1}^{N_i'(t_{i,h-1})} \exp\left\{-\beta_i(t_{i,h} - t_{i,h-1} + t_{i,h-1} - t'_{i,k})\right\} \\ &= B_i(h) - \sum_{k=N_i'(t_{i,h-1})+1}^{N_i'(t_{i,h})} \exp\left\{-\beta_i(t_{i,h} - t'_{i,k})\right\}\end{aligned}$$

From this, we obtain the recursive relation:

$$B_i(h) = \exp\left\{-\beta_i(t_{i,h} - t_{i,h-1})\right\}(B_i(h-1)) + \sum_{k=N_i'(t_{i,h-1})+1}^{N_i'(t_{i,h})} \exp\left\{-\beta_i(t_{i,h} - t'_{i,k})\right\}$$

This recursive relation allows us to compute the log-likelihood in linear time, as the sum term inside has $O(1)$ terms. The base case $B_i(1)$ is simply:

$$B_i(1) = \sum_{k=1}^{N_i'(t_{i,1})} \exp\left\{-\beta_i(t_{i,1} - t'_{i,k})\right\}$$

which in our data can be computed in $O(1)$ time as there are normally very few arrival times at station $i$ smaller than or equal to $t_{i,1}$.

We then have that the log-likelihood for station $i$ has the form:

$$l_i = \sum_{j=1}^{N_i(T)} \log(\lambda_i + \alpha_i B_i(j)) - \Lambda_i(T)$$

To derive $\Lambda_i(t)$ for an arbitrary $t > 0$, we use a similar technique to model 2, integrating the intensity function $\lambda_i^*(t)$ over $[0, t]$ by instead splitting the interval into $[0, t'_{i,1}], (t'_{i,1}, t'_{i,2}], ..., (t'_{i,N_i(t)}, t]$ to ensure the number of terms in the sum inside $\lambda_i^*(t)$ remains constant when we integrate over it. Similar cancellation as before yields:

$$\Lambda_i(t) = \lambda_i t - \frac{\alpha_i}{\beta_i} \sum_{k=1}^{N_i'(t)} \left[e^{-\beta_i(t - t'_{i,k})} - 1\right]$$

## 4.4 Self-Exciting and Mutually Exciting Processes (Model 4)

For our fourth model, the intensity function of departure times from station $i$ is excited by arrivals to this station as well as departures from it. The conditional intensity function is thus a combination of the previous two models:

$$\lambda_i^*(t) = \lambda_i + \sum_{t'_{i,k}<t} \alpha'_i e^{-\beta'_i(t-t'_{i,k})} + \sum_{t_{i,k}<t} \alpha_i e^{-\beta_i(t-t_{i,k})}$$

where we have $\lambda_i, \alpha_i, \beta_i, \alpha'_i, \beta'_i > 0$, $\alpha_i < \beta_i$ and $\alpha'_i < \beta'_i$. One can see that the intensity function has both a self-exciting and mutually exciting summation term.

We again have that the log-likelihood is given by

$$l_i = \sum_{j=1}^{N_i(T)} \log(\lambda_i^*(t_{i,j})) - \Lambda_i(T)$$

To obtain a recursive formula for this log-likelihood, we simply combine the formulas derived in previous models, defining:

$$A_i(j) = \sum_{l=1}^{j-1} e^{-\beta_i(t_{i,j}-t_{i,l})}$$

$$B_i(h) = \sum_{k=1}^{N'_i(t_{i,h})} \exp\left\{-\beta'_i(t_{i,h} - t'_{i,k})\right\}$$

We then use the recursive formulas:

$$B_i(h) = \exp\left\{-\beta'_i(t_{i,h} - t_{i,h-1})\right\}(B_i(h-1)) + \sum_{k=N'_i(t_{i,h-1})+1}^{N'_i(t_{i,h})} \exp\left\{-\beta'_i(t_{i,h} - t'_{i,k})\right\}$$

$$A_i(j) = e^{-\beta_i(t_{i,j}-t_{i,j-1})}(1 + A_i(j-1))$$

Now, our log-likelihood can be re-written in terms of these formulas by seeing that:

$$l_i = \sum_{j=1}^{N_i(T)} \log(\lambda_i + \alpha_i A_i(j) + \alpha'_i B_i(j)) - \Lambda_i(T)$$

Additionally, by combining the results for the compensators of models 2 and 3, we obtain the formula for $\Lambda_i(t)$ for any $t > 0$:

$$\Lambda_i(t) = \lambda_i t - \frac{\alpha'_i}{\beta'_i} \sum_{k=1}^{N'_i(t)} \left[e^{-\beta'_i(t-t'_{i,k})} - 1\right] - \frac{\alpha_i}{\beta_i} \sum_{j=1}^{N_i(t)} \left[e^{-\beta_i(t-t_{i,j})} - 1\right]$$

## 4.5 Mutually Exciting Processes Based on Journey Duration (Model 5)

One facet of the data we have not yet utilised in our models is the duration time of journeys. There is reason to believe this would affect the behaviour of departure times, due mainly to two behaviours: Firstly, riders sometimes check back in faulty bikes to take a new one. Secondly, as mentioned in the section for model 3, riders often return bikes before their journeys reach 30 minutes, taking a new one immediately, to avoid paying a £2 fee for journeys over 30 minutes. These two behaviours are modelled in model 5, in which the intensity function of each station is excited by arrivals at that station, and particularly so when the journey durations are around 1 minute and 30 minutes. This choice is because the minimum journey duration recording in our data is 1 minute. Model 5 thus has the following intensity function:

$$\begin{aligned}\lambda_i^*(t) = \lambda_i &+ \sum_{t'_{i,k}<t} \alpha_i \exp\left\{-\delta_i(d'_{i,k}-1)\right\} \exp\left\{-\beta_i(t-t'_{i,k})\right\} \\ &+ \sum_{t'_{i,k}<t} \alpha'_i \exp\left\{-\delta'_i|d'_{i,k}-30|\right\} \exp\left\{-\beta'_i(t-t'_{i,k})\right\}\end{aligned}$$

where $d'_{i,k}$ is the duration time of the journey corresponding to the arrival time $t'_{i,k}$, and we have $0 < \lambda_i, \alpha_i, \delta_i, \beta_i, \alpha'_i, \delta'_i, \beta'_i$, $\alpha_i < \beta_i$ and $\alpha'_i < \beta'_i$. As mentioned above, $d'_{i,k} \geq 1$ in the data, which justifies the expression used. In this model, stations are now mutually exciting point processes with excitation of the departure times spiking immediately after returning a bike and around the 30-minute mark, which models the behaviour described above.

The compensator for this model is given by:

$$\Lambda_i(t) = \lambda_i t - \frac{\alpha_i}{\beta_i} \sum_{k=1}^{N'_i(t)} \left[e^{-\delta_i(d'_{i,k}-1)} e^{-\beta_i(t-t'_{i,k})} - 1\right] - \frac{\alpha'_i}{\beta'_i} \sum_{k=1}^{N'_i(t)} \left[e^{-\delta'_i|d'_{i,k}-30|} e^{-\beta'_i(t-t'_{i,k})} - 1\right]$$

via integration of the intensity function similarly to the method described in model 3, simply treating $(d'_{i,k}-1)$ and $|d'_{i,k}-30|$ as constants in the integration, since these do not change inside the intervals $[0, t'_{i,1}], (t'_{i,1}, t'_{i,2}], ..., (t'_{i,N_i(t)}, T]$ over which we integrate the intensity function. This holds as the $(d'_{i,k}-1)$ and $|d'_{i,k}-30|$ terms are constant in between any two arrival times $t'_{i,j}$ and $t'_{i,j+1}$.

The log-likelihood for this model is again:

$$l_i = \sum_{j=1}^{N_i(T)} \log(\lambda_i^*(t_{i,j})) - \Lambda_i(T)$$

If we define $C_i(h)$ and $D_i(h)$ in the following manner:

$$C_i(h) = \sum_{k=1}^{N'_i(t_{i,h})} \exp\left\{-\delta_i(d'_{i,k}-1)\right\} \exp\left\{-\beta_i(t_{i,h}-t'_{i,k})\right\}$$

$$D_i(h) = \sum_{k=1}^{N'_i(t_{i,h})} \exp\left\{-\delta'_i|d'_{i,k}-30|\right\} \exp\left\{-\beta'_i(t_{i,h}-t'_{i,k})\right\}$$

we have the recursive relations:

$$C_i(h) = \exp\left\{-\beta_i(t_{i,h} - t_{i,h-1})\right\} C_i(h-1) + \sum_{k=N_i'(t_{i,h-1})+1}^{N_i'(t_{i,h})} \exp\left\{-\delta_i(d_{i,k}' - 1)\right\} \exp\left\{-\beta_i(t_{i,h} - t_{i,k}')\right\}$$

$$D_i(h) = \exp\left\{-\beta_i'(t_{i,h} - t_{i,h-1})\right\} D_i(h-1) + \sum_{k=N_i'(t_{i,h-1})+1}^{N_i'(t_{i,h})} \exp\left\{-\delta_i'|d_{i,k}' - 30|\right\} \exp\left\{-\beta_i'(t_{i,h} - t_{i,k}')\right\}$$

allowing for computation of our log-likelihood in linear time, which can now be written as:

$$l_i = \sum_{j=1}^{N_i(T)} \log\left(\lambda_i + \alpha_i C_i(j) + \gamma_i D_i(j)\right) - \Lambda_i(T)$$

allowing us to use the same procedure as before to compute the log-likelihood in linear time.

# 5 Goodness of fit

In this section, we describe methods to compute optimal parameters for our data and to assess the fit of each model by using the so-called simple p-value test and the Kolmogorov-Smirnov test statistic.

## 5.1 Computing Optimal Parameters

To compute the optimal parameters for models 2 to 5, we use maximum likelihood estimation. As we have methods to compute the log-likelihood for each model efficiently, we can find parameters that (locally) maximise this function with a numerical optimisation method called the Nelder-Mead algorithm from the Python optimisation package SciPy. This method was chosen as it has demonstrated excellent performance when applied to the modelling of point processes (Tench, Fry, Gill, 2016). The Nelder-Mead method allows for unconstrained multi-dimensional optimisation of a function without knowledge of its derivative, which makes it applicable to our models. It is widely used across many of the sciences due to its ease of use and simplicity (Scholarpedia, 2009).

The Nelder-Mead algorithm works by first specifying an initial guess for our parameters, which is then used to optimise the function by using the geometric properties of simplices, as outlined in (Scholarpedia, 2009). As we require bounds on our parameters, it is essential for the algorithm to stay within these parameter bounds when optimising. One method to constrain the solution is to transform the parameter values to enforce the bounds implicitly. For example, for model 2 we require $\lambda_i > 0$ and $0 < \alpha_i < \beta_i$, which can be encoded using the transformation:

$$\lambda_{trans} = e^{\lambda_i},\ \alpha_{trans} = e^{\alpha_i},\ \beta_{trans} = e^{\alpha_i} + e^{\beta_i}$$

We then let the algorithm compute the values of $\tilde{\lambda}$, $\tilde{\alpha}$ and $\tilde{\beta}$ that maximise the likelihood function $L(e^{\tilde{\lambda}}, e^{\tilde{\alpha}}, e^{\tilde{\alpha}} + e^{\tilde{\beta}})$. Thus, the parameters that maximise the likelihood function $L(\lambda_{true}, \alpha_{true}, \beta_{true})$ are $\{e^{\tilde{\lambda}}, , e^{\tilde{\alpha}}, e^{\tilde{\alpha}} + e^{\tilde{\beta}}\}$ and for this solution we automatically have that $\lambda_{true}, \alpha_{true}, \beta_{true} > 0$ and $\alpha_{true} < \beta_{true}$.

## 5.2 Simple p-value Test

We now give a method of evaluating goodness of fit of our models. The process outlined below describes a method to evaluate the fit of a Poisson point process, which is the formulation of our first model. However, the Random Time Change Theorem (1) provides an elegant method for assessing fit of each of our subsequent models using this same procedure. Indeed, if we first transform the departure times $t_{i,1}, ..., t_{i,N_i(T)}$ from station $i$ by the *estimated* compensator for station $i$, $\Lambda_i(t)$ (using our computed parameters that maximise the likelihood), the transformed times $\Lambda_i(t_{i,1}), ..., \Lambda_i(t_{i,N_i(T)})$ form a Poisson point process with *unit* rate under the assumption that our model is correct. Thus, the same test as below can be used on the transformed times with $\hat{\lambda} = 1$ to assess fit of each of our models.

**Evaluating fit of a Poisson point process to data**
As seen in the preliminaries, the *inter-arrival times* $\{\tau_1, \tau_2, \tau_3, ...\} = \{t_1, t_2 - t_1, t_3 - t_2, ...\}$ for a Poisson process with rate $\lambda$ are independent and each follow an Exponential($\lambda$) distribution. Thus, given an estimate $\hat{\lambda}$ for $\lambda$, we can calculate the approximate p-values for each inter-arrival time $\tau_i$ in our data by $p_i = P(X > \tau_i)$ for $X \sim$ Exponential($\hat{\lambda}$). Note that the p-values are the

probability that the observations would be at least as extreme as observed. Under the assumption that our model is correct, i.e. that the point process is Poisson with rate $\hat{\lambda}$, the $\tau_i$ are independent observations of an Exponential($\hat{\lambda}$) distribution. Thus, the calculated p-values should follow a Uniform$(0, 1)$ distribution. We can thus assess fit of the Poisson model for a point process by comparing the obtained p-values to a Uniform$(0, 1)$ distribution, either via a QQ-plot or by computing the Kolmogorov-Smirnov test statistic, which we introduce below.

## 5.3 Kolmogorov-Smirnov Test Statistic

The Kolmogorov-Smirnov (KS) test statistic is a **non-parametric test statistic** used to quantify the differences in the shape of a sample empirical cumulative distribution function, $F_S(x)$, to a known theoretical c.d.f., $F_T(x)$ (Clapham, 2016). The null hypothesis of this test is

$$H_0 = \text{data follows a known distribution } T$$

and the Kolmogorov-Smirnov test statistic is defined to be the greatest vertical distance between the two cumulative distribution functions. In other words:

$$\text{KS} \coloneqq \sup_{x \in [0,1]} |F_S(x) - F_T(x)|$$

Indeed, if the null hypothesis were true, the value of the KS-statistic would be zero. Thus a lower value for this statistic indicates data that better fits the theoretical (expected) distribution T.

In the context of our models, we are interested to find out whether the p-values for each station follow a Uniform$(0, 1)$ distribution. After calculating the p-values for each station and obtaining its empirical c.d.f., we can compare it to the c.d.f. of a Uniform$(0, 1)$ distribution to calculate the KS statistic for each station, for each model. Thus models with lower median KS statistics for each station can be said to fit the data better.

## 5.4 Train-test split

For each of our models, we will separate the dataset into train data and test data. This assures that we test the fit of our model on data it has never seen before, preventing over-fitting. The split we have chosen is a $12/4$ split of our data which is composed of 16 weeks. That is, we take the first 12 weeks of data to train the models and test the fit on the last 4 weeks. The test set was made up of $3,164,821$ events and 789 unique starting stations, whilst the test set captured $796,282$ events and the same number of unique starting stations.

# 6 Results

In this section, we outline the results of the above described tests to our data. The main results come from the box plots (Figures 8 and 9), which concisely show the relative performance of the models. The QQ and scatter plots show interesting properties of the data, such as where each model goes wrong. Lastly, in Figure 10 and 11, the $KS$ scores for each model are plotted against the popularity of each station, showing how the popularity of a station influences the fit of our models to it.

## 6.1 Quantile Analysis

The figures below shows QQ-Plots of the p-values of the train and test sets for each model, comparing the theoretical quantiles of a Uniform$(0, 1)$ distribution against the observed (empirical) quantiles of the calculated p-values. Each grey line represents the empirical quantiles of one station, whilst the blue line represents the *average* of all stations. The closer these lines are to the line $y = x$ (in black), the better the fit is for the model.

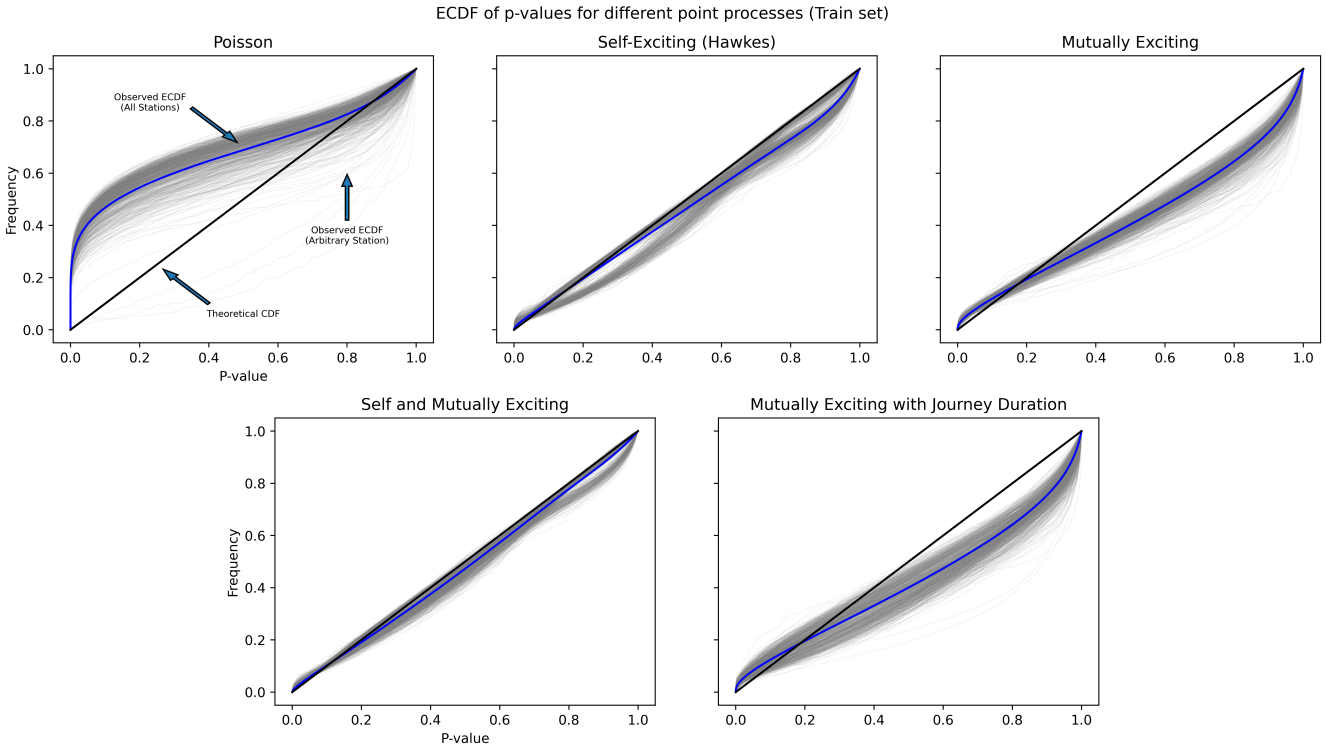


Figure 6: QQ-Plots of the p-values of the train set compared to a uniform distribution.

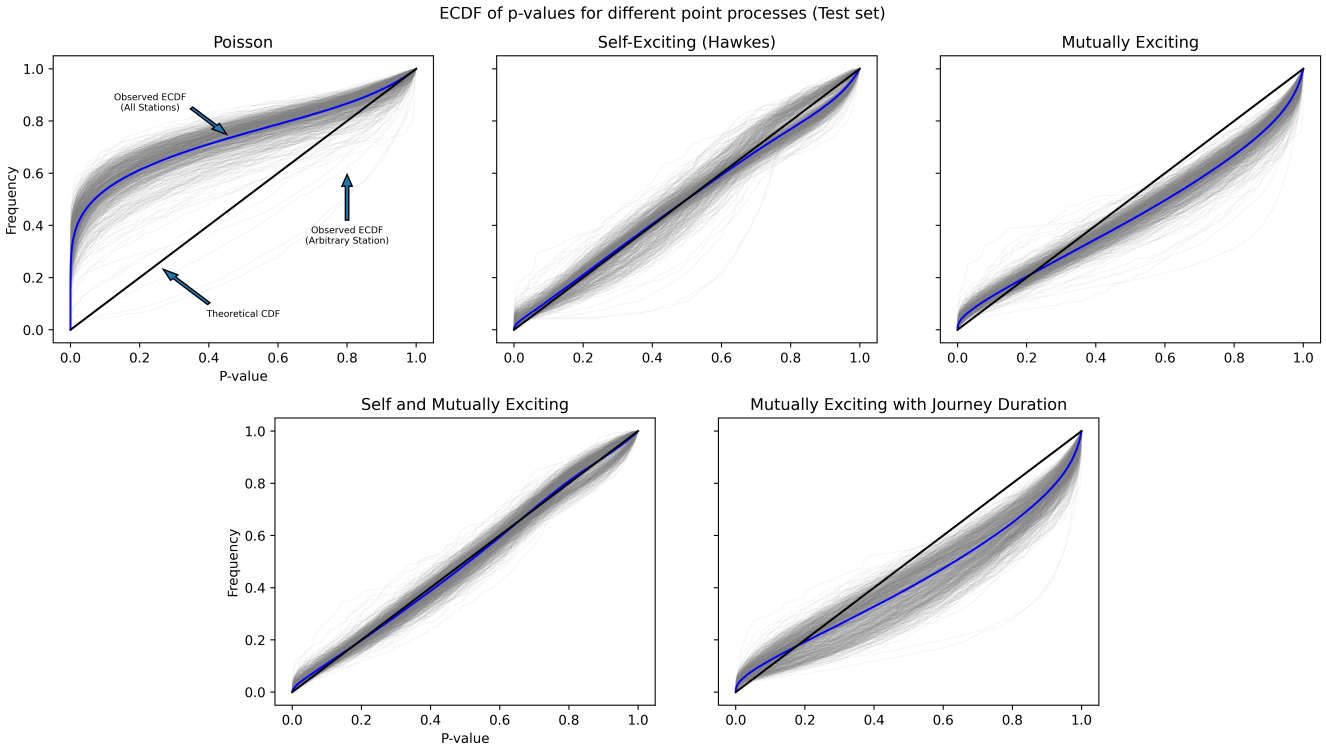


Figure 7: QQ-Plots of the p-values of the test set compared to a uniform distribution.

For model 1, the empirical quantiles deviate significantly from the theoretical ones for most stations, strongly indicating that the computed p-values do not follow a uniform distribution and thus that the model fit is poor. The curves tend to be over the line $y = x$, indicating that many p-values are smaller than expected, which is caused by inter-arrival times longer than predicted by the simple Poisson process. This could be caused by the lack of cycles being rented at night.

Models 2 and 3 generally show better behaviour, as the quantile lines tend to match the theoretical line more closely. The lines are slightly below the line $y = x$, indicating that there might be a heavy tail towards larger p-values. This indicates some shorter inter-arrival times than expected, likely due to the clustering behaviour that each model ignores (either SE or ME). The lines for model 4 fit much more closely the theoretical line, with an even distribution around it. This likely indicates that the model does not ignore any important clustering behaviour as it has both an SE and an ME component. The lines for model 5 seem similar in shape to model 3, but with a larger variance, indicating that the model fit is poorer.

Generally, the fit for the training set and the test set are similar, which indicates that the models do not suffer from over-fitting. Surprisingly, model 4 performs better in the test set than the training set. Its fit is excellent, as the blue line very closely matches the black line, with grey lines tightly spread around the black line in a symmetric fashion, which could also indicate that the model is *unbiased*.

## 6.2 KS Scores

Figures 8 and 9 below display a quantitative assessment of fit of each model to the data by showing box plots of the KS-scores of the train and test sets for each model, for each of the stations. The yellow line indicates the median, in the usual box plot fashion, and the points outside the whiskers indicate *outliers*. The lower the median KS-scores are, the better the fit is for the models.

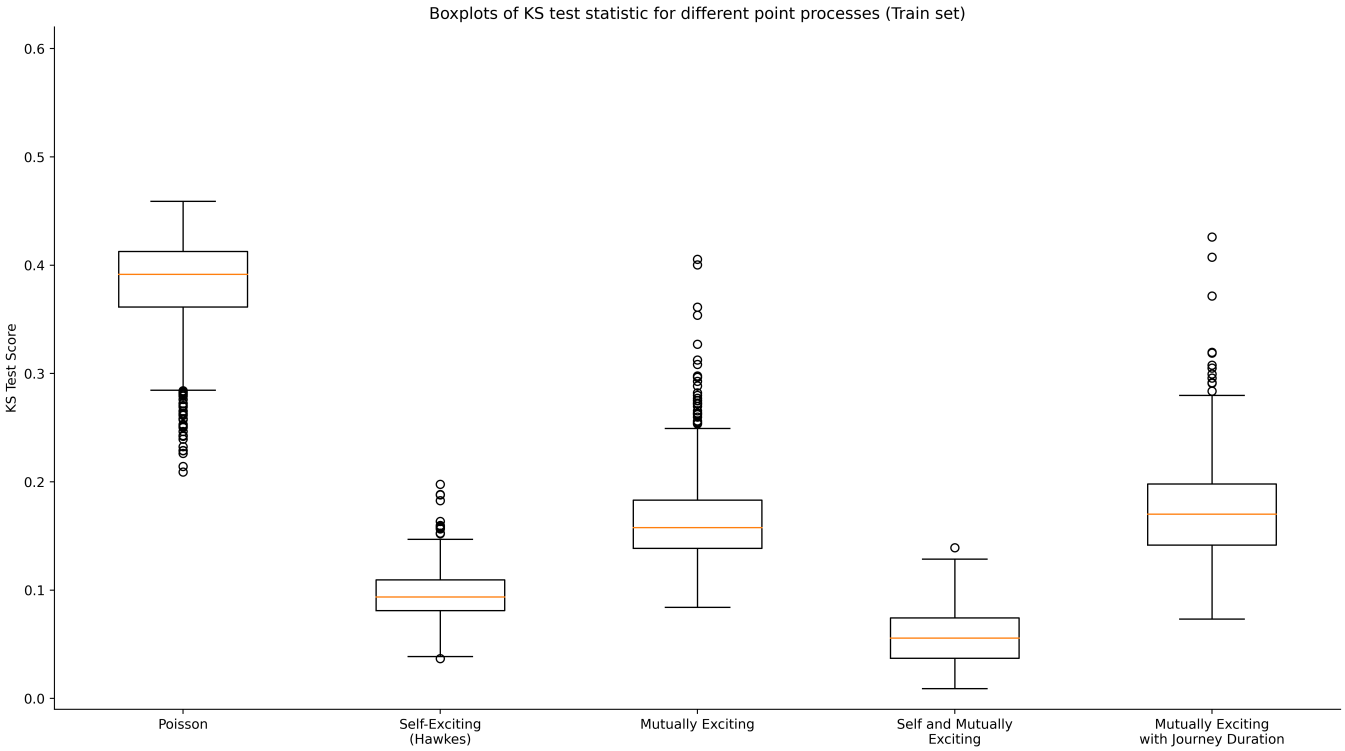


Figure 8: Train set KS-scores for each model.

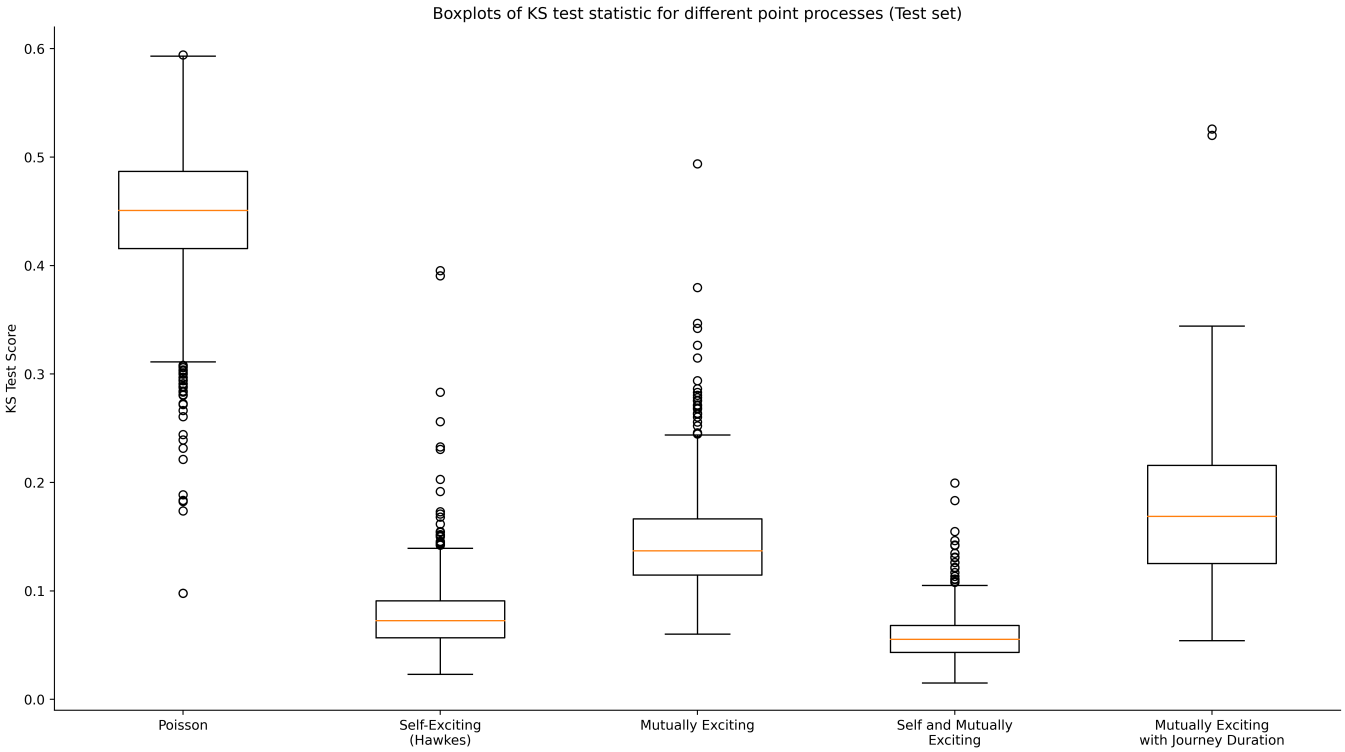


Figure 9: Test set KS-scores for each model.

We can see the Poisson model generally has the worst fit, as is indicated by its high median. However, there are many outliers for this model, with some stations having low KS-scores under this model. This could indicate that some stations indeed behave as simple Poisson point processes, and do not show much SE or ME behaviour. Model 2 already has excellent fit, with a median KS-score close to 0.1. Model 3 has slightly worse fit, possibly indicating that self-exciting behaviour is more key to the dynamics of the system than mutually exciting behaviour. Model 4 has the best fit, as it includes both key behaviours of the system. Model 5 generally has a worse fit than model 3 and a larger variance, which could indicate that the duration times of journeys are not as important to the behaviour of the system as previously hypothesized.

The final figures we consider to assess fit of our models show how the KS-scores depend on the popularity of each station. Here, we plot the KS-scores of both the test and train sets against the number of departures from each station over the observation time of our data, and plot a line of best fit, which is shown in orange.

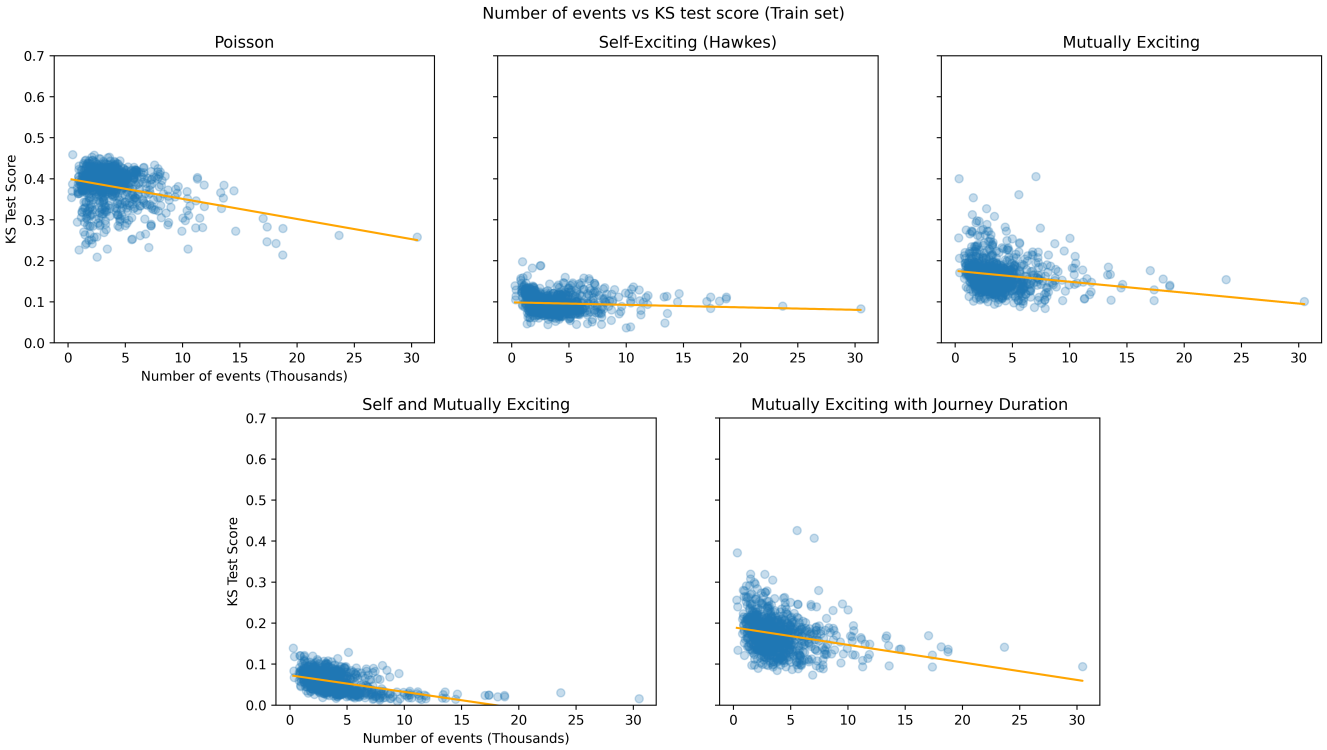


Figure 10: Scatter plots of train set KS-scores against station popularities for each model. The orange line is a linear fit computed using linear regression.

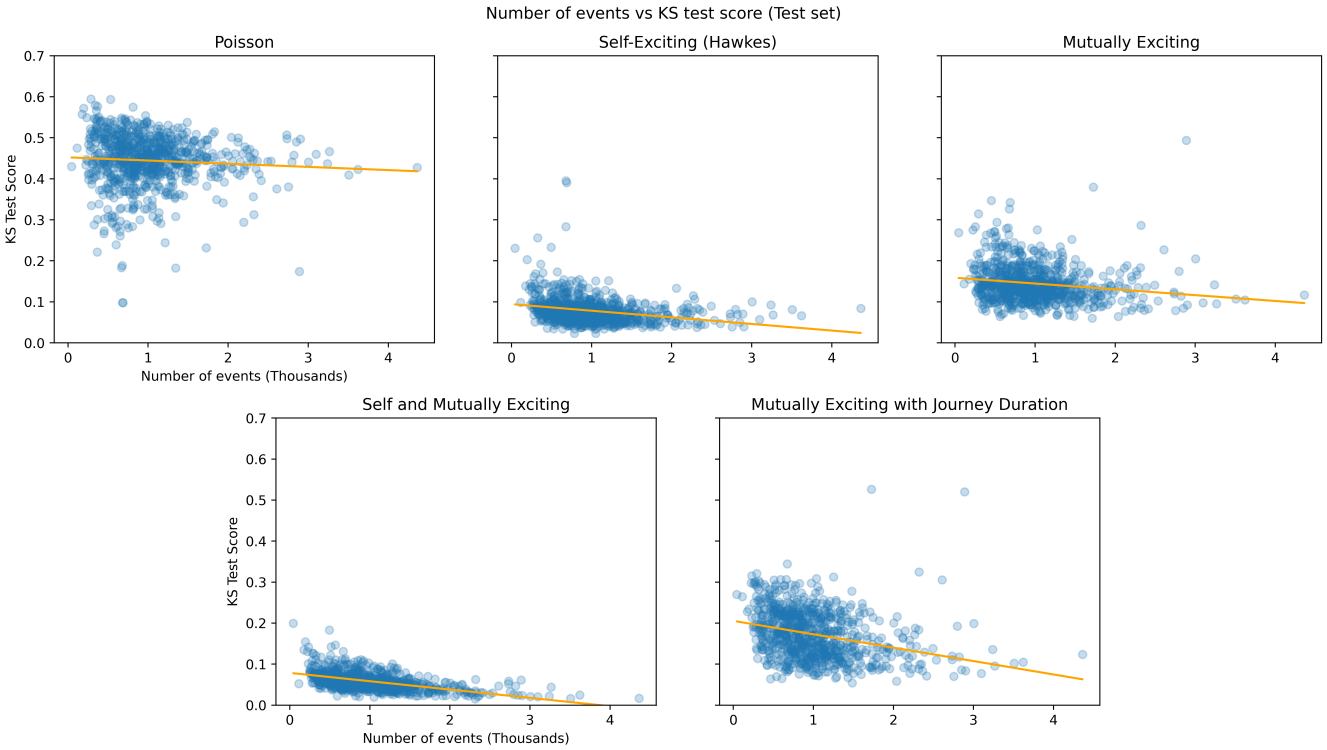


Figure 11: Scatter plots of test set KS-scores against station popularities for each model. The orange line is a linear fit computed using linear regression.

We can see that there is a downward trend in Figures 10 and 8 for most of our models, indicating that stations with higher popularity have a better fit. One likely cause of this is that there is a greater amount of data to describe the behaviour of these stations, thus the fit is better.

Additionally, more popular stations could display more SE behaviour due to generally being used more often by groups. They could also display more ME behaviour due to being empty more often because of their popularity. These behaviours would make more popular stations fit our SE and ME models better.

Finally, the sharpest downward trend in the above graphs is for model 5, which includes duration times in the intensity function. This could indicate that the duration times are more key to the behaviour of more popular stations, and are not so important for less popular stations.

# 7 Discussion

The theory of self-exciting and mutually exciting point processes is a very powerful one, and has yielded fruitful results for modelling our dataset. Having proposed five models with varying levels of complexity, we have derived recursive forms for the log-likelihood of our models and found optimal parameters via the Nelder-Mead optimisation method. We then assessed the performance of these models by making use of the Random Time Change theorem by transforming the departure times for each station by the corresponding compensator functions and computing KS test statistics for each model.

The conclusion of our analysis is that the SE and ME models are an appropriate choice for our data, as the fit for models 2 and 3 is very good, and excellent for model 4. This is reasonable, as one bike being taken out of a station might imply a group is taking out multiple bikes leading to clustering behaviour (ME) and one bike arriving at an empty station might create a new journey soon as someone takes advantage of it (SE).

## 7.1 Model Performances

As shown by Figures 7 and 8, model 1 was ineffective at modelling our data, which is likely because the system displays self-exciting and mutually exciting behaviour which is ignored by this model. Model 2 fit the data much better, achieving lower KS-scores than model 3, which likely indicates that self-exciting behaviour is more key to the dynamics of the system than mutually exciting behaviour. Both behaviours taken into account, however, produced the best fit for the data as model 4 has the best KS-scores out of all the models. Finally, contrary to our predictions, model 5, which was designed as an improvement to model 3, had slightly worse performance than the latter. This could be either because the duration times are not as instrumental to the dynamics as previously believed, or because the multiplication of two exponential terms inside the intensity function makes it decay very fast, which makes this function difficult to fit to the data. The Further Work section explores different models that could have been used instead. The downward trend for model 5 in Figure 10 could suggest that the duration times of journeys are indeed important for certain stations whilst for other they have none or little effect.

## 7.2 Limitations

There are some limitations to our above analysis which are important to recognise. Firstly, limitations to do with the dataset include:

- The time span of observation of our data is only 16 weeks, potentially providing limited information about the systems which would be better captured by a longer observation time.
- The resolution of the data is only up to the minute, which could hide some key information about the system.
- The data does not include all possible factors that could influence departures, such as the weather, the time of the day, or even the day of the week. Indeed, there is reason to believe the time of the day greatly influences the rate of departures, as shown in the histogram **??**. However, our models keep the same background intensity for bikes even at night, which is likely inappropriate.

- The amount of data is not the same for all stations, which could have produced the trends observed in Figure 11.

Secondly, limitations to our models and analysis include:

- The intensity function for model 5 may not have been an accurate description of how duration times influence mutually exciting behaviour at stations. Different models are discussed in the next sub-section.
- We have only used one method of assessing goodness of fit of our models. One further method we could have used is the Lewis test, which is detailed in the Appendix.

## 7.3 Further work

Given additional time, we would have explored how daily patterns could be added to the conditional intensity function, as was suggested is reasonable above. This could be achieved by modelling the background rate $\lambda_i(t)$ in the intensity as a function changing over time according to certain parameters.

Another possible area for further work is to pursue different options for models which include the duration times of journeys, including a version of Model 5 which also has a self-exciting term, and models which have a different form for how the duration time of journeys affects excitation. One version that was tried was the following:

$$\lambda_i^*(t) = \lambda_i + \sum_{t'_{i,k}<t} \frac{\alpha_i}{d'_{i,k}} \exp\left\{-\beta_i(t - t'_{i,k})\right\} + \sum_{t'_{i,k}<t} \frac{\alpha'_i}{1 + |d'_{i,k} - 30|} \exp\left\{-\beta'_i(t - t'_{i,k})\right\}$$

This model has a weaker decay for the excitation around durations of 1 and 30, which was hypothesized as a cause for the weak performance of model 5. However, this model showed very poor performance when tested. This could indicate that the duration times are in fact not as instrumental to the dynamics of the system as previously believed.

Another possible model that was considered is one in which different parameters are modelled for journeys with different duration times. Due to the fee for journeys over 30 minutes, it is reasonable to assume that most journeys under 30 minutes are for commuting and most journeys over it are for leisure. Thus, these two different journey reasons could lead to different behaviour of the system. This is modelled using a delta-function:

$$\gamma'_{i,k} \coloneqq \begin{cases} 1 & \text{if } d'_{i,k} > 30 \\ 0 & \text{otherwise.} \end{cases}$$

we then have that the intensity function would be given by:

$$\lambda_i^*(t) = \lambda_i + \sum_{t'_{i,k}<t} \gamma'_{i,k}\alpha_i \exp\left\{-\beta_i(t - t'_{i,k})\right\} + \sum_{t'_{i,k}<t} (1 - \gamma'_{ik})\alpha'_i \exp\left\{-\beta'_i(t - t'_{i,k})\right\}$$

This model displays mutually exciting behaviour which depends upon the journey duration in a binary manner, which could be a better model for our data.

Indeed, there are many different types of models that can be created to describe this situation, but generally simpler models are favoured, as they are more robust and universal. Thus model 4 is likely a very good model to describe this data, as its median KS-score is very low and it is relatively simple.

## 7.4 Applications

Understanding the patterns and behaviors of departure times of bike journeys for each station is crucial for service providers such as Transport for London (TfL), who provide the Santander bikes, to maintain the efficiency of their bike-sharing system. **Starvation** and **congestion** are frequently occurring problems of bike sharing (Foschi, 2020). Starvation occurs when no more bikes are available for rental at a station, and congestion occurs when stations are full, preventing riders from returning their bikes at the station. These are so-called *competing* objectives, since mitigating congestion might increase starvation and vice versa. The proposed models find applications mainly in simulating future behaviour of the Santander cycle stations, helping in dynamic bike allocation and optimal bike re-balancing across the stations, preventing bike shortages or surpluses across stations.

# Appendix

## A Properties of Poisson point processes

Here we enumerate useful properties of Poisson point processes used in proving that inter-arrival times follow an Exponential distribution.

- **Time homogeneity**
  Denote $P(k, \tau)$ to be the probability of $k$ arrivals in an interval of duration $\tau$. During the time interval of that given length, there is a random number of arrivals (events) following a particular probability distribution with pmf $P(k, \tau)$. Time homogeneity means this probability distribution only depends only on the **length** of the time interval, and **not on the exact location** of the interval on the time axis itself. Two interval of length $\tau$ will have the exact same statistical behavior of arrivals described by the Poisson($\lambda\tau$) distribution (MITOpenCourseWare, 2012).

- **Small interval probabilities**
  For a very small time interval of length $\tau$:

$$P(k, \tau) \approx \begin{cases} \lambda\tau & \text{if } k = 1 \\ 1 - \lambda\tau & \text{if } k = 0 \\ 0 & \text{if } k > 1 \end{cases} \tag{1}$$

  The last equation assumes that it is very unlikely for two or more events to happen during the same time interval which is very small, and therefore its probability is negligible.

- **Memorylessness of inter-arrival times**
  Because inter-arrival times follow an Exponential distribution, they are memoryless: the time to the next arrival is independent of the past. This can be represented mathematically as:

$$P(\tau_i > s + t \mid \tau_i > s) = P(\tau_i > t)$$

  For example, say we are waiting for the arrival of the next bus. Just because we have waited for s minutes and have not seen the bus, it does not mean the bus is any more likely to appear in the next t seconds. The fact that we have waited for a really long time or just come to the bus stop does not make any difference in the arrival time of the bus.

We now show the proof that the inter-arrival times of a simple Poisson process follow an Exponential distribution:

$$\begin{aligned} P(\tau_1 > t) &= P(0 \text{ event in time interval } [0, \text{t}]) \\ &= P(N(t) = 0) \\ &= \frac{e^{-\lambda t}(\lambda t)^0}{0!} \\ &= e^{-\lambda t} \end{aligned} \tag{2}$$

so $P(\tau_1 \leq t) = 1 - e^{-\lambda t}$ which means $\tau_1 \sim \text{Exp}(\lambda)$

$$\begin{aligned} P(\tau_2 > t \mid \tau_1 = s) &= P(0 \text{ event in the interval } [s, s+t] \mid \tau_1 = s) \\ &= P(0 \text{ events in the interval } [s, s+t]) \\ &= P(0 \text{ events in the interval } [0, t]) \text{ (**Time Homogeneity**)} \\ &= P(N(t) = 0) \\ &= e^{-\lambda t} \end{aligned} \tag{3}$$

so $P(\tau_2 \leq t) = 1 - e^{-\lambda t}$ which means $\tau_2$, and similarly for $\tau_3, \tau_4, .. \sim$ Exponential($\lambda$)

## B Lewis Test

The Lewis Test is another test which can be used to assess fit of a Poisson process, which is generally more powerful than the simple p-value test (Laub, Taimre, Pollett, 2015). It relies on the fact that if $t_1^*, \ldots, t_N^*$ are the arrival times for a Poisson process with unit rate, then $\{t_1^*/t_N^*, \ldots, t_{N-1}^*/t_N^*\}$ are distributed as the order statistics of a uniform $[0, 1]$ random sample (Laub, Taimre, Pollett, 2015). By using the random time change theorem to obtain $t_j^* = \Lambda_i(t_j)$ which follow the distribution of a Poisson process with unit rate by the null hypothesis, we can test whether each of our models are good fits for the data by taking the KS statistics of this test for each station (comparing quantiles of our transformed time ratios with theoretical quantiles for the corresponding order statistics of a uniform distribution).

## C Station location data

The first five rows of the station location dataframe are shown below:

| | Station.Id | StationName | longitude | latitude |
|---|---|---|---|---|
| 0 | 1 | River Street, Clerkenwell | -0.109971 | 51.5292 |
| 1 | 2 | Phillimore Gardens, Kensington | -0.197574 | 51.4996 |
| 2 | 3 | Christopher Street, Liverpool Street | -0.084606 | 51.5213 |
| 3 | 4 | St. Chad's Street, King's Cross | -0.120974 | 51.5301 |
| 4 | 5 | Sedding Street, Sloane Square | -0.156876 | 51.4931 |

Table 4: First 5 rows of the station location dataframe.

From this data, we used code provided by our supervisor to extract the exact location between the stations in Km by computing the geodesics joining the coordinates of pairs of stations.

## D Code

All of the code used in this project can be found in the following GitHub repository:

https://github.com/carlosaccp/Santander-SE-ME

# References

Clapham, M. E. (Jan. 2016). *10: Check Kolmogorov-Smirnov*. URL: https://youtu.be/ZO2RmSkXK3c.

Foschi, R. (2020). "A Point Processes Approach to Bicycle Sharing Systems' Design and Management". *Available at SSRN 3655288*.

Laub, P. J., Taimre, T., Pollett, P. K. (2015). *Hawkes Processes*. DOI: 10.48550/ARXIV.1507.02822. URL: https://arxiv.org/abs/1507.02822.

MITOpenCourseWare (Nov. 2012). *Poisson Process I*. URL: https://youtu.be/jsqSScywvMc.

Ogata, Y. (1988). "Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes". *Journal of the American Statistical Association* 83.401, pp. 9–27. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478560.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford science publications. OUP Oxford, pp. 320–322. ISBN: 9780198507659. URL: https://books.google.co.uk/books?id=M-3pSCVxV5oC.

Price-Williams, M., Heard, N. (2020). "Nonparametric self-exciting models for computer network traffic".

Rizoiu, M.-A., Lee, Y., Mishra, S., Xie, L. (2017). *A Tutorial on Hawkes Processes for Events in Social Media*. DOI: 10.48550/ARXIV.1708.06401. URL: https://arxiv.org/abs/1708.06401.

Scholarpedia (Feb. 2009). *Nelder-Mead algorithm*. URL: http://www.scholarpedia.org/article/Nelder-Mead_algorithm.

Tench, S., Fry, H., Gill, P. (2016). "Spatio-temporal patterns of IED usage by the Provisional Irish Republican Army". *European Journal of Applied Mathematics* 27.3, pp. 377–402. DOI: 10.1017/S0956792515000686.