Analytical Modelling of Point Process and Application to Transportation

Chapter · June 2018

DOI: 10.1007/978-3-319-90403-0_19

CITATION
1

READS
52

1 author:

Minh Le Kieu
University of Auckland
28 PUBLICATIONS **268** CITATIONS

SEE PROFILE

Analytical Modelling of Point Process and Application to Transportation

Le Minh Kieu

Abstract This chapter aims to explain the inference mechanisms of the expected number of passengers arriving at transit stops. These questions are crucial in tactical planning and operational control of public transport to estimate the impact and effectiveness of different planning and control strategies. The existing literature offers limited number of approaches for these problems, which mainly focus more on the prediction of aggregated passenger counts. We propose two analytical models to model the arrivals of passengers: The first model is a non-homogeneous Poisson Process (NHPP); the second model is a time-varying Poisson Regression (TPR) model. Finally, numerical experiments and case study show the performance of the proposed models using simulated data. The analysis of estimated model's parameters using domain knowledge also provide good insights into the factors that impact the patronage level of buses in New South Wales, Australia.

1 Introduction

Passenger demand plays an essential role in tactical planning and operational control in transportation, especially in public transport, because transit vehicles have to stop for passengers boarding and alighting. Transit tactical planning and operational control, as defined in [9], concerns the decisions to design the exact transit services, e.g. frequency of services and timetables; and the decisions to control the operating service, especially in real time. The questions of modelling the expected number of passengers arrivals at transit stops are essential for these studies. For instance, the total or mean waiting time is often used as the main objective function for public transport tactical planning and operation studies [10, 3, 8, 9], which in turn is estimated using a knowledge of passenger demand.

The expected number of passenger arrivals can be explicitly linked to the estimation of aggregated passenger counts within a time period. Literature currently offers two major lines of research for this problem, one for long-term and the other for short-term passenger demand estimation. Long-term demand estimation models aim to complement long-term transit planning practice, such as in four-step demand modelling [19], route planning and frequency setting [9]. These models are developed to anticipate the approximation of passenger demand in the long-term for transit strategic planning, rather than the tactical planning and operational control problem discussed in this chapter. The other line of research, short-term demand estimation model, favours the use of data-driven and black-box methods, mainly aim for predictions. Examples of them include Neural Network [4, 20], Support Vector Machine [23] and the time-series analysis models [18]. While these methods showed their accuracy and robustness, the majority of them aim to provide predictions rather than an analytical connections between passenger demand and explanatory variables. For transit tactical planning and operational studies, data-driven models for short-term prediction may not be as useful as analytical models, because analytical models can be a part of a holistic framework, where researchers can estimate the passenger demand given the changes in explanatory variables. Existing data-driven methods generally use aggregated counts at previous time steps to predict the count at the next time step by relying on the underlying dynamic relationship between adjacent time steps.

One question which is of interest is *how* passengers arrive at transit stops. Transport researchers are generally interested in modelling and simulating the exact passenger arrival times at transit stops. This information is helpful for various purposes, for instance, to estimate the total travel time for passengers from the moment of arrival at transit stops to the moment of alighting a transit vehicles. Existing studies in transit planning and operational control usually assume a known passenger arrival rate, which is the number of passengers arriving at a transit stop per time unit. The arrival rate allows a convenient simulation of passenger arrivals under one of two approaches: (a) deterministic or (b) stochastic point process. The deterministic approach assumes that passenger arrive uniformly to transit stops, so that the number of boarding/arrived passengers is simply the

Le Minh Kieu
Data61, CSIRO e-mail: leminh.kieu@data61.csiro.au

product of the passenger arrival rate and the time headway between consecutive vehicles. The approach has been used in many earlier studies such as [10] and [13]. [6] and [7] also use a variation of this approach, where a dimensionless parameter is used to represent the marginal increase in vehicle delay resulted from a unit increase in headway. The stochastic point process approach assumes that passengers arrive randomly at stops with a stable arrival rate. In the majority of existing studies, this point process is a Homogeneous Poisson Process (HPP), which aims to model the passenger arrival times using only the arrival rate and the time interval between consecutive arrivals, regardless of the interval starting time. HPP is widely used to model systems with stochastic events, such as modelling the presence of connected vehicle in traffic [25] or traffic incidents [1]. An emerging number of existing studies in public transport have also adopted this stochastic approach, such as [12], [24] and [17]. There is considerable evidence that assumptions of stochastic HPP process for passenger arrivals is reasonable for high-frequency services, such as those with scheduled headway to 10-15 minutes [9]. At longer headways, there is another line of research concerning passengers who time their arrivals with the schedule and service reliability [11, 2]. In this study, we assume that passengers do not consult the schedule prior to arrival at transit stops, thus the use of a stochastic point process such as HPP remains valid.

Existing stochastic processes in literature of public transport assume a stable passenger arrival rate or intensity that does not change over time. A common approach to include time into consideration is to define exogenous time intervals. In each interval, the passenger arrival rate is constant. This approach has limited accuracy, because the passenger arrival process is not fully continuous time-dependent, but rather multiple independent HPP superimposed [22]. Non-homogeneous Poisson Process (NHPP), which allows the arrival rate to be continuous time-dependent, is a substantial advance from the HPP in terms of versatility and accuracy to the model passenger arrival process. NHPP models are not popular in public transit studies, but have been used elsewhere, such as software reliability [14] and finance [5].

This chapter proposes two analytical methods to model expected arrival rate of passengers arriving at transit stops. After the literature review, the first part of the chapter concerns the modelling of exact passenger arrival times using a time-varying Point Process model. Another aspect of the chapter concerns that of the modelling of aggregated counts of passenger demand, using a time-varying Poisson Regression model. This model aims to count *how many* passengers will be at a stop in a specific time period under certain conditions. Only aggregated counts of passenger demand are required to train this model. Finally, we also show the model calibration process using synthetic simulated data.

2 Modelling the exact arrival times with Point Process

In this section, we briefly recap the the fundamentals of point processes and the celebrated *Poisson process*, which would be used to 'count' and further evaluate the passenger demands. The following section serves as the building block for realistic modelling of passenger demands in later sections, to include periodicities in demands.

2.1 A representation of point processes

A point process is a mathematical construct to record times at which event happens, which we shall denote by $T_1, T_2, \ldots$. For example T_1 represents the time when passenger 1 arrives at a bus stop, T_2, represents the following passenger arrival and so on. T_k can usually be interpreted as the time of occurrence of the k-th event, in this case - the k-th arrival. In this paper, we refer to T_i as event times. Formally, we define a counting process N_t as a random function defined on time $t \geq 0$, and takes integer values $1, 2, \ldots$. We define $N_0 = 0$. N_t is piecewise constant and has jump size of 1 at the event times T_i. The Poisson process can be defined as follows:

Definition 1. Poisson process: Let $(Q_k)_{k\geq 1}$ be a sequence of independent and identically distributed Exponential random variables with parameter λ and event times $T_n = \sum_{k=1}^{n} Q_i$. The process $(N_t, t \geq 0)$ defined by $N_t := \sum_{k\geq 1} \mathbb{1}_{\{t\geq T_k\}}$ is called a *Poisson process* with intensity λ.

Memoryless property.

Note that the sequence of Q_k are known as the *inter-arrival times*, and it can be interpreted as follows in terms our modelling context: the first passenger arrives at time Q_1, the second arrives at Q_2 after the first, so on and so forth. One can show that this construct means that passenger arrives at an average rate of λ per unit time, since the expected time between event times is $\frac{1}{\lambda}$. Suppose we were waiting for an arrival of an event, say another bus passenger arrival to a bus stop, the inter-arrival times of which follow an Exponential distribution with parameter λ. Assume that r time units have elapsed and during this period no events have arrived, i.e. there are no events during the time interval $[0, r]$. The probability that we will have to wait a further t

time units is given by

$$p(Q > t+r \,|\, Q > r) = \frac{p(Q > t+r, Q > r)}{p(Q > r)} = \frac{p(Q > t+r)}{p(Q > r)} = \frac{\exp(-\lambda(t+r))}{\exp(-\lambda r)} = \exp(-\lambda t) = p(Q > t). \quad (1)$$

Eq. (1) is said to have no memory and it is one of the special property of the Poisson process. Usually memorylessness is a property of certain distribution rather than a process. It usually refers the waiting time distribution until a certain event; and does not depend on how much time has elapsed already.

Moment generating functions.

We now look at a particular kind of transformed average. The moment generating function φ of a random variable X, is defined as $\varphi_X(s) := E[e^{sX}]$. We now compute the moment generating function of a Poisson distribution $X \sim Pois(\lambda)$:

$$\varphi_X(s) = E[e^{sX}] = \sum_{k=0}^{\infty} e^{sk} p(X = k) = \sum_{k=0}^{\infty} \frac{e^{sk} e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^s)^k}{k!} = e^{\lambda(e^s - 1)}. \quad (2)$$

The moment generating functions are important because each distribution possesses a unique moment generating function. This means that we can infer the distribution from the moment generating function. In addition, the moment generating function of a sum of independent random variables is the product of the moment generating function of the individual random variables.

2.2 Non-homogeneous Poisson Process

The Poisson process, as we defined it so far, is simply characterised by a *constant* arrival rate λ. It is equivalent to an assumption, for example, that public transport passengers arrival rate to stops is the same regardless of the time being midnight or peak periods. It is more useful to extend the Poisson process to a more general point process in which the arrival rate varies as a function of time. Note that the intensity usually depends on the arrival time, not just on the interarrival time. We can define this type of process as non-homogeneous Poisson process (NHPP).

Definition 2. The point process N is said to be an inhomogeneous Poisson process with intensity function $\lambda(t) \geq 0$ with $t \geq 0$, if

$$\begin{aligned} &p(N_{t+h} = n+m \,|\, N_t = n) = \lambda(t)h + o(h) && \text{if} \quad m = 1, \\ &p(N_{t+h} = n+m \,|\, N_t = n) = o(h) && \text{if} \quad m > 1, \\ &p(N_{t+h} = n+m \,|\, N_t = n) = 1 - \lambda(t)h + o(h) && \text{if} \quad m = 0. \end{aligned} \quad (3)$$

Note that if the point process N be a NHPP with intensity function $\lambda(t)$, then $N(t)$ follows a Poisson distribution with parameter $\int_0^t \lambda_u \, du$, i.e. $p(N_t = n) = \frac{1}{n!} \exp\left(-\int_0^t \lambda_u \, du\right) \left(\int_0^t \lambda_u du\right)^n$. One can also show that the number of points in the interval $[s,t]$ follows a Poisson distribution with parameter $\int_s^t \lambda_u \, du$, i.e. $p(N_t - N_s = n) = \frac{1}{n!} \cdot \exp\left(-\int_s^t \lambda_u \, du\right) \left(\int_s^t \lambda_u du\right)^n$.

We can see that the exact event times are needed to calculate moments in the NHPP setting. This next section proposes a public transport demand model and aims to simulate the dynamic and stochastic arrival process of public transport passengers.

2.3 The proposed time-varying intensity function for dynamic and stochastic passenger arrival process.

We propose a parametric form for the rate of demand of passengers:

$$\lambda_t = p c^p t^{p-1} + \varepsilon, \quad (4)$$

where $c > 0$ and $p \in \mathbb{R}$. The paramter ε is usually taken to be fixed and acts as a parameter such that the rate never goes negative (bounded away from zero), since a negative rate of demand is non-sensical. Note that this function is rich enough for

several reasons. When the parameter $p = 1$, it reduces to a constant and we know from above that this specifies the parameter for the Exponential random variables. If this is respected then the data follows a Poisson process. If on the other hand, under the case then $p < 1$, this gives a decreasing curve (see plot). We interepret this as the rate of demand is decreasing. Finally, our choice of intensity function can also handle the case when $p > 1$ - this corresponds to the increasing rate of demand. We summarize the following description below:

- it reduces to a constant when $p = 1$, and hence is able to recover Poisson process should the data respects this,
- when $p < 1$, the rate of demand is decreasing,
- when $p > 1$, the rate of demand is increasing.

Figure 2.3 shows a plot of this intensity. It can be easily noted that this is a generalisation of the HPP, where the rate can be constant (similar to HPP) or varies over time.

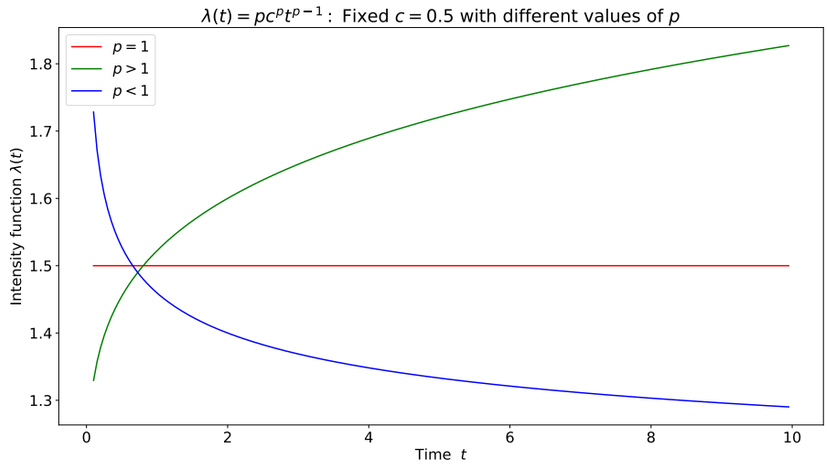

Fig. 1 *A proposed NHPP model with time-varying intensity function.*

2.4 Likelihood function for nonhomogeneous Poisson process

One of the main problems in modelling nonhomogeneous Poisson process is inferring its parameters given data so that we have a calibrated model for the demand of passenger arrivals. Let N_t be a counting process on $[0,T]$ for $T < \infty$ and let $\{T_1, T_2, ..., T_n\}$ denote a set of event times of N_t over the period $[0,T]$. Then the data likelihood L (see [21] for instance) is a function of parameter set θ is:

$$L(\theta) = \prod_{j=1}^{n} \lambda(T_j) e^{-\int_0^T \lambda_x dx}. \tag{5}$$

Let Θ be the set of parameters of the modulating the nonhomogeneous Poisson process. The maximum likelihood estimate can be found by maximizing the likelihood function in Eq. 5 with respect to the space of $\theta \in \Theta$. Concretely, the maximum likelihood estimate $\hat{\theta}$ is defined to be $\hat{\theta} = \arg\max_{\theta \in \Theta} l(\theta)$. It is customary to maximize the log of the likelihood function:

$$l(\theta) \ = \log L(\theta) \ = -\int_0^T \lambda_x dx + \sum_{j=1}^{N(T)} \log \lambda(T_i) \tag{6}$$

This negative log-likelihood can then be minimized with standard optimization packages.

3 Modelling the aggregated passenger demand with Time-varying Poisson Regression

In this section, we argue that a *collective* point process framework can also be formulated as a time-varying Poisson Regression model to estimate the count of arriving passengers to public transport stops. Aggregated counts of passengers are assumed to follow a Poisson distribution, which is consistent with the collective assumption in a Poisson Process (Definition 2). We then further propose a time-varying formulation of Poisson Regression to model the aggregated passenger counts at different time of the day.

3.1 A representation of Generalized Linear Model: Poisson Regression

One of the most common type of regression, the ordinary least squares assumes that the dependent variable Y is normally distributed around the expected value, and can take any real value, even negative values. Another type of regression, the Logistic Regression assumes a binary 0-or-1 dependent variable. These models are often unsuitable for count data, such as aggregated passenger counts, where the data is intrinsically non-negative integer-valued.

Poisson Regression is widely considered as the benchmark model for count data. It assumes the dependent variable Y has a Poisson distribution, and assumes the logarithm of Y can be modelled by a linear combination of X. It is a type of Generalized Linear Model (GLM). Let k be the number of independent variables (regressors). X is a 1-dimension vector $X = (X_1, X_2, \dots X_k)$, which can be both continuous or categorical variables. Poisson Regression can be written as a GLM for counts:

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k = x^T \beta \tag{7}$$

The dependent variable Y has a Poisson distribution, that is $y_i \sim Poisson(\mu_i)$ for $i = 1, ..., N$. The Poisson distribution has only one parameter μ that decides both conditional mean and variance. The conditional mean $\mathbb{E}(y|x)$ and conditional variance $Var(y|x)$ are equal in the Poisson regression model. The following exponential mean function can be written:

$$\mathbb{E}(y|x) = \mu = \exp(x^T \beta) \tag{8}$$

Under the GLM framework and assuming that an n independent sample of pairs of observations (y_i, x_i), the regression coefficient β_j can be estimated using Maximum Likelihood Estimation (MLE). It is worth reiterating that MLE aims to find parameters that maximize the probability that the specified model has generated the observed sample. Given the observed data, we can define the joint probability distribution of the sample as the product of individual conditional probability distributions.

$$f(y_1, \ \dots, \ y_N | x_1, \ \dots, \ x_N; \beta) = \prod_{i=1}^{N} f(y_i | x_i; \beta) \tag{9}$$

As per the previous section, equation 9 is often called *likelihood function*, which is often written in a shorter form:

$$L = L(\beta; y_1, \ \dots, \ y_N, \ x_1, \ \dots, \ x_N) \tag{10}$$

MLE aims to maximise this likelihood function with regard to parameters $\hat{\beta}$:

$$\hat{\beta} = \arg_\beta \max L(\beta; y_1, \ \dots, \ y_N, \ x_1, \ \dots, \ x_N) \tag{11}$$

It is often more convenient to maximise the logarithmic transformation of this likelihood function, as it replaces products by sums and allows the use of the central limit theorem. We define the log-likelihood function of Poisson Regression as:

$$\begin{aligned} \ell(\beta; Y, \ X) &= \log \prod_{i=1}^{N} f(y_i | x_i; \beta) \\ &= \sum_{i=1}^{N} \log f(y_i | x_i; \beta) \\ &= \sum_{i=1}^{N} -\exp(x_i' \beta) + y_i x_i' \beta - \log(y_i \ !) \end{aligned} \tag{12}$$

The estimated regression coefficient β_j that maximizes the value of the log-likelihood function, is found by computing the k first derivatives of the log-likelihood function with respect to $\beta_1, \beta_2, .., \beta_k$ and setting them equal to zero.

$$s_N(\beta; y,\ x) = \frac{\partial \ell(\beta; y, x)}{\partial \beta} = \sum_{i=1}^{N} [y_i - \exp(x_i' \beta)] x_i \tag{13}$$

We define $\hat{\beta}$ as the value of β that solves the first order conditions:

$$s_N(\hat{\beta}; y,\ x) = 0 \tag{14}$$

The system of k equations in 13 has to be solved using an numerical iterative algorithm due to the non-linearity of β. There are a number of existing algorithms in literature that have been well implemented in various statistical packages, such as Newton-Raphson, Broyden-Fletcher-Goldfarb-Shanno (BFGS), Nelder-Mead and Simulated Anneling method.

3.2 Time-varying Poisson Regression model

As we are concerned with the time dimension in the passenger arrival process, the arrival patterns can be considered as a time series Y_t. Autoregressive-based approaches for time-series, such as [18], or Neural Network based [4] approaches show high accuracy and robustness, but focus on short-term demand prediction, rather than developing an analytical formulation which is more useful for statistical studies. This section focuses on proposing an analytical model for public transport planning and operational control. Thus we introduce here a time-varying formulation of Poisson Regression to capture the variations of passenger arrivals to transit stops. We call this model as the Time-varying Poisson Regression (TPR) model.

We are interested in modelling the counts of passenger demand throughout the time of the day. One can observe from aggregated passenger demand data that this count variable has a periodic sinusoidal pattern with two demand peaks at AM and PM rush hours, while gradually reduces to a plateau during off-peak periods. This bimodality distribution of passenger demand is well observed and analysed in literature [15]. A natural modelling approach to capture this sinusoidal pattern is to use a Fourier series:

$$f(x) = \frac{1}{2} a_0 + \sum_{n=1}^{\infty} a_n \cos(nx) + \sum_{n=1}^{\infty} b_n \sin(nx), \tag{15}$$

where

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) dx, \tag{16}$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx) dx, \tag{17}$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx) dx. \tag{18}$$

Here we assume the dependent variable Y is both Poisson distributed and time-dependent, that is $y_t \sim Poisson(\mu_t)$ where $t = 1, ..., N$ are a time-of-day variable. The time-varying formulation of our Poisson Regression model can be written as:

$$\log(\lambda_t) = \alpha_0 + \sum_{k=1}^{K} \left[\beta_h \cos(k \frac{2\pi}{T} t) + \gamma_h \sin(k \frac{2\pi}{T} t) \right] \tag{19}$$

The harmonic terms $\sin(k\frac{2\pi}{T}t)$ and $\cos(k\frac{2\pi}{T}t)$ are added to capture the daily demand patterns. K is the number of harmonics, in which larger K would generally increase the accuracy, but also the complexity of the model. If t is in minutes, T equals 24*60 minutes.

We further increase the adaptability of the model to observed passenger demand data by adding time-invariant independent variables into the model in equation 19. These variables do not have a time-varying formulation. Many variables in practice can be classified into this group, such as weather, day-of-the-week, events or travel cost. For generality, The TPR model can be formulated as:

$$\log(\mu_t) = \alpha_0 + \sum_{h=1}^{H} \left[\beta_h \cos(k \frac{2\pi}{T} t) + \gamma_h \sin(k \frac{2\pi}{T} t) \right] + \sum_{v=1}^{V} \xi_v x_v \tag{20}$$

where V is the number of time-invariant independent variables. Larger V would generally increase the model complexity. The question to whether a time-invariant variable x_i is used in the model is to be decided by considering its correlation to other variables, and its contribution to the prediction of the dependent variable $\log(\mu_t)$.

The TPR model in equation 20 has both time-varying and time-invariant independent variables. The next section will discuss the parameter estimation procedure of this model using MLE.

4 Simulated experiments

In this section, we describe the numerical experiments of NHPP and TPR models using synthetic simulated data. We first generate the synthetic data using predefined parameters, and then fit this simulated data to the proposed NHPP models. The models perform well if they can get back the predefined parameters.

4.1 Non-homogeneous Poisson Process (NHPP)

This subsection discusses the simulation of data from NHPP with predefined parameters as well as the parameter estimation process for NHPP.

Simulation of a Nonhomogeneous Poisson Process using predefined parameters

Given predefined parameters, we briefly explain how we can apply the thinning method [21] to simulate a NHPP. Thinning is method to imitate the trajectory of the counting process over time. Given a NHPP with time-dependent intensity function λ_t, we choose a constant λ^* such that

$$\lambda_t \leq \lambda^*, \quad \text{for all } t, \quad 0 \leq T, \tag{21}$$

for some maturity $T < \infty$. We then simulate a homogeneous Point process with the designated rate λ^* through a sequence of independent and identically distributed exponential distributed random variable, each having a theoretical mean of $(\lambda^*)^{-1}$. We then look at simulated event times of the homogeneous Poisson process and assign some of these to be the event times of the nonhomogeneous Poisson process with intensity function λ_t. We let an event time at a particular time t in the homogeneous Poisson process be also an event time in the nonhomogeneous Poisson process with probability $\frac{\lambda(t)}{\lambda^*}$, independent of the history up to and including time t, and assign no event time otherwise. Hence, the set of event times of the nonhomogeneous Poisson process constructed is a subset of the event times from the homogeneous Poisson process. The resulting pseudo-algorithm reads as follows:

1. Set $T_0 \leftarrow 0$ and $T^* \leftarrow 0$ where T^* denotes the event times of homogeneous Poisson process with intensity λ^*
2. For $j = 1, 2, \ldots n$: generate an exponential random variable $\mathscr{E}$ with mean $(\lambda^*)^{-1}$ and set $T^* = T^* + \mathscr{E}(\lambda^*)$. We then generate a unit uniform random variable and accept the event time $(T_i = T^*)$ if $U < \frac{\lambda(T^*)}{\lambda^*}$, and reject otherwise. The sequence T_i generated from this algorithm is the event times from a nonhomogeneous Poisson process with rate λ_t.

Numerical experiments

We set our parameters for the NHPP model in Equation 4 as in Table 1 as follows:

Table 1 Parameters for NHPP

Variables	Value
p	0.75
c	0.3

The aforementioned thinning simulation is therefore performed for the intensity function $\lambda_t = 0.304 \cdot t^{-0.25} + \varepsilon$. The simulated arrival times are then used to estimate the parameters for the proposed NHPP model in Equation 4. The calibrated parameters should be as close as possible to the predefined parameters in Table 1. Figure 2 shows the calibration results. The calibrated parameters are very similar to the predefined parameters.

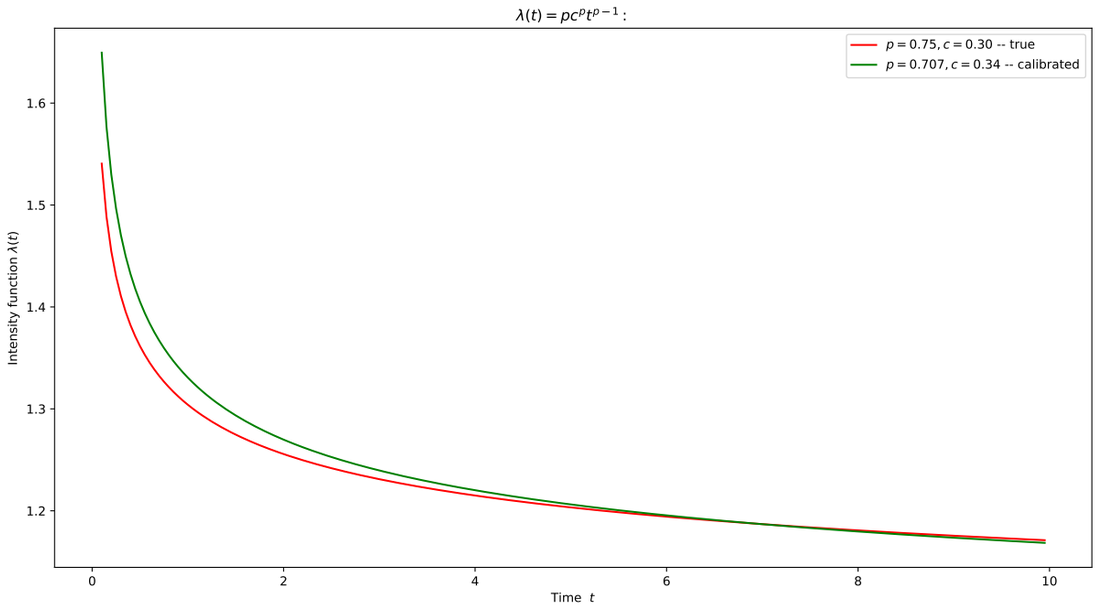

Fig. 2 Calibrated and true trajectory of the proposed NHPP intensity function

4.2 Time-varying Poisson Regression (TPR)

This sub-section describes the generation of synthetic simulated data and the parameter estimation process for time-varying Poisson Regression model

Data generation process

The TPR model has $1 + 2 \times K + V$ parameters, where K is the number of harmonics and V is the number of time-invariant independent variables. The complexity of the model depends on the values of K and V. In this section, we generate the synthetic data using 3 harmonics ($K = 3$) and 3 time-invariant variables ($V = 3$). The time-invariant variables x_i are normally distributed with zero mean, and standard deviation of 0.1, 0.2 and 0.3, respectively. Table 2 shows the chosen parameters for the synthetic simulation.

Table 2 Parameters for synthetic simulation data

Variables	Value	Note
α_0	1	Intercept
β_1	-1	Harmonic 1
γ_1	1	Harmonic 1
β_2	-1	Harmonic 2
γ_2	1	Harmonic 2
β_3	1	Harmonic 3
γ_3	-1	Harmonic 3
ξ_1	0.5	$x_1 \sim \mathcal{N}(0, 0.1)$
ξ_2	0.5	$x_2 \sim \mathcal{N}(0, 0.2)$
ξ_3	0.5	$x_3 \sim \mathcal{N}(0, 0.3)$

We simulate 100 days of data, with the time varies from 4AM to 10PM everyday and each sample is an aggregated passenger count for a 15-minute interval. Figure 3 shows the simulated passenger demand for the first 3 days. The x-axis is the passenger count and the y-axis is every time window for the first 3 days of the dataset.

We use this synthetic simulated data to estimate the parameters for 4 TPR models, from simple to complex model. The details for each model are as follows:

• **H1V1**

The first model is a simple model with 1 level of harmonic and 1 time-invariant variable.

$$\log(\lambda_t) = \alpha_0 + \beta_1 \cos(\frac{2\pi}{T}t) + \gamma_1 \sin(\frac{2\pi}{T}t) + \xi_1 x_1 \tag{22}$$

Table 3 shows the parameter estimates for Model 1.

Table 3 Estimated parameters for Model 1. Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Coefficients	Estimate	Std. Error	z value	$\Pr(>\|z\|)$	
α_0	2.4169	0.0043	564.761	$<$ 2E-16	***
β_1	-0.0316	0.0062	-5.139	2E-07	***
γ_1	0.8776	0.0047	185.005	$<$ 2E-16	***
ξ_1	0.4064	0.0322	12.628	$<$ 2E-16	***

• **H0V3**

The second model ignores the effect of the harmonics. This model only includes 3 time-invariant variables.

$$\log(\lambda_t) = \alpha_0 + \sum_{v=1}^{3} \xi_v x_v \tag{23}$$

Table 4 shows the parameter estimates for H0V3.

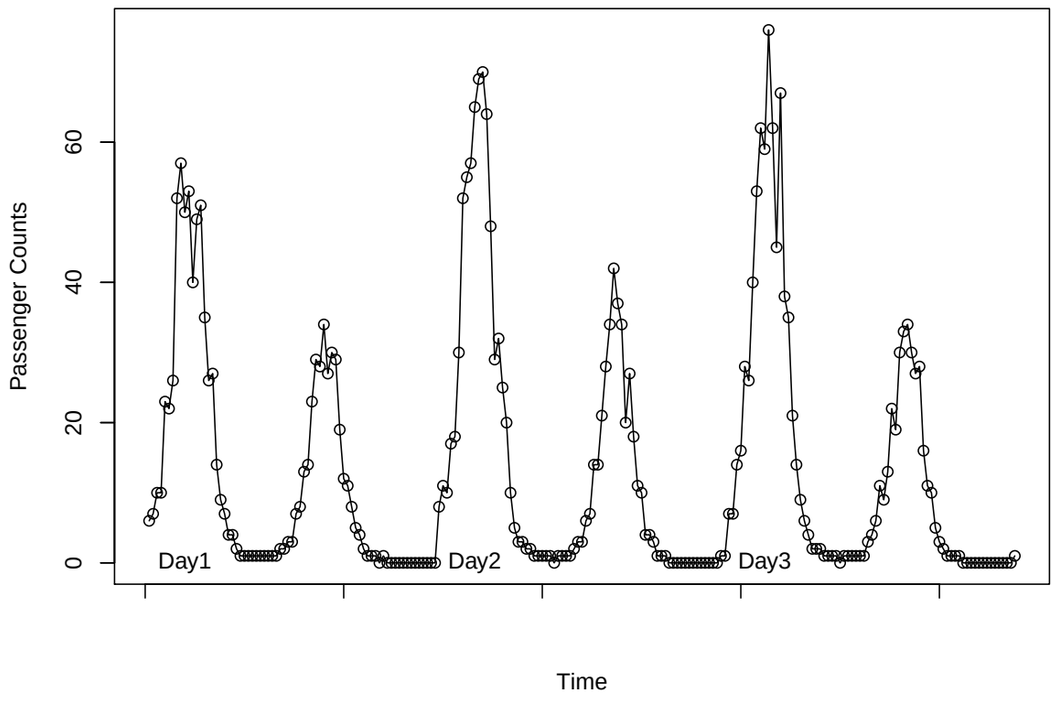

Fig. 3 Synthetic simulated data of passenger demand

Table 4 Estimated parameters for H0V3. Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Coefficients	Estimate	Std. Error	z value	Pr($>\|z\|$)	
α_0	2.5691	0.0033	787.4	< 2E-16	***
ξ_1	0.2484	0.0321	7.739	1E-14	***
ξ_2	0.5940	0.0161	36.824	< 2E-16	***
ξ_3	0.4111	0.0106	38.607	< 2E-16	***

• H3V0

The third model ignores the effect of the time-invariant variables. This model only includes the 3 harmonic levels.

$$\log(\lambda_t) = \alpha_0 + \sum_{h=1}^{H} \left[\beta_h \cos(k\frac{2\pi}{T}t) + \gamma_h \sin(k\frac{2\pi}{T}t)\right] \tag{24}$$

Table 5 shows the parameter estimates for H3V0.

Table 5 Estimated parameters for H3V0. Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Coefficients	Estimate	Std. Error	z value	Pr($>\|z\|$)	
α_0	0.63622	0.03264	19.5	<2e-16	***
β_1	-1.65975	0.05754	-28.84	<2e-16	***
γ_1	1.26252	0.0208	60.7	<2e-16	***
β_2	-1.27614	0.03227	-39.55	<2e-16	***
γ_2	1.27385	0.02118	60.15	<2e-16	***
β_3	0.92572	0.01676	55.22	<2e-16	***
γ_3	-0.83519	0.01336	-62.54	<2e-16	***

• H3V3

The last model includes 3 harmonic levels and 3 time-invariant variables.

$$\log(\lambda_t) = \alpha_0 + \sum_{h=1}^{H} \left[\beta_h \cos(k\frac{2\pi}{T}t) + \gamma_h \sin(k\frac{2\pi}{T}t)\right] + \sum_{v=1}^{V} \xi_v x_v \tag{25}$$

Table 6 shows the parameter estimates for Model H3V3.

Table 6 Estimated parameters for H3V3. Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Coefficients	Estimate	Std. Error	z value	Pr($>\|z\|$)	
α_0	0.64099	0.03144	20.39	<2e-16	***
β_1	-1.61123	0.05556	-29	<2e-16	***
γ_1	1.24552	0.02028	61.43	<2e-16	***
β_2	-1.25812	0.03142	-40.04	<2e-16	***
γ_2	1.24861	0.02058	60.67	<2e-16	***
β_3	0.93607	0.01662	56.34	<2e-16	***
γ_3	-0.85728	0.01304	-65.73	<2e-16	***
ξ_1	0.50175	0.03191	15.72	<2e-16	***
ξ_2	0.50383	0.01596	31.56	<2e-16	***
ξ_3	0.50248	0.01076	46.68	<2e-16	***

Model comparison

The results from Table 3 to 6 show the model performance. It is clear that H3V3 has closest parameters to the actual parameters for synthetic simulation. We further evaluate the goodness-of-fit of each model by comparing their Akaike Information Criterion (AIC) statistics in Table 7.

Table 7 Goodness-of-fit of the proposed models

Model	Degree of Freedom	AIC
H1V1	4	135816.44
H0V3	4	173589.48
H3V0	7	26920.61
H3V3	10	23441.78

As expected, H3V3 shows the best fit among the proposed models. It is because the model incorporates all the determinants in the data, including 3 harmonics and 3 time-invariant variables. H1V1 and H0V3 have significantly lower fits due to the lack of harmonic variables, in which H1V1 has a slightly better fit compared to H0V3 due to the inclusion of one harmonic. The time-invariant variables further increase the goodness-of-fit of modelling. One can see this fact by comparing the AIC statistic of H3V0 and H3V3 because the only different between them is the time-invariant variables.

We also simulate one day worth of new aggregated data to evaluate the performance of each Poisson Regression model. The data is simulated using the same parameters in Table 2 for 73 time periods of 15 minutes each. The new simulated data is used in H1V1 to H3V3 to predict the value of Counts. Figure 4 shows the new data and the estimation results from H1V1 to H3V3. One can easily see that H0V3 does not capture the sinusoidal pattern of the data. Model 1 captures some pattern with limited accuracy, such as the fact that the demand in earlier time periods are larger than those in later time periods. H3V0 well captures the sinusoidal pattern of the data, even the difference between two peaks periods around 8:00 and 16:00. Only H3V3 captures both the sinusoidal pattern and the deviation of the sinusoidal pattern introduced by time-invariant variables. In fact, H3V3 provides very close estimation to the simulated data.

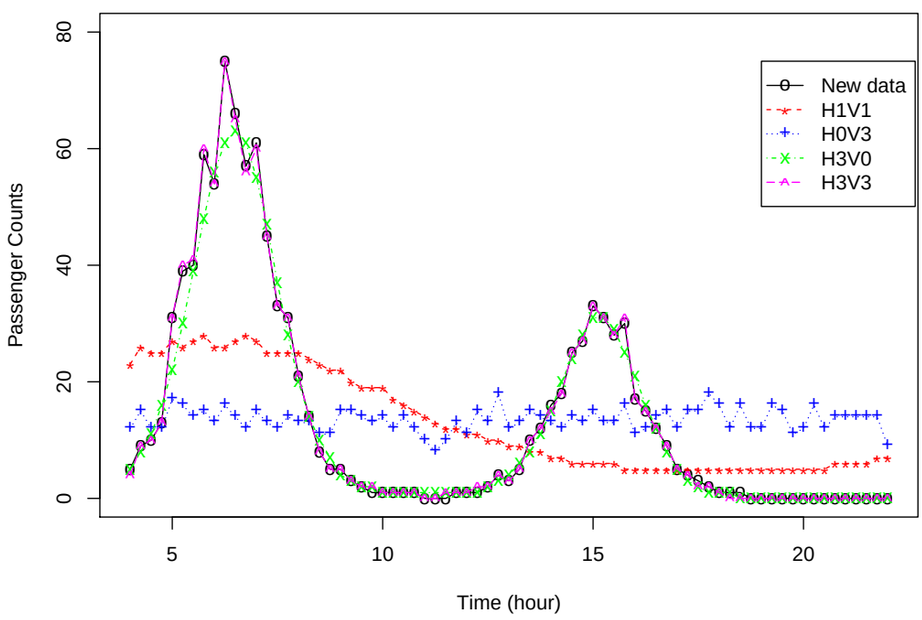

Fig. 4 *Comparison of different Poisson Regression model performance on simulation data*

5 Case study

This section describes a case study where the proposed models are implemented using an observed dataset. We use domain knowledge in Transportation to decide the explanatory variables and to process the data for the models.

5.1 Case study site and dataset

This chapter uses an aggregated Smart Card data from New South Wales (NSW), Australia for the case study. Smart Card is a microchip card, typically the size of a credit card, which has been widely used for ticketing purpose around the world. Examples of Smart Card in public transport are the Oyster Card in London, Opal Card in Sydney, or Myki Card in Melbourne. This chapter uses a 14-day Smart Card data. The data consists of over 2.4 million of Smart Card transactions over large metropolitan areas in NSW, including Sydney, Newcastle and Wollongong City from February to March 2017. The data consists of all bus transactions in the aforementioned metropolitan areas. Each data record contains the following fields:

- $Card_{ID}$: the unique Smart Card ID, which has been hashed into a unique number
- T_{on}: the time when the passenger with $Card_{ID}$ boards a bus
- T_{off}: the time when the passenger with $CardID$ aligns a bus
- S_{on}: the stop/station ID of T_{on}
- S_{off}: the stop/station ID of T_{off}

We only focus our case study on estimating aggregated passenger counts using the Time-varying Poisson Regression (TPR) model proposed in Section 3 because the timestamps in the Smart Card are the boarding and alighting times of a passengers to a bus, rather than the passenger arrival times that are required for the model in Section 2. The objective is to estimate an aggregated count of passengers per time period for each travel choices between a pair of origin and destination. Transit providers can use this proposed TPR model to estimate the change in passenger demand given the changes in explanatory variables such as travel time or transfer time.

The next few subsection describes the required steps to process the input data for the proposed TPR model.

5.2 Journey reconstruction algorithm

For each Smart Card record from each individual passenger, the first step is to reconstruct the full public transport journey with transfers from origin to destination from individual Opal card transactions. This step is essential because Smart Card data only includes the tap-on and tap-off, while we are interested in modelling a completed journeys between a origin and a destination. A completed journeys would naturally give us the following explanatory variables for the TPR model:

- Travel time tt: the time gap between the first tap-on and the last tap-off of a journey
- Transfer time tf: the time gap between a tap-off from a bus to a tap-on to another bus to continue the journey
- Travel distance d: the Euclidean distance between the first tap-on and the last tap-off
- Distance from the origin to CBD d_o: the Euclidean distance from the origin to the Sydney CBD
- Distance from the destination to CBD d_d: the Euclidean distance from the destination to the Sydney CBD

The journey reconstruction algorithm is based on the time and distance gap between individual tap-on and tap-off. Figure 5 shows the proposed journey reconstructing algorithm that based on [16]. We revise the algorithm proposed in [16] by adding the distance gap Δd, which is set to be 500 meter. Δd is added to ensure that the transfer time will only be spent on walking and waiting, rather than any other side activity using a private vehicles.

The time gap Δt is defined to be less than 60 minutes, because in Sydney passengers will receive a discount if they make a transfer within 60 minutes from the last tap-off, so the majority of passengers would continue their journeys within this time frame. The following steps describes the trip reconstruction process.

- Step 1: Querry all the Opal transactions of an individual passenger i. A binary indicator RID is assigned as zero.
- Step 2: For each transaction in the above database, the corresponding transaction is discarded if it is a tap-on reversal, where tap-on and tap-off are at the same location
- Step 3: If RID equals zero, a variable $OriginLocation$ is defined and set as equal to the current tap-on. We also assign a new unique $JourneyID$, change RID to one and move to the next transaction. Otherwise we move to Step 4.

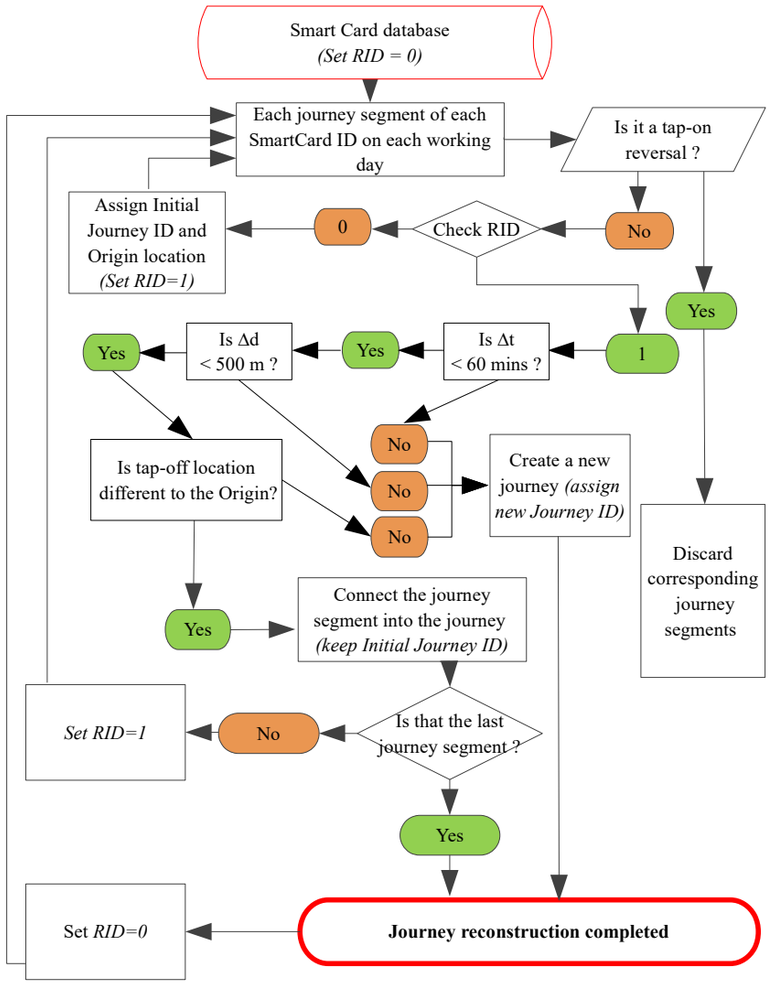

Fig. 5 Journey reconstruction algorithm

- Step 4: Now with *RID* equals one, the current transaction will be assigned the current *JourneyID* if it satisfies three conditions: (1) time gap between the current tap-on and the last tap-off δt is less than 60 minutes, (2) the distance gap δd is less than 500 m, and (3) the current tap-off is different to *OriginLocation*. Otherwise, we assign a new *JourneyID* and set *RID* equals zero.
- Step 5: The journey reconstruction process for the passenger i is finished after the last transaction of the day, otherwise we move to the next transaction.

5.3 Data processing

After journey reconstruction, the remaining data processing in preparation of the inputs for TPR is self-explanatory. Variables tt, tf, d, d_o and d_d are directly calculated from each completed journeys. We then aggregate the completed journeys according to their start time and their *AlternativeID* to produce passenger demand counts. The *AlternativeID* is an indicator of the route choice. It has been defined in a way such that passengers from the same area who make similar choices will have the same *AlternativeID*. Table 8 shows an example of the data used for the case study.

The *AlternativeID*, as shown in Table 8, has been coded in the format: [Origin ZoneID, Destination ZoneID, Mode, Route of the first tap-on, Zone of the first tap-on, Zone of the first tap-off, Route of the last tap-on, Zone of the last tap-on, Zone of the last tap-off]. The Count is total number of passengers who travelled within the same time period, and made the same travel decision as shown in *AlternativeID*.

Table 8 Examples of processed data for the case study

Time	*AlternativeID*	d	d_o	d_d	tt	tf	Count
2:45:00	205_1306_B.N60.205.1306	15947.24	327.688	16223.4	2379	0	20
2:45:00	1701_971_B.N90.1701.83_B.N80.96.971	13963.24	15291.12	10497.86	3660	240	1
2:45:00	2764_144_B.N10.2764.144	9810.014	9865.593	291.957	1104	0	10
2:45:00	1059_1306_B.N60.1059.1306	4439.947	19943.25	16223.4	720	0	3
2:45:00	105_1571_B.520.105.1571	11487.84	1140.505	11981.43	1370	0	6
17:00:00	247_301_B.428.247.301	2319.39	2582.726	4866.599	520	0	12
17:00:00	81_4579_B.616X.81.4579	19921.91	1559.606	20985.03	2298	0	19
17:00:00	242_183_B.428.242.140_B.333.107.183	3738.204	2334.387	1408.359	1740	300	1
17:00:00	305_321_B.428.305.321	1560.36	4570.752	4835.734	450	0	2
17:00:00	81_4568_B.616X.81.3903_B.617X.3903.4568	29567.6	1559.606	30539.1	3750	150	2
17:00:00	6414_6344_B.320.6414.6344	6342.787	111994.2	116865.5	1350	0	6

5.4 Case study modelling results

We use the five explanatory variables, as described in Section 5.1, as the time-invariant variables of the TPR model, as described in Section 3. The dataset is randomly divided into the training dataset, which includes 90% of data points, and the testing dataset, which includes the remaining 10%. We develop TPR models with 3, 4, 5 harmonics and 5 time-invariant variables. Thus the models are named H3V5, H4V5 and H5V5, similar to Section 4.2. We then compare them using Root Mean Square Error (RMSE) as the criteria, which can be calculated as follows:

$$RMSE = \sqrt{\frac{1}{D}\sum_{i=1}^{D}(c_i - \bar{c}_i)^2} \tag{26}$$

Where c_i and $\bar{c}_i$ are the actual and estimated count, respectively. D is the total number of data points in the testing dataset. Thus RMSE measures the mean error of our prediction compares to the observed value. The models are trained using the training dataset, and then tested using the testing dataset.

Table 9 Estimation errors with different TPR models

Model	RMSE
H3V5	7.29
H4V5	6.84
H5V5	6.67

H5V5 shows better performance than H3V5 and H4V5. Table 10 shows the estimated parameters of H5V5. Most of the parameters are significant.

The values and especially the signs of the explanatory variables d_o, d_d, d, tt and tf provide insights into the bus passenger demand in NSW, Australia. The positive sign of d_o and d show that the further passengers are from the Sydney CBD and the longer travel distance, the more likely that a journey by bus will be made. Similarly, the negative sign of d_d shows that if the journey ends near the CBD, the less likely that a journey by bus will be made. This is because the Sydney CBD is well serviced by other public transport modes such as train, light rail and ferry, so bus travels are more for further areas. The negative signs of travel time tt and transfer time tf show that passengers care about these factors. If transit providers can provide services with shorter travel time and transfer time, the patronage for bus will be increased. Passengers concern most about distance of travel and transfer time, which is showed by the fact that the estimated coefficients d and tf are significantly larger than others.

Table 10 Estimated parameters for H5V5. Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Coefficients	Estimate	Std. Error	z value	$\Pr(>\|z\|)$	
α_0	1.6030	0.0038	418.9300	<2e-16	***
β_1	-0.2198	0.0059	-37.3720	<2e-16	***
γ_1	-0.0262	0.0038	-6.9340	<2e-16	***
β_2	-0.1925	0.0027	-72.6250	<2e-16	***
γ_2	-0.0043	0.0054	-0.7850	0.4330	
β_3	0.1262	0.0025	50.9330	<2e-16	***
γ_3	0.1108	0.0049	22.7230	<2e-16	***
β_4	-0.0882	0.0027	-33.2090	<2e-16	***
γ_4	0.2382	0.0032	75.6170	<2e-16	***
β_5	-0.0938	0.0016	-57.8180	<2e-16	***
γ_5	0.0456	0.0015	30.2900	<2e-16	***
d_o	0.0017	0.0001	24.0960	<2e-16	***
d_d	-0.0015	0.0001	-22.2250	<2e-16	***
d	0.0365	0.0001	281.0640	<2e-16	***
tt	-0.0071	0.0000	-147.7890	<2e-16	***
tf	-0.0226	0.0001	-194.6990	<2e-16	***

6 Discussion and conclusion

The inference of expected number of passengers arrivals at transit stops are essentially important for transit tactical planning and operation control studies. We propose a non-homogeneous Poisson Process (NHPP) framework to model the exact records of passenger arrival times. Simulation and calibration for this model are discussed. To estimate the aggregated count of passengers arriving at transit stops, this chapter proposes a time-varying Poisson Regression (TPR) model, given the time and over explanatory variables. This model uses aggregated counts of passenger demand within a time period and several other variables to estimate the passenger counts. The numerical experiments using synthetic simulated data show the calibration process for parameters of both NHPP and TPR.

We also use domain knowledge to implement the TPR model on a case study using observed Smart Card data in New South Wales, Australia. The Transportation domain knowledge is used to define the important explanatory variables for the TPR model, and to process the data. The variables travel time, transfer time, and distance are the most important to explain the bus passenger demand. Domain knowledge has also been used to obtain great insights into the factors that impact the patronage level of buses in NSW, Australia. By analysing the values and signs of variables d_o, d_d, d, tt and tf, we have found that passengers are more likely to use bus when the journey is long, and starts further from the Sydney CBD. They are less likely to use bus if the travel time or transfer time are large; and if the journey is also provided by other modes of transport such as train, light rail or ferry.

The proposed analytical models are useful as a part of a transit tactical planning and operational control framework to estimate the passenger demand at transit stops. Future work includes the use of observed data, a more involved formulation for NHPP model and possibly an inclusion of the autoregressive term for the TPR model.

References

[1] Baykal-Gürsoy M, Xiao W, Ozbay K (2009) Modeling traffic flow interrupted by incidents. European Journal of Operational Research 195(1):127–138

[2] Bowman LA, Turnquist MA (1981) Service frequency, schedule reliability and passenger wait times at transit stops. Transportation Research Part A: General 15(6):465–471

[3] Cats O, Larijani A, Koutsopoulos H, Burghout W (2011) Impacts of holding control strategies on transit performance: Bus simulation model analysis. Transportation Research Record: Journal of the Transportation Research Board (2216):51–58

[4] Celikoglu HB, Cigizoglu HK (2007) Public transportation trip flow modeling with generalized regression neural networks. Advances in Engineering Software 38(2):71–79

[5] Cizek P, Härdle WK, Weron R (2005) Statistical tools for finance and insurance. Springer Science & Business Media

[6] Daganzo CF (2009) A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. Transportation Research Part B: Methodological 43(10):913–921

[7] Daganzo CF, Pilachowski J (2011) Reducing bunching with bus-to-bus cooperation. Transportation Research Part B: Methodological 45(1):267–277
[8] Delgado F, Muñoz J, Giesen R, Cipriano A (2009) Real-time control of buses in a transit corridor based on vehicle holding and boarding limits. Transportation Research Record: Journal of the Transportation Research Board (2090):59–67
[9] Desaulniers G, Hickman MD (2007) Public transit. Handbooks in operations research and management science 14:69–127
[10] Eberlein XJ, Wilson NH, Barnhart C, Bernstein D (1998) The real-time deadheading problem in transit operations control. Transportation Research Part B: Methodological 32(2):77–100
[11] Fonzone A, Schmöcker JD, Liu R (2015) A model of bus bunching under reliability-based passenger arrival patterns. Transportation Research Part C: Emerging Technologies 59:164–182
[12] Fu L, Yang X (2002) Design and implementation of bus-holding control strategies with real-time information. Transportation Research Record: Journal of the Transportation Research Board (1791):6–12
[13] Fu L, Liu Q, Calamai P (2003) Real-time optimization model for dynamic scheduling of transit operations. Transportation Research Record: Journal of the Transportation Research Board (1857):48–55
[14] Hossain SA, Dahiya RC (1993) Estimating the parameters of a non-homogeneous poisson-process model for software reliability. IEEE Transactions on Reliability 42(4):604–612
[15] Kieu LM, Bhaskar A, Chung E (2014) Public transport travel-time variability definitions and monitoring. Journal of Transportation Engineering 141(1):04014,068
[16] Kieu LM, Bhaskar A, Chung E (2015) Passenger segmentation using smart card data. IEEE Transactions on intelligent transportation systems 16(3):1537–1548
[17] Kieu LM, Bhaskar A, Cools M, Chung E (2017) An investigation of timed transfer coordination using event-based multi agent simulation. Transportation Research Part C: Emerging Technologies 81:363–378
[18] Ma Z, Xing J, Mesbah M, Ferreira L (2014) Predicting short-term bus passenger demand using a pattern hybrid approach. Transportation Research Part C: Emerging Technologies 39:148–163
[19] McNally MG (2007) The four-step model. In: Handbook of Transport Modelling: 2nd Edition, Emerald Group Publishing Limited, pp 35–53
[20] Pekel E, Soner Kara S (2017) Passenger flow prediction based on newly adopted algorithms. Applied Artificial Intelligence 31(1):64–79
[21] Ross SM (1997) Introduction to Probability Models, sixth edn. Academic Press, San Diego, CA, USA
[22] Sayarshad HR, Chow JY (2016) Survey and empirical evaluation of nonhomogeneous arrival process models with taxi data. Journal of Advanced Transportation 50(7):1275–1294
[23] Sun Y, Leng B, Guan W (2015) A novel wavelet-svm short-time passenger flow prediction in beijing subway system. Neurocomputing 166:109–121
[24] Toledo T, Cats O, Burghout W, Koutsopoulos HN (2010) Mesoscopic simulation for transit operations. Transportation Research Part C: Emerging Technologies 18(6):896–908
[25] Wang X (2007) Modeling the process of information relay through inter-vehicle communication. Transportation Research Part B: Methodological 41(6):684–700

7 Acknowledgement

This study is supported by Strategic Research Funding at the Data61—CSIRO, Australia and partly by the NSW Premier's Innovation Initiatives Project. The data used in this study is provided by the Transport for NSW, the transport operator of NSW, Australia. The conclusions in this paper reflect the understandings of the authors, who are responsible for the accuracy of the data.