Self and Mutually-Exciting Point Processes on Networks, with Applications to Bike-Sharing Systems

Written By: James Boucher, Qinyu Liu, Tarun Mistry, Kian Shayeghi, Juno Shin

Supervised By: Dr F. Sanna Passino

May-June 2023

Abstract

This project investigates the application of self and mutually-exciting point processes to modelling event times in bike-sharing systems. In particular, we will be focusing on Hawkes processes, wherein future events are dependent on the entire history of the system. By examining the interplay between self-excitation components based on departure times, and mutual-excitation components based on arrival times, our aim is to identify an effective model for predicting bike journey start times. The estimation of model parameters is performed using maximum likelihood estimation, specifically a recursive form of the log-likelihood (detailed in this paper). Since the data consisted of 4 million events, this was of crucial importance in calculating these parameters efficiently and allowed us to reduce complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$. The study compares four models using real-world data from the Santander Cycles bike-sharing system and explores how effectively these models capture the system's dynamics. The research outcomes contribute to the field of point process modeling, which informs decision-making processes in various domains such as urban mobility planning and resource allocation. By enhancing our understanding of event time dependencies in complex systems, this project offers practical implications for improving the accuracy and efficiency of self and mutually-exciting models.

Contents

1 Introduction

Point processes, or counting processes, are used to measure when particular events take place by observing the behaviour of a system over a period of time. For example, if we were interested in scoring tendencies in basketball games, our "points" would be the times at which a basket is scored. In order to model the future behaviour of this system, we could look at the average number of points scored over a period of time, and predict that the rate will stay constant over the future. This idea is the foundation of the Poisson Process, the simplest type of point process. However, in practice, this model often ends up being inaccurate, mainly because certain events occurring in a system change the probability of events happening in the future, e.g. it would be reasonable to assume that a team stealing the ball would cause a "spike" in their probability of scoring a basket shortly after, or we could say that a team on a streak of scoring is more likely to make their next shot. These are examples of mutual and self-excitation in point processes. In this paper, we will be focusing on a specific type of self-exciting point process: the Hawkes Process, which is dependent on the history of events within a system. Put simply, Hawkes Processes cause "jumps" in probability immediately after events, followed by an exponential decay.

This type of model is useful for predicting Earthquakes (after an earthquake, the probability of aftershocks is likely to instantaneously increase and decay over time), Epistemology (once a subject is infected with a pathogen, the probability of finding pathogens in others around them increases) and even violence following political turmoil [1] [2].

In this project, however, we will be working on modeling the arrival times of Santander bikes at docking stations: the famous bike-sharing system introduced to London in 2010 by the then-mayor Boris Johnson.

We will use four models; namely a Poisson process, a self-exciting process (in particular, a Hawkes process), a mutually-exciting process, and a joint self and mutually-exciting process, in order to model these arrival times. In the first part of the report, we outline the necessary mathematical theory to estimate parameters, implement time-efficient code and evaluate our models. In the second half, we showcase the results of our code graphically and evaluate the validity and performance of each model.

2 Point Processes

2.1 Prerequisites

(The following definitions are from [3], [4], [5]) We consider a collection of points falling in some given space (e.g. time, which we will exclusively work on within this report) as an informal idea of a point process. Formally we have the following:

Definition 2.1 (Point Process). *If a sequence of (real-valued) random variables given by $T = \{T_1, T_2, T_3...\}$ takes finite, non-zero values with $T_1 \leq T_2 \leq T_3 \leq ...$ and the total number of points in a bounded interval is almost surely finite then such a sequence is called a **point process**.*

There are a number of ways to uniquely define a point process. One way is to consider the distribution of time until the next event following the previous one. This gives us the idea of a counting process as a sum of such distributions.

Definition 2.2 (Counting Process). *The random variable $N(t) \in \mathbb{N}$ is called a **counting process** if it takes the value of the number of events occurring between time 0 and time t.*

The terms *counting process* and *point process* refer to the same idea, differing in the sense that the prior considers the total number of events up to a point in time whereas the latter considers the event times themselves. Throughout this paper, the term most appropriate to the context of each section will be used.

Definition 2.3 (Conditional Intensity Function). *Suppose we have an associated history of events up to time t of a counting process N. Then if we can find a non-negative function satisfying*

$$\lambda(t|\mathscr{H}_t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(N(t+\Delta t) - N(t) = 1|\mathscr{H}_t)}{\Delta t}$$

*where we define $\mathscr{H}_t$ to be the history of events up to time t, we call such a function a **conditional intensity function** for the point process.*

We can now define a precursor of the Hawkes process: the Poisson process.

Definition 2.4 (Poisson Process). *Let $(X_i)_{i \geq 1}$ be a sequence of i.i.d exponential random variables with parameter λ known. Let the event times be defined by $T_n = \sum_{n \geq i \geq 1} X_i$. Then the counting process defined by*

$$\sum_{i \geq 1} \mathbb{1}_{\{t \geq T_i\}}$$

*is known as the **Poisson process** with parameter λ. The (conditional) intensity function for the Poisson process is*

$$\lambda(t|\mathscr{H}_t) = \lambda(t) = \lambda$$

where $\lambda \in \mathbb{R}$ constant.

The Poisson process is certainly a useful starting point in calculating arrival times, however, it has limitations which make it inapt for our data. In particular, it assumes a constant rate of events. The Hawkes process extends this idea by conditioning based on the history of previous events. In order to define such a process, we need to examine some key ideas first.

Definition 2.5 (**Self-Exciting Process**). *A **self-exciting process** is a point process in which the occurrence of an event increases the probability of events occurring in the near future. More specifically, it is a point process with conditional intensity function given by*

$$\lambda(t|\mathscr{H}_t) = \lambda_0 + \sum_{T_i<t} \phi(t - T_i)$$

*with $\lambda_0 > 0$ the **baseline intensity** and ϕ the **kernel function**.*

In the context of our data, we are designing a model which assumes that in the period shortly after someone picks up a bike from a station, the probability that someone else picks up a bike is increased. In a general self-exciting process, the model may only depend on a particular subset of the history of events. We will primarily be working with the Hawkes process in this report, which is a special kind of self-exciting process.

Definition 2.6 (**Hawkes Process**). *We call a self-exciting point process in which the conditional intensity function depends on the **entire** history of events a **Hawkes Process**.*

The dependence on previous events is introduced through the kernel function. We will exclusively work with an exponential kernel function of the form $\alpha e^{-\beta(t-T_i)}$ with $\alpha, \beta > 0$ where $\alpha, \beta \in \mathbb{R}$. Intuitively, this represents a "jump" of size α at the time of an event, followed by decay at an exponential rate of β, repeating at each event. In order to prevent our model from 'exploding' to infinity, we condition $\beta > \alpha$, i.e. the function decays at a rate quick enough to keep it from growing uncontrollably.

Definition 2.7 (**Mutually-Exciting Process**). *A **mutually-exciting process** is a point process in which the probability of an event occurring (event a) is increased through the occurrence of a different* *type of event (event b). Its conditional intensity function is given by*

$$\lambda(t|\mathscr{H}_t) = \lambda_0 + \sum_{T'_i<t} \phi(t - T'_i)$$

where the T'_i are the times at which event b occurs. Since we are working with the Hawkes Process, we will use the CIF

$$\lambda(t|\mathscr{H}_t) = \lambda_0 + \sum_{T'_i<t} \gamma e^{-\delta(t-T'_i)}$$

For our data, we fitted a mutually-exciting model by assuming that the arrival of a bike at a station increased the probability of a bike being picked up.

2.2 The Time-Rescaling Theorem

Definition 2.8 (Compensator). *The **compensator** of a point process with conditional intensity function $\lambda(t|\mathscr{H}_t)$ is the function*

$$\Lambda(t) = \int_0^t \lambda(x)dx$$

The main consequence of this result is that for any point process, we can define the compensator regardless of the existence of the conditional intensity function. The time-rescaling theorem is dependent on this definition.

Theorem 1 (Time-Rescaling Theorem). *Suppose we have realised events $0 \leq t_1, t_2, ..., t_n \leq T$ for a given point process N. Consider the transformed times given by $\Lambda(t_1), \Lambda(t_2), ..., \Lambda(t_n)$. These transformed times form a Poisson Process with a unit rate.*

Proof. This proof is attributed to [6]. For simplicity we relabel $X_k = \Lambda(t_k) - \Lambda(t_{k-1})$, also define $X_T = \int_{t_n}^T f(x|\mathscr{H}_t)dx$. We aim to show that the X_k are i.i.d. exponential variables with $\lambda = 1$ for $k = 1, 2, ..., n$. Consider the joint probability density

$$f(X_1, X_2, ..., X_n \cap X_{n+1} \geq X_T) = f(X_1, X_2, ..., X_n)\mathbb{P}(X_{n+1} \geq X_T|X_1, X_2, ..., X_n)$$

The two events $\{X_{n+1} \geq X_T|X_1, ..., X_n\} = \{t_{n+1} \geq T|t_1, ..., t_n\}$ are equivalent, hence we can evaluate part of the right-hand side expression

$$\begin{aligned}\mathbb{P}(X_{n+1} \geq X_T | X_1, ..., X_n) &= \mathbb{P}(t_{n+1} \geq T | t_1, ..., t_n)\\ &= e^{-\int_{t_n}^{T} \lambda(x|\mathscr{H}_{t_n})dx}\\ &= e^{-X_T}\end{aligned}$$

where the final line follows from our earlier definition of X_T. Now we use the multivariable change of variable formula (Port, 1994 as cited in [6])

$$f(X_1, X_2, ..., X_n) = |J| f(t_1, t_2, ..., t_n \cap N(t_n) = n)$$

where $|J|$ is the determinant of the Jacobian matrix. Since we have X_k as a function of $t_1, t_2, ..., t_k$ we can say our Jacobian is upper triangular, hence its determinant is the product of its diagonal elements. This arises via the inverse differentiation theorem (Protter and Morrey, 1991 as cited in [6])

$$J_{kk} = \frac{\partial t_k}{\partial X_k} = \lambda(t_k | \mathscr{H}_{t_k})^{-1}$$

Now substituting this into the equation formed, we obtain

$$\begin{aligned}f(X_1, X_2, ..., X_n) &= \left[\prod_{k=1}^{n} \lambda(t_k|\mathscr{H}_{t_k})^{-1} \cdot \prod_{k=1}^{n} \lambda(t_k|\mathscr{H}_{t_k})\right] e^{\left\{-\int_{t_{k-1}}^{t_k} \lambda(t|\mathscr{H}_t)dt\right\}}\\ &= \prod_{k=1}^{n} e^{-(\Lambda(t_k)-\Lambda(t_{k-1}))}\\ &= \prod_{k=1}^{n} e^{-X_k}\end{aligned}$$

Finally, substituting this into the joint probability density function gives

$$f(X_1, X_2, ..., X_n \cap X_{n+1} \geq X_T) = e^{-X_T} \prod_{k=1}^{n} e^{-X_k}$$

which gives us that $X_1, X_2, ..., X_n$ are i.i.d. exponential with rate 1, completing the proof. □

The Time-Rescaling theorem is incredibly useful, since it allows us to convert any point process into a Poisson process. We can also make use of the fact that, given $t_0 = 0, \Lambda(t_1), \Lambda(t_2) - \Lambda(t_1), ..., \Lambda(t_n) - \Lambda(t_{n-1}) \overset{i.i.d.}{\sim}$ Exponential(1). This result allows us to test the fit of a point process model by applying the Kolmogorov-Smirnov test to a transformed Poisson process.

We will use Maximum Likelihood Estimation to compute estimates for our model parameters. In order for this to be possible, we must derive the likelihood function for each of our models. More specifically, we will be making use of the log-likelihood in each case.

3 Log-Likelihood and Maximum Likelihood Estimation

(The following makes use of [3], [5], [8], [9])

3.1 General Form of the Log-Likelihood

For our counting process, we let $N(t) = n$, with CIF $\lambda(t)$. $t_1, ..., t_n$ are observations of our counting process up to time t. In order for this to be a valid pdf, it must integrate to 1. Subsequently, our exact density is equal to

$$\frac{\lambda(t_i)}{\int_0^t \lambda(u)\, du}$$

First consider the function $f(T_1, T_2, ..., T_n) = \prod_{i=1}^{n} f(T_i|T_1, T_2, ..., T_{i-1})$. We see that

$$\begin{aligned}
\lambda(t) &= \frac{f^*(t)}{1 - F^*(t)} \\
&= \frac{\frac{\partial}{\partial t}F^*(t)}{1 - F^*(t)} \\
&= -\frac{\partial}{\partial t} log(1 - F^*(t)) - \int_{T_n}^{t} \lambda(s)\, ds \\
&= -[log(1 - F^*(t)) - log(1 - F^*(T_n))]
\end{aligned}$$

We note that $F^*(T_n) = 0$ as $T_{n+1} > T_n$ hence

$$\int_{T_n}^{t} \lambda(s)\, ds = -log(1 - F^*(t))$$

Following some rearranging, this gives us $F^*(t) = 1 - e^{-\int_{T_n}^{t} \lambda(s)\, ds}$. By combining the relationship between $\lambda(t)$, $f^*(t)$ and $F^*(t)$ we obtain

$$\begin{aligned}
f^*(t) &= \lambda(t)(1 - F^*(t)) \\
&= \lambda(t)e^{\int_{T_n}^{t} \lambda(s)\, ds}
\end{aligned}$$

Plugging this into the likelihood function

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f^*(t) \\
&= \prod_{i=1}^{n} \lambda(T_i) e^{-\int_{T_{i-1}}^{T_i} \lambda(u)\, du} \\
&= \prod_{i=1}^{n} \lambda(T_i) e^{-\int_{T_0}^{T_n} \lambda(u)\, du}.
\end{aligned}$$

Finally, by defining our $T := T_n$ we get our desired final form of the likelihood

$$L(\theta) = \prod_{i=1}^{n} \lambda(T_i) e^{-\int_{T_0}^{T} \lambda(u)\, du}.$$

Equivalently, we can use the compensator to derive the likelihood. With our observations $0 < t_1, ..., t_n < t$ we can formulate the likelihood

$$\begin{aligned}
L(\theta) &= \mathbb{P}(N(t) = n) \cdot \mathbb{P}(t_1, t_2, ..., t_n | N(t) = n) \\
&= e^{-\Lambda(t_n)} \cdot \frac{\Lambda(t)^n}{n!} \cdot n! \prod_{i=1}^{n} \frac{\lambda(t_i)}{\Lambda(t)} \\
&= e^{-\Lambda(t_n)} \cdot \prod_{i=1}^{n} \lambda(t_i)
\end{aligned}$$

We proceed by considering a small change in time, Δt, and we assume that at most one event can occur during this time. So, on each interval $y(t) \equiv N(t + \Delta t) - N(t)$; the time series is an independent Poisson series with mean $\lambda(t)\Delta t$. Note that observing $N(t_k)$ for $0 < t_k < t$ is equivalent to observing the series

$$y(t_k) = \begin{cases} 1, & \text{if } t_k = t_1, t_2, \ldots, t_n \\ 0, & \text{otherwise} \end{cases}$$

Thus given the observation times $0 < t_1, t_2, ..., t_n < t$, we obtain the likelihood

$$\begin{aligned}
L(\theta) &= \prod_{t < t_k} e^{\lambda(t_k)\Delta t} \lambda(t_k)^{y(t_k)} \\
&= e^{\sum_{t_k < t} \lambda(t_k)\Delta t} \prod_{i=1}^{n} \lambda(t_i)
\end{aligned}$$

If we observe what happens as $\Delta t \to 0$

$$\begin{aligned}&\lim_{\Delta t \to 0} e^{\sum_{t_k < t} \lambda(t_k)\Delta t} \prod_{i=1}^{n} \lambda(t_i) \\ &= e^{\int_0^t \lambda(u)\, du} \prod_{i=1}^{n} \lambda(t_i)\end{aligned}$$

Then we can rewrite the notation into a more technical form $dN(t) \equiv y(t) = N(t + \Delta t) - N(t)$ so that the log-likelihood can be written as

$$\begin{aligned}logL(\theta) &= -\int_0^t \lambda(u)\, du + \sum_{i=1}^{n} log\lambda(t_i) \\ &= -\int_0^t \lambda(u)\, du + \int_0^T log\lambda(t)\, dN(t) \\ (=l)\end{aligned}$$

3.2 Log-Likelihood and MLE for the Poisson Process

Using the general form, our log-likelihood is

$$\begin{aligned}l &= -\int_0^t \lambda du + \sum_{i=1}^{n} log\lambda \\ &= -\lambda t + log\lambda\end{aligned}$$

To find the maximum, we differentiate with respect to λ

$$\frac{dl}{d\lambda} = -t + \frac{n}{\lambda}$$

and set $\frac{dl}{d\lambda} = 0$, which yields

$$\hat{\lambda} = \frac{n}{t}$$

Finally, we check that this is a maximum

$$\frac{d^2 l}{d\lambda^2}|_{\lambda=\hat{\lambda}} = -\frac{n}{\hat{\lambda}^2} < 0 \quad \square$$

Intuitively, this makes sense: the parameter which is most likely to have generated our data according to the Poisson model is simply the number of events over the time period, i.e. the "rate". We proceed by outlining the log-likelihood for the next 3 models, but finding the MLE is not as simple in these cases, ergo we make use of the optimisation function in R.

3.3 Log-Likelihood for the Self-Exciting (Hawkes) Process

We have to utilise the negative log-likelihood to allow us to use the optimisation function in R, due to the fact that it returns minima, and we would like the maxima of our log-likelihood functions. This is equal to

$$-\log(L(\theta)) = \int_0^T \lambda(t)dt - \sum_{i=1}^{N(T)} log(\lambda(T_i)).$$

Recursive Form of Negative Log-Likelihood

The log-likelihood function $L(\theta)$ for the interval $[0, t_k]$ can be written as

$$\begin{aligned} l &= \sum_{i=1}^{k} \log(\lambda^*(t_i)) - \int_0^{t_k} \lambda^*(u)du \\ &= \sum_{i=1}^{k} \log(\lambda^*(t_i)) - \Lambda(t_k) \end{aligned}$$

We split $[0, t_k]$ into k disjoint intevals $[0, t_1], (t_1, t_2], ..., (t_{k-1}, t_k]$. Then,

$$\begin{aligned} \Lambda(t_k) &= \int_0^{t_k} \lambda^*(u)du \\ &= \int_0^{t_1} \lambda^*(u)du + \sum_{i=1}^{k-1} \int_{t_i}^{t_{i+1}} \lambda^*(u)du \end{aligned}$$

Since λ^* is exponentially decaying for the Hawkes process, we have

$$\begin{aligned} \Lambda(t_k) &= \int_0^{t_k} \lambda(u)du + \sum_{i=1}^{k-1} \int_{t_i}^{t_{i+1}} \sum_{t_j<u} \alpha e^{-\beta(u-t_j)}du \\ &= \lambda t_k + \alpha \sum_{i=1}^{k-1} \int_{t_i}^{t_{i+1}} \sum_{j=1}^{i} e^{-\beta(u-t_j)}du \\ &= \lambda t_k + \alpha \sum_{i=1}^{k-1} \sum_{j=1}^{i} \int_{t_i}^{t_{i+1}} e^{-\beta(u-t_j)}du \\ &= \lambda t_k - \frac{\alpha}{\beta} \sum_{i=1}^{k-1} \sum_{j=1}^{i} [e^{-\beta(t_{i+1}-t_j)} - e^{-\beta(t_i-t_j)}] \end{aligned}$$

We can now cancel out the terms in the double summation

$$\begin{aligned}\Lambda(t_k) &= \lambda t_k - \frac{\alpha}{\beta}\sum_{i=1}^{k-1}[e^{-\beta(t_k-t_i)} - e^{-\beta(t_i-t_i)}] \\ &= \lambda t_k - \frac{\alpha}{\beta}\sum_{i=1}^{k-1}[e^{-\beta(t_k-t_i)} - 1]\end{aligned}$$

Now we substitute in λ^* and Λ

$$l = \sum_{i=1}^{k}\log[(\lambda + \alpha\sum_{j=1}^{i-1} e^{-\beta(t_i-t_j)}] - \lambda t_k + \frac{\alpha}{\beta}\sum_{i=1}^{k}[e^{-\beta(t_k-t_i)} - 1]$$

and define $\mathbf{A}(i)$ such that

$$\begin{aligned}\mathbf{A}(i) &= e^{-\beta t_i + \beta t_{i-1}}\sum_{j=1}^{i-1} e^{-\beta t_{i-1}+\beta t_j} \\ &= e^{-\beta(t_i - t_{i-1})}(1 + \sum_{j=1}^{i-2} e^{-\beta(t_{i-2}-t_j)}) \\ &= e^{-\beta(t_i - t_{i-1})}(1 + \mathbf{A}(i-1))\end{aligned}$$

Finally, if we add the base case $\mathbf{A}(1) = 0$ we end up with

$$l = \sum_{i=1}^{k} log(\lambda + \alpha\mathbf{A}(i)) - \lambda t_k + \frac{\alpha}{\beta}\sum_{i=1}^{k}[e^{-\beta(t_k-t_i)} - 1]$$

The implications of this result are of the utmost importance to our models, due to the fact that our dataset consists of around 4 million observations. The standard form of the likelihood has a complexity of $O(n^2)$, as opposed to this recursive form of the log-likelihood, which has a complexity of $O(n)$, allowing for much greater computational efficiency in the optimisation process.

3.4 Log-Likelihood for the Mutually-Exciting Process

In the mutually-exciting case, we are observing how the arrivals of bikes at particular stations affect departures from that station. Hence, we must introduce a new conditional intensity function for this case

$$\lambda(t) = \lambda_0 + \sum_{t'_k < t} \gamma e^{-\delta(t - t'_k)}$$

where λ_0 is the baseline intensity as before, and the t'_k are the arrival times of bikes at the observed station. Note that γ, δ behave as α, β did in the self-exciting model, and as such are under the same constraints. To derive the recursive form of the log-likelihood, we define

$$\begin{aligned}\mathbf{B}(i) &= \sum_{t'_k<t_i} e^{-\delta(t_i-t'_k)} \\ &= e^{-\delta(t_i-t_{i-1})} \cdot \sum_{t'_k<t_{i-1}} e^{-\delta(t_{i-1}-t'_k)} + \sum_{t_{i-1}<t'_k<t_i} e^{-\delta(t_i-t'_k)} \\ &= e^{-\delta(t_i-t_{i-1})}\mathbf{B}(i-1) + \sum_{t_{i-1}<t'_k<t_i} e^{-\delta(t_i-t'_k)}\end{aligned}$$

where we reintroduce t_i as the departure times of the bikes from the observed station. Simply put, the $\mathbf{B}(i)$ takes into account the number of arrivals at the station between consecutive departures. With base case $\mathbf{B}(1) = \sum_{t'_k<t_1} e^{-\delta}(t_1 - t'_k)$, we can now define our log-likelihood

$$l = \sum_{i=1}^{k} log(\lambda + \gamma\mathbf{B}(i)) - \lambda t_k + \frac{\gamma}{\delta}\sum_{t'_k<t_k}[e^{-\delta(t'_k-t_k)} - 1]$$

Of course, this can also be computed in linear time as in the self-exciting case.

3.5 Log-Likelihood for the Self and Mutually-Exciting Process

For this model, we simply combine the log-likelihoods from the self and mutually-exciting cases, and optimise once more, this time with 5 parameters $(\lambda, \alpha, \beta, \gamma, \delta)$.

$$l = \sum_{i=1}^{k} log[\lambda t_i + \alpha\mathbf{A}(i) + \gamma\mathbf{B}(i)] - \lambda t_k + \frac{\alpha}{\beta}\sum_{i=1}^{N(t_k)} e^{-\beta(t_k-t_i)} + \frac{\gamma}{\delta}\sum_{i=1}^{N'(t_k)} e^{-\beta(t_k-t'_i)}$$

Where t_i and t'_i are departures and arrival times respectively. This is the most general form of the model; if we set $\alpha = \beta = 0$, we are simply left with the mutual-excitation component, and if we set $\gamma = \delta = 0$, we are left with the self-exciting model.

3.6 Reparametrisation

In order to make sure that our parameters fit the constraints we have specified ($\lambda > 0, \beta > \alpha$, etc.), we must amend our model so that our parameters span the real line. This is because the optimisation function in R works over the entire real line, i.e. the optimal parameters it returns could be any real number, potentially going against our constraints. For this reason, we introduce parameters

$$\tilde{\theta} = (\tilde{\lambda}, \tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta})$$

$$\theta = (\lambda, \alpha, \beta, \gamma, \delta) \text{ where } \lambda, \alpha, \gamma > 0, \beta > \alpha, \delta > \gamma$$
$$\tilde{\lambda} = log(\lambda)$$
$$\tilde{\alpha} = log(\alpha)$$
$$\tilde{\beta} = log(\beta - \alpha)$$
$$\tilde{\gamma} = log(\gamma)$$
$$\tilde{\delta} = log(\delta - \gamma)$$

Now, $\tilde{\theta} \in \mathbb{R}^5$ as required. The self and mutually exciting cases follow trivially. Once we have obtained estimates for our parameters using this algorithm, we can plug them into our conditional intensity functions for each station. The next step in testing our model utilises the idea of compensators and the time-rescaling theorem.

4 Utilisation of Compensators and The Recursive Form

4.1 Compensators

To better understand the concept of compensators, let us first consider a basic example of a counting process: the Poisson process. Remember that we have a counting process $N(t)$ which counts the number of events from time 0 to time t. Since a counting process is locally integrable, we know that its compensator exists (See reference [4]). If we have $N(t)$ events to occur from time 0 to time t, then we have the compensator $\Lambda(t) = \lambda t$ in the Poisson case.

More generally, we would like to consider counting processes that have intensity functions which are dependent on time, $\lambda(t)$. We take our mutually-exciting process as an example. In this case, we know that the intensity function $\lambda(t)$ can be denoted by

$$\lambda(t) = \lambda + \sum_{t'_k < t} \gamma e^{-\delta(t - t'_k)}$$

From the earlier definition, we know that the formula of the compensator $\Lambda(t)$ is given by the equation $\Lambda(t) = \int_0^t \lambda(v)dv$. Hence, $\Lambda(t)$ can be denoted by the formula below

$$\begin{aligned}\Lambda(t) &= \int_0^t \lambda(v)dv \\ &= \lambda(t) - \frac{\gamma}{\delta} \sum_{t'_k < t} [e^{-\delta(t - t'_k)} - 1]\end{aligned}$$

Thus, the formula of the compensator at time t_i is

$$\begin{aligned}\Lambda(t_i) &= \lambda t_i - \frac{\gamma}{\delta} \sum_{t'_k < t_i} e^{-\delta(t_i - t'_k)} + \frac{\gamma N'(t_i)}{\delta} \\ &= \lambda t_i - \frac{\gamma}{\delta} B(i) + \frac{\gamma}{\delta} N'(t_i)\end{aligned}$$

where $B(i)$ denotes the formula $\sum_{t'_k < t_i} e^{-\delta(t_i - t'_k)}$.

The compensator is a very useful concept in the field of stochastic processes because it provides many advantages for further research within the field. Firstly, the compensator allows us to reduce the effect of "noise" in the data and consequently helps us provide more accurate parameter estimations. Secondly, compensators allow us to construct martingales $(M(t))$, which can be defined by $M(t) = N(t) - \lambda(t)$. Martingales are crucial in many fields of probability (See relevant materials [10]). Also, compensators of point processes have applications in financial mathematics (see relevant materials [11]). Note that specific conditions of the compensator are dependent on context.

4.2 Recursive Form

We now move on to the recursive form of the compensator, which allows us to calculate the value of the compensator at each point of our data in linear time. In the case of compensators of the self-exciting process, the recursive form is

$$\Lambda(t_i) = \lambda t_i - \frac{\alpha}{\beta}\mathbf{A}(i) + \frac{\alpha}{\beta}N(t_i)$$

Where $N(t_i)$ is the number of events up to time t_i, and $\mathbf{A}(i) = \sum_{t_k<t_i} e^{-\beta(t_i-t_k)}$. Since we have already computed $\mathbf{A}(i)$, we can simply plug in these values and calculate the compensator values in linear time. In order to utilise the time-rescaling theorem, we want to find the difference between the compensator values at consecutive event times

$$\Lambda(t_i) - \Lambda(t_{i-1}) = \lambda(t_i - t_{i-1}) - \frac{\alpha}{\beta}[\mathbf{A}(i) - \mathbf{A}(i-1) - 1]$$

We can also generalise this difference in compensators to the other cases. From here, in order to calculate our p-values (further discussed in the next chapter), we recall the time-rescaling theorem

$$\Lambda(t_i) - \Lambda(t_{i-1}) \sim \text{Exponential}(1)$$

so our p-values are simply

$$e^{-(\Lambda(t_i)-\Lambda(t_{i-1}))}$$

This form is particularly useful when the intensity $\lambda(t)$ itself is a stochastic process that depends on the history of $N(t)$, as it is often possible to write $\lambda(t)$ in a recursive form, which then allows $\Lambda(t)$ to be computed recursively as well.

Note that the specific form of the compensator and the advantages of the recursive form can vary depending on the specifics of the counting process and intensity function in question. In a similar way, the recursive form of compensators is convenient for us to calculate KS scores, which we will introduce in the next chapter.

5 Kolmogorov-Smirnov Tests and Scores, P-Values

The Kolmogorov–Smirnov test is a nonparametric test (meaning it does not assume any specific distribution for the data) where the Empirical Cumulative Distribution Function (ECDF) of a sample and the CDF of our model are compared to determine whether the two models differ from each other.

Definition 5.1 (The Kolmogorov-Smirnov test and score). *The **Kolmogorov-Smirnov** (or **KS**) test calculates the maximum vertical distance between the two CDFs. This quantity is known as the **KS score**.*

Of course, a lower KS score indicates that our model is a good fit for the data, as it shows that our model does not deviate too much from the Theoretical CDF.

Hypothesis testing using p-values

Another way of testing the fit of our model is through the observation of our p-values which we have calculated prior to this section. For a good model, the p-values should be uniformly distributed. The intuition behind this is clear; if our model follows the distribution exactly, then

$$\mathbb{P}(\text{ observing an event of probability} \leq p) = p$$

which is exactly the CDF of the uniform distribution between 0 and 1.

- H_0 - The p-values follow a uniform distribution
- H_1 - The p-values do not follow a uniform distribution

If our observation is at least as extreme as our confidence level, we reject the null hypothesis at that level.

Advantages of the KS test

- The KS test makes use of the empirical continuous distribution function, which, in general, makes it more powerful than much other goodness of fit tests as it makes more direct use of individual events.
- The distribution of KS scores does not depend on the underlying CDF. [12]

6 Results of Models

6.1 Validity of Models

In this section, we consider the work of Ying, Xue 2019 [13]

Definition 6.1 (Overfitting). ***Overfitting*** *is the case where a model's performance on the training set is not maintained in the test set. If overfitting occurs, the model fails to capture the trend in the broader data set.*

Causes of Overfitting

The causes of overfitting are often complicated. However, we can roughly categorise them:

1. **Noise learning on the training set**: This situation makes the "noise" have an impact on the model which may significantly affect future predictions

2. **Hypothesis complexity**: The trade-off in complexity, which is a key concept in statistical and machine learning, is a compromise between Variance and Bias. It refers to a balance between accuracy and consistency. When a model has too many inputs, the model becomes more accurate but on average becomes less consistent.

3. **Size of training set**: If we use a training set that is too large, our model may end up being too specific to the training set and as a result will not generalise well to any test sets.

Solutions

To reduce the effects of overfitting, various strategies are proposed to address these causes.

1. **Early-stopping**: This strategy is used to avoid the phenomenon of "learning speed slow-down", which causes the accuracy of a model to stop improving and potentially start deteriorating after a certain point, due to noise learning.

2. **Data-expansion**: An expanded data set can improve the accuracy of predictions to a great extent, especially in complicated models.

3. **Regularization**: An overfitting model tends to take most (or all) features of a data set into consideration, without considering that some of them may have little to no impact on the results. In order to limit these cases, we can scale the weights of the features accordingly.

Given the form of the data set we have, a disproportionately large training set size is the most likely cause of overfitting for our model. Therefore, data expansion is the most effective method to avoid overfitting.

More on data-expansion

Data augmentation is widely used and proved to be effective as a general strategy to improve models' generalization performance in many application areas, such as pattern recognition However, the enormous size of the data will definitely increase calculation time. To address this issue, there are 4 approaches to expanding the training set:

1) Acquire more training data.
2) Add random noise to the training set.
3) Re-acquire some data from the existing data set through some processing
4) Produce some new data based on the distribution of existing data set

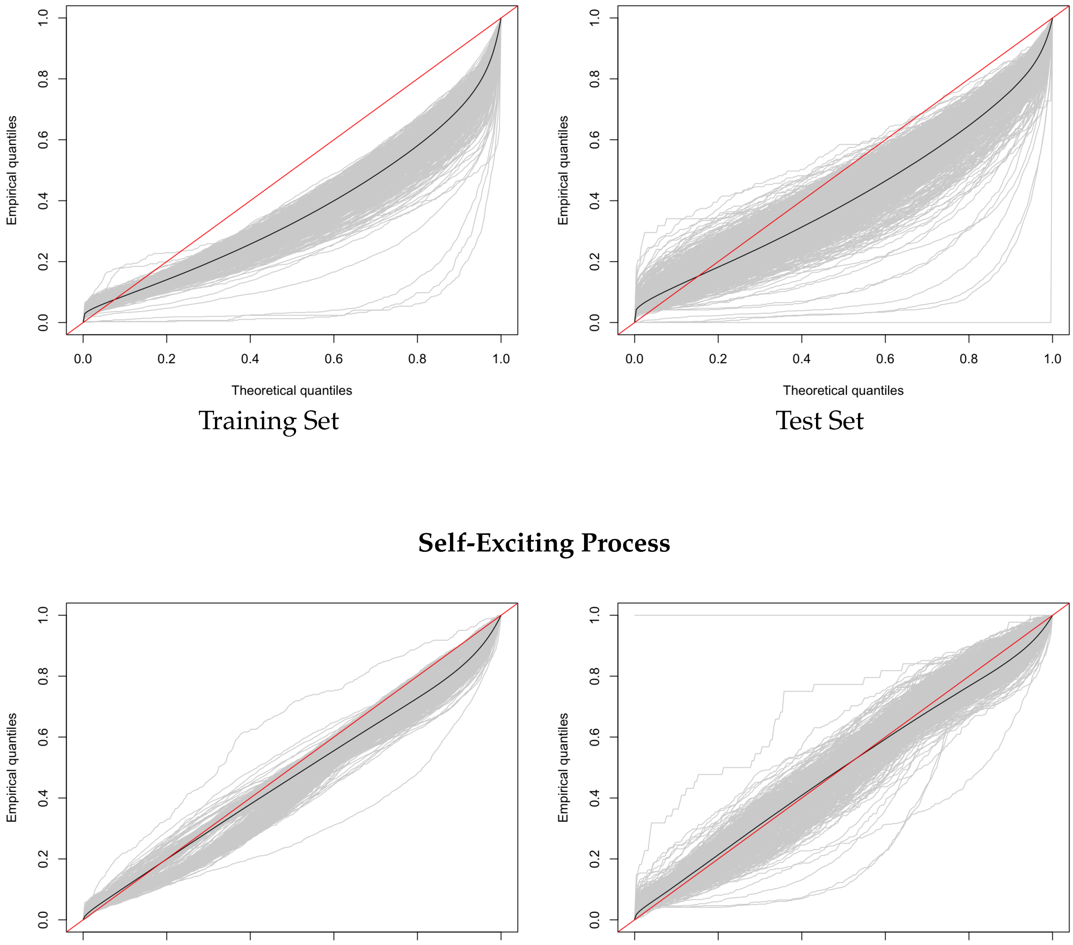
In the case of our data, the application of our models was performed through a conversion of the provided times, followed by an addition of random noise

$$t = \frac{t - t_{min}}{60} + \epsilon$$

Where t_{min} denotes the time of the first event. The addition of the random noise $\epsilon \sim \text{Uniform}(0, 1)$ also means our times appear to be more "continuous", which is useful as the conditional intensity function is a continuous function. Converting the times into minutes elapsed since the first event made the data much more readable than when it was in UNIX time. It is also worth noting that the the bike-sharing system is docked, i.e. the bikes must be left at a particular station with a fixed location. In further models, it may be worth considering factors affecting the accessibility of these stations e.g. the proximity of the stations to people.

6.2 Performance of the models

Figure 1: Q-Q Plots of Training and Tests sets of all 4 Models



Poisson Process

Training Set

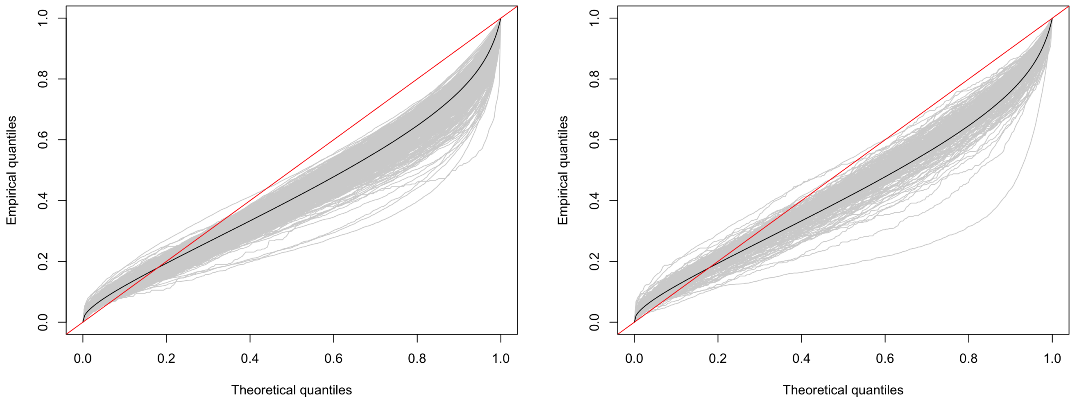
Test Set

Self-Exciting Process

Training Set

Test Set

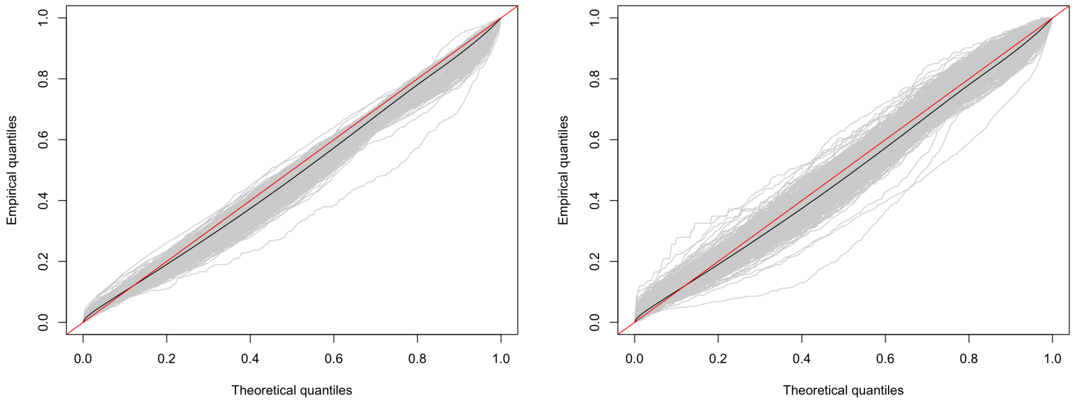
Mutually-Exciting Process



Training Set

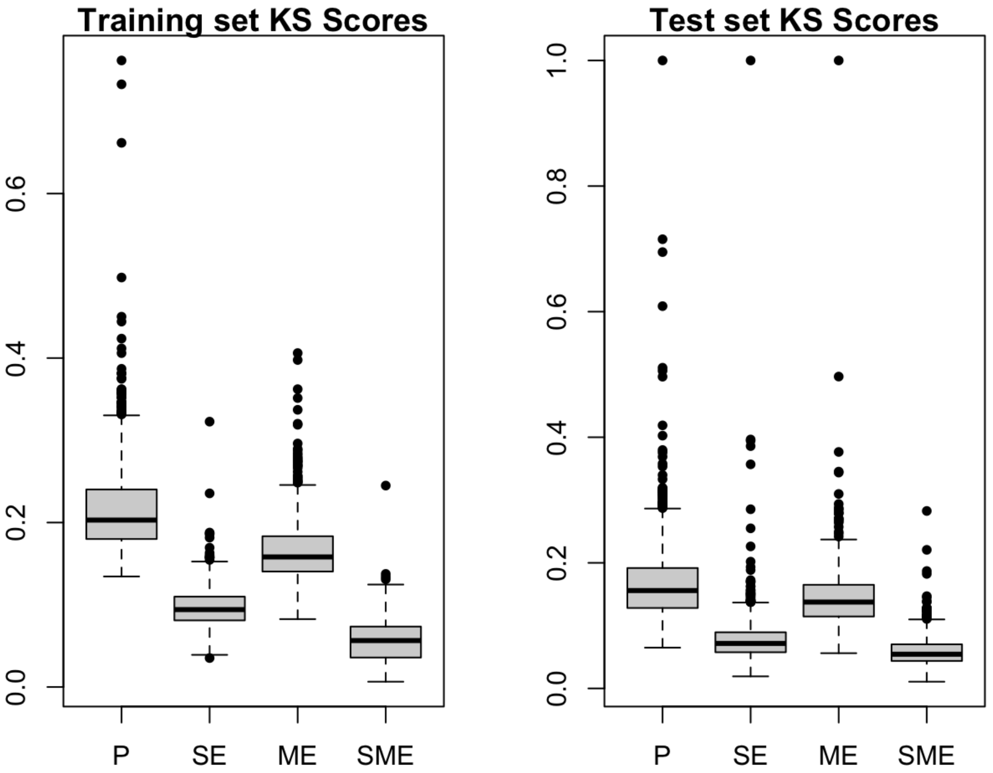
Test Set

Self and Mutually-Exciting Process



Training Set

Test Set



P = Poisson, SE = Self-Exciting, ME = Mutually-Exciting, SME = Self and Mutually-Exciting

Figure 2: Box Plot of KS Scores for Training and Test sets

From **Figure 1**, we see the Q-Q plots created by the models are similar for both the training set and the test. Furthermore, from **Figure 2**, we notice that the KS scores created from the training set and the test are very similar for all 4 models.

Comparing the results

From the Hypothesis testing that we constructed earlier, we always have to reject H_0, because we can never ascertain the true distribution. However, if the p-values of a model are close to the uniform distribution, we reject the null hypothesis with less confidence which suggests that the model is a good fit. Also, low KS scores mean that the vertical distance between the theoretical distribution and the ECDFs are small which also suggests that the model is a good fit. Figures 1 and 2 both appear to suggest that the Self and Mutually-Exciting model is the best fit, closely followed by the Self-Exciting model. The Poisson model is not a good fit, as it has larger KS scores than all other models on average, and the Q-Q Plots suggest that the p-values of the training and test sets do not closely follow a uniform distribution.

7 Conclusion

We can rank the models from most to least appropriate based on our findings in the previous section:

1) Self and Mutually-Exciting
2) Self-Exciting
3) Mutually-Exciting
4) Poisson

Of course, the optimal parameters for the self and mutually-exciting model provide the best fit, as expected. The self-exciting model being superior to the mutually-exciting model in this context tells us that future pickup times for bikes are better determined by creating a model based on past pickup times, rather than past arrival times. This also makes sense; when people pick up a bike from a station, in many cases it is more likely that other people will follow, possibly due to the end of work/school. Arrival times may be a bit more sporadic, as people have journeys of varying lengths, furthermore people who work in the same places will not necessarily live close to one another, i.e. they will arrive at different stations. A situation in which the mutually-exciting model would be a better fit would be at stations where bikes are scarce, as people may be waiting for arrival to pick up a bike. Lastly, the Poisson is always going to be the worst-fitting model, as even in the worst case, the optimal parameters of the other models will result in the kernel function being wiped out, leaving the baseline intensity which will be equal to the constant rate of our poisson process.

Some potential drawbacks of our model include the implicit time-homogeneity property; i.e. we have assumed that the day-to-day (or week-to-week) behaviour of this system doesn't change. This may cause problems if we were to observe the behaviour of the system over a year, as one would expect more people to cycle during the summer/spring than in the winter.

We have included the link to our Git repository in the appendix which contains all of the code written for this project, along with a README file explaining various sections of the code.

After taking all the data into consideration, we can conclude that self and mutually-exciting point processes, in particular the Hawkes process, can be used to provide accurate models for bike-sharing systems. Potential further improvements include a more specialised kernel function, or consideration of other events which may be mutually-exciting.

8 Appendix

Here is the GitHub link to our code: https://github.com/ks920ic/M2R/tree/master

9 References

[1] Lima R. Hawkes Processes Modeling, Inference and Control: An Overview. *Siam Review*. 2023; 65(2): 365-367. https://epubs.siam.org/doi/epdf/10.1137/21M1396927

[2] Johnson N, Hitchman A, Phan D. Smith L. Self-Exciting Point Process Models for Political Conflict Forecasting. *Euro. Jnl of Applied Mathematics*. 2018;29: 685–707. doi:10.1017/S095679251700033X

[3] Laub PJ, Taimre T, Pollett PK. Hawkes Processes. ArXiv [Preprint] 2015. Version 1. https://arxiv.org/abs/1507.02822

[4] Lowther G. *Compensators of Counting Processes*. https://almostsuremath.com/2011/12/27/compensators-of-counting-processes/ [Accessed 16th June 2023]

[5] Rizoiu MA, Lee Y, Mishra S, Xie L. A Tutorial on Hawkes Processes for Events in Social Media. ArXiv [Preprint] 2017. Version 2. https://arxiv.org/abs/1708.06401

[6] Brown EN, Barbieri R, Ventura V, Kass RE, Frank LM. The Time-Rescaling Theorem and Its Application to Neural Spike Train Data Analysis. *Neural Comput*. 2002;14(2):325-46. doi: 10.1162/08997660252741149

[7] Port SC. *Theoretical Probability for applications*. New York: Wiley; 1994.

[8] Pawitan Y. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press; 2001.

[9] Rassmussen JG. Temporal Point Processes and the Conditional Intensity Function. ArXiV [Preprint] 2018. Version 1. https://arxiv.org/pdf/1806.00221.pdf

[10] Harrison JM, Pliska SR. Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their Applications*. 1981;11(2): 215-260. https://www.kellogg.northwestern.edu/research/math/papers/454.pdf

[11] Hayes A. *Martingale System: What It Is and How It Works in Investing*. https://www.investopedia.com/terms/m/martingalesystem.asp [Accessed 16th June 2023]

[12] Srimani S, Parai M, Ghosh K, Rahaman H. A Statistical Approach of Analog Circuit Fault Detection Utilizing Kolmogorov–Smirnov Test Method. *Circuits, Systems, and Signal Processing*. 2012;40: 2091–2113. https://doi.org/10.1007/s00034-020-01572-x

[13] Xue Y. An Overview of Overfitting and its Solutions. *IOP Conf. Series: Journal of Physics*. 2019;1168(2): 2019. doi:10.1088/1742-6596/1168/2/022022

[14] Murdoch DJ, Tsai YL, Adcock J. P-Values are Random Variables. *The American Statistician*. 2008; 62(3): 242-245. https://www.jstor.org/stable/pdf/27644033.pdf

[15] Pillow JW. Time-Rescaling Methods for the Estimation and Assessment of Non-Poisson Neural Encoding Models. *Adv. Neural Information Processing Systems*. 2009;22(1): 1473-1481. https://pillowlab.princeton.edu/pubs/Pillow09$_{T}RandNonPoissModels_{N}IPS.pdf$