Predicting Short-Term Public Transport Demand via Inhomogeneous Poisson Processes

Aditya Krishna Menon
Data61/CSIRO and the Australian National University
108 North Rd, Acton
Canberra, ACT 2601
aditya.menon@data61.csiro.au

Young Lee
Data61/CSIRO and the Australian National University
13 Garden Street, Eveleigh
Sydney, NSW 2015
young.lee@data61.csiro.au

ABSTRACT

Forecasting short term passenger demand for public transport is a core problem in urban mobility. Typically, this is addressed using Poisson regression or homogeneous Poisson processes. However, such approaches have several limitations, including susceptibility to noise at fine time granularities, and the inability to capture complex non-stationary trends. In this paper, we show how such short term demand can be accurately modelled with an inhomogeneous Poisson process, using a neural network as the underlying intensity. This choice of intensity subsumes existing models as special cases, and is powerful enough to capture certain stylised facts of real-world demand. Experiments on real-world bus arrival data from a large metropolitan area in Australia validate our approach.

CCS CONCEPTS

• **Information systems** → **Data mining**;

KEYWORDS

point process; neural network; urban mobility

1 INTRODUCTION

Public transport is a core ingredient of sustainable urban mobility [18]. The long-term demand for such services requires modelling of various socio-economic factors [4]. Our interest here is in the *short-term* passenger demand, such as the number of passengers expected to arrive at a bus stop within a small time window (e.g. 5–10 minutes). This time-varying short-term demand is crucial to aid in the daily scheduling of services [10, 11, 16].

At first glance, this problem appears to admit a trivial solution: one can discretise time into fixed buckets, and then estimate the average demand for that bucket from several days' worth of historical data. This approach is intuitive, has prior precedent [10, 11, 16], and is largely unassailable when the buckets are sufficiently large (say 30 minutes). However, for smaller buckets, the approach is prone to noise. Figure 1 provides an example of estimates of the mean and 95% confidence interval for a single bus stop, estimated on a months' worth of data (to be described shortly), using buckets of varying granularity. Evidently, at 30 minute buckets, there is a clear predictable pattern with modest variance; however, at 5 minute buckets, there is significant noise, foiling this approach.

The above motivates approaches that can estimate the rate of arrivals in small (possibly *infinitesimal*) time windows. This points to the family of *point process* models [5], which have seen success in spatial urban analytics problems such as modelling ambulance demand [20] and accidents [7]. Closer to our temporal demand forecasting, Allende-Bustamante et al. [1] used a self-exciting (Hawkes) process to model demand at a metro station, while Moreira-Matias et al. [14] used an inhomogeneous Poisson process to model demand for taxis. While effective, such approaches have some limitations; for the former, it is difficult for Hawkes processes to capture cyclical behaviour of a high followed by low demand, while for the latter, the choice of simple *intensity function* (formalised shortly) prohibits more fine-grained predictions.

This raises the motivating question for this work: can we design a stochastic process model that accurately captures the characteristics of short-term passenger demand? We answer this in the affirmative via an inhomogeneous Poisson process with an intensity function governed by a *single-layer neural network*. This choice of intensity is general enough to encapsulate certain stylised facts of transportation demand, such as nonlinear trend characteristics. This theoretical flexibility is borne out in practice: we present experiments on real-world bus arrival data from a large city in Australia to validate our approach. In particular, we show it to be a viable alternative to Poisson regression for *short-term* forecasts.

To make the above claims more precise, we first formalise the underpinnings of point processes and Poisson regression.

2 BACKGROUND AND NOTATION

We begin with some relevant background and notation.

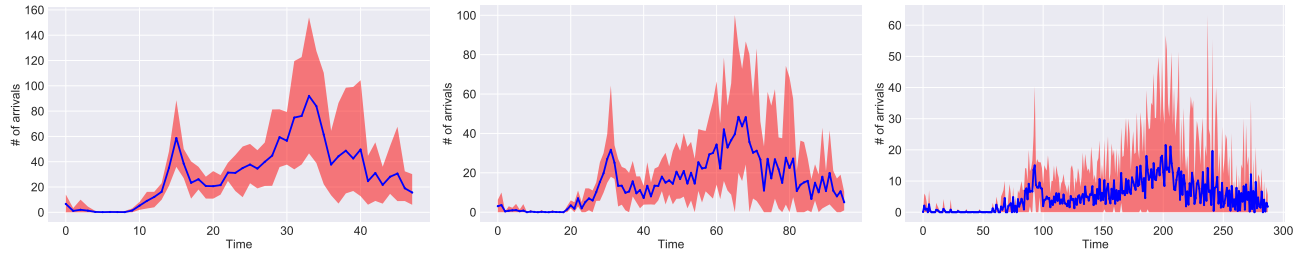
Poisson processes. Suppose we wish to estimate the number of events occurring in a time interval. Consider the stochastic process $(\mathsf{N}_t)_{t \geq 0}$ where the random variables N_t count the number of events that have occurred upto and including time t. This forms an *inhomogeneous Poisson process* (*IPP*) with (locally integrable) *intensity function* $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ if [5]

- $\mathsf{N}_t - \mathsf{N}_s \sim \text{Poisson}(\Lambda(s,t))$, where $\Lambda(s,t) = \int_s^t \lambda(x)\,dx$
- $\mathsf{N}_t - \mathsf{N}_s \perp\!\!\!\perp \mathsf{N}_{t'} - \mathsf{N}_{s'}$ for any $s < t < s' < t'$.

The function Λ is known as the *intensity measure*. When $\lambda(t)$ is a constant, we have the *homogeneous Poisson process* (*HPP*).

CIKM'17 , November 6–10, 2017, Singapore, Singapore

ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3133058



(a) 30 minutes. **(b) 15 minutes.** **(c) 5 minutes.**

Figure 1: Examples of number of arrivals (blue: mean, red: 95% confidence interval) at various time resolutions for a single bus stop.

To use an IPP, one needs to specify a particular form for $\lambda(t)$, and then derive a means of estimating it from data. The former requires domain-specific consideration, which for our problem is deferred to §3. The latter may be done by maximum likelihood estimation (MLE). Suppose our intensity is parameterised by some $\theta \in \Theta$. Given event times $\mathcal{T} = \{t_i\}_{i=1}^N$, for $T \doteq t_n$, the negative log-likelihood of θ is [5, Equation 2.1.9]

$$\mathcal{L}_{\text{IPP}}(\theta; \mathcal{T}) = \sum_{i=1}^{N} -\log \lambda(t_i; \theta) + \int_0^T \lambda(t; \theta)\, \mathrm{d}t. \tag{1}$$

For a suitably differentiable parametrisation, one can estimate θ by minimising Equation 1 using any gradient-based optimiser, such as L-BFGS. The integral above – which is the expected number of events in the entire time interval $[0, T]$ – usually does not have a closed form, but may be approximated by quadrature [6].

Poisson regression. In its most general form, Poisson regression [12] involves predicting nonnegative integer targets $\mathsf{Y} \in \mathbb{N}_+$ for instances $\mathsf{X} \in \mathcal{X}$ in some measurable space $\mathcal{X}$. This is done by positing the existence of a measurable $\mu \colon \mathcal{X} \to \mathbb{R}_+$ such that

$$\mathsf{Y} \mid \mathsf{X} = x \sim \text{Poisson}(\exp(\mu(x))).$$

Given samples $\{(x_i, y_i)\}_{i=1}^N$, and assuming μ is parametrised by some $\theta \in \Theta$, the negative log-likelihood of θ is

$$\mathcal{L}_{\text{PR}}(\theta; \mathcal{T}) = \sum_{i=1}^{N} -y_i \cdot \mu(x_i) + \exp(\mu(x_i)).$$

One can use Poisson regression to model event times $\mathcal{T}$ as per the previous section. The key is to discretise the time interval $[0, T]$ into a number of fixed intervals $(j \cdot \Delta, (j+1) \cdot \Delta]$ for resolution $\Delta \in (0, 1)$. The events can then be viewed as a sample $\{(x_i, y_i)\}$, where x_i simply denotes the interval under consideration, and y_i the number of events that occur in that interval. Under Poisson regression, a nonparametric estimate of $\mu(x)$ would simply be the number of events in the interval corresponding to x. A parametric estimate of $\mu(x)$ can further be viewed as an approximation to the IPP objective with piecewise constant intensity $\lambda = \exp \circ \mu$ [2].

Other demand models. Time series models such as the Autoregressive Integrated Moving Average (ARIMA) have been used to predict public transport demand [17]; these have the advantage of simplicity, but require tuning a window size, and can be sensitive to the choice of this parameter.

3 IPPS WITH NEURAL INTENSITY

The nature of public transportation demand is that it is time-varying. More specifically, even a cursory visual inspection of passenger arrival data over time (see Figure 1) reveals several regularities, or *stylised facts* of passenger demand. For instance, we can see clearly defined trends wherein arrivals peak around the morning and evening, and subsequently decay. The precise nature of these trends will of course vary amongst different bus stops, being in turn determined by e.g. the schedule of buses serving that stop.

The challenge in applying IPPs to capture such characteristics is in choosing a sufficiently flexible family of λ. Typical choices for λ involve (exponentials of) polynomials or Fourier series [8, pg. 34]. While the latter can capture periodic, nonlinear trends, they result in a highly non-convex likelihood; further, it is natural to ask whether there are other means of capturing nonlinearity.

To this end, we formulate a flexible family of intensities in the form of a *one-layer neural network*:

$$\lambda(t; \theta) = g\left(a + \sum_{k=1}^{K} b_k \cdot f(c_k \cdot t + d_k)\right), \tag{2}$$

where $g(\cdot), f(\cdot)$ are non-negative activation functions, and K is some number of *hidden units*. The parameters to estimate are $\theta = \{a, (b_k, c_k, d_k)_{k=1}^K\}$; this estimation may be done by minimising Equation 1. Intuitively, each term $f(c_k \cdot t + d_k)$ may be seen as a (learned) hidden representation of the input time. By choosing K sufficiently large, we can model complex non-linear demand.

Recall that passenger demand has a strong cyclical component. To encode this feature into our neural network, there are two choices. The first is to use a periodic activation function, such as $f(z) = \sin z$. The second is to encode the periodicity explicitly into the model, by working with the intensity $\tilde{\lambda}(t) = \lambda(\text{REM}(t))$, where $\text{REM}(t)$ computes the remainder of the time-stamp with respect to some fixed offset, so that e.g. each day is assumed to have the same intensity. We employed the latter as it was found to produce good results, possibly owing to the mitigation of non-convexity.

Special cases. Equation 2 captures as special cases several existing approaches. First, the HPP model is recovered by letting $b_k = c_k = d_k = 0$, so that $\lambda(t)$ is a constant. Second, for identity activation and $f(\cdot)$ a suitable indicator function, we get a $\lambda(t)$ that is piecewise constant on intervals. In fact, the result is simply Poisson regression on the resulting intervals. Third, when $g(y) = \exp(y)$ and $f(z) = \sin(z)$, we obtain the exponentiated Fourier series model.

A basis function view. We can view Equation 2 more generally as an instance of the intensity family

$$\lambda(t;\theta) = g\left(a + \sum_{k=1}^{K} b_k \cdot \phi_k(t)\right), \tag{3}$$

where ϕ_k is a *basis function*. In this view, the use of a neural network can be contrasted to the use of kernel methods to model $\lambda(t)$, as per the recent work of Flaxman et al. [9]: a kernel method can be seen as using *fixed* basis functions, while a neural network adaptively *learns* the basis functions from the data. An advantage of the neural network approach, beyond increased modelling power, is scalability: the bane of naïve kernel methods is their quadratic time complexity. The random Fourier features approximation to kernel methods [15] can be seen as a neural network with sin and cos activations, where the weights from the input to hidden layer are clamped to suitably distributed random values.

For $\phi_k(t) = \exp\left(-b_k \cdot (t - c_k)^2\right)$, Equation 3 also captures the inhomogeneous model of Allende-Bustamante et al. [1] – indeed, for this choice, we have a radial basis function network [3]. However, they proposed to use EM to fit the model, while we directly optimise the likelihood. If the basis functions are fixed, and one uses $g(y) = \exp(y)$, this has the advantage of involving a convex optimisation.

Existing work. The idea of using neural network intensities for general point processes is not new; several authors [13, 19] have recently proposed the use of recurrent neural networks in conjunction with self-exciting point processes. Inference in these more complex models is much more involved, however. Further, self-excitation is not clearly suited to model cyclical trends.

4 EXPERIMENTS

We now validate our approach on a (proprietary) real-world dataset.

4.1 Description of data

We use real-world bus stop arrival data from a large Australian metropolitan area. The data contains the precise times that passengers alight[1] a bus at a particular stop (identified by a unique ID) over the period of four weeks in 2017. We employ the data for the 100 bus stops that serve the majority of passengers. We exclude weekends from the analysis, as these have qualitatively different characteristics than weekdays.

4.2 Methods compared

Our baseline method is nonparametric Poisson regression, as summarised in §2. Concretely, suppose we wish to estimate the number of events occurring in some time interval. Suppose we partition each day into intervals $I_1, \ldots, I_K$, e.g. with a granularity of 5 minutes. Let us further assume that across days, the number of arrivals in an interval I_j is governed by a Poisson random variable with mean e^{μ_j}. Then, we can estimate the mean number of arrivals in each interval, averaging the arrivals over days. Such a model has been previously used to predict arrivals at bus stops [11].

We compare this method against the Poisson process in its homogeneous (**HPP**), and inhomogeneous versions (**IPP**) using different activation functions: sigmoid $f(z) = (1 + e^{-z})^{-1}$ (suffix **Sig**), and inverse square $f(z) = (1 + z^2)^{-1}$ (suffix **InvSq**). For all IPP methods, we set $g(\cdot)$ to be the identity activation and the number of hidden units $K = 2$, as this was found to give reasonable performance. We additionally experimented with fixed basis functions (per Equation 3), but omit details owing to space constraints.

[1] Strictly, we do not know exactly when passengers *arrive* to the bus stop; their alightment time is an upper bound on this arrival time.

4.3 Evaluation protocol

We compare various methods by means of how well they can predict future passenger arrivals at a bus stop. We use two weeks' worth of passenger arrival data to train the various models. These are then used to make predictions of arrivals in the next two weeks. Predictions are made at a 5 minute granularity; recall that for an IPP, we predict the number of arrivals in a time interval $(s, t]$ via $\int_s^t \lambda(x)\,dx$. This quantity is estimated by numerical quadrature.

For each prediction, we compute two error metrics: the mean absolute error (**MAE**), and the mean square error (**MSE**). We further compute the % improvement in these metrics of each method over Poisson regression. Recall that we make predictions for 100 distinct bus stops. For each bus stop, we thus get MAE and MSE improvement scores. We summarise the mean, and the 5% and 95% quantile of these scores over all bus stops.

4.4 Results and discussion

We summarise our results by means of answers to central questions.

What improvement is offered over Poisson regression? Table 1 confirms that all methods significantly improve upon Poisson regression, which is reassuring. Of note is that the benefit is consistent over all bus stops, as the 5% quantile improvement is still significant, being on the order of 10%. Further, our IPP methods improve slightly but consistently over the HPP method, showing the benefit of capturing the nonlinear trend in demand.

Which bus stops see most improvement? Figure 2 shows that in general, for bus stops that have only a few passengers, the improvement offered by IPP over Poisson regression is significant. This is intuitive, as when there are many passengers, we expect even the 5 minute demand to be reasonably reliable to estimate via Poisson regression. By contrast, with fewer passengers, data sparsity hampers the effectiveness of this method, and calls for parametric methods such as IPP.

What is the impact of time granularity? To further verify our intuition from Figure 1 that Poisson regression is not suitable at finer time granularities, we compare its performance to that of the best performing IPP (with sigmoid activation) as the time granularity is varied. We see from Figure 3 that at granularities above 15 minutes, Poisson regression is comfortably superior in terms of MAE; this is unsurprising, as at such granularity there is sufficient data to reliably estimate the mean number of arrivals. However, below this mark, the IPP is clearly superior.

How fast is estimation? An incontrovertible advantage of Poisson regression is its speed – on average over all bus stops, we find that model estimation takes a mere **0.05 seconds**. By contrast, our IPP models take between **12 seconds** to **25 seconds** to estimate. While significantly slower in relative terms, in absolute terms this is still acceptable, and given the vast performance improvement resulting from IPP, we believe the resulting tradeoff will be appealing in practical urban mobility applications.

(a) MAE % improvement over Poisson regression

Method	Mean	5% quantile	95% quantile
IPP-Sig	50.9	16.5	78.1
IPP-InvSq	49.7	15.8	77.7
HPP	49.5	15.3	77.5

(b) MSE % improvement over Poisson regression

Method	Mean	5% quantile	95% quantile
IPP-Sig	62.9	33.4	85.4
IPP-InvSq	62.4	32.6	85.4
HPP	62.1	29.0	85.3

Table 1: Comparison of various methods to predict bus stop demand at 5 minute time windows.

Figure 2: IPP performance improvement (mean and 95% C.I.) as a function of passenger count at a bus stop.

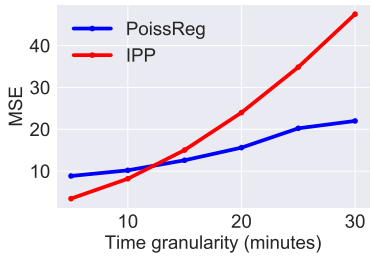


Figure 3: Comparison of Poisson regression and IPP (mean and 95% C.I.) at varying time granularities.

5 CONCLUSION

We have shown how to forecast short term user demand for public transport via a inhomogeneous Poisson process, using a neural network as the underlying intensity. Compared to approaches such as Poisson regression or homogeneous Poisson processes, this can deal with noise inherent in fine time granularities, and can capture the complex non-stationary trend of such demand. Experiments on real-world bus arrival data from a large metropolitan area in Australia validate our approach. In particular, we show benefits over standard Poisson regression in performing *short term* predictions.

There are several directions for future work, such as employing a stochastic intensity (known as a Cox process) to further capture uncertainty, systematically exploring a suitable class of activations for public transport data, and investigating alternate means of estimating parameters for the non-convex likelihood function under a neural network. More broadly, we hope to further study the use of neural networks to model complex time series data.

ACKNOWLEDGMENTS

We thank Transport for NSW for their support in this research.

REFERENCES

[1] J. L. Allende-Bustamante, J. I. Yuz, and J. Rodó. 2016. Application of point processes estimation to a Metro system. In *2016 Australian Control Conference (AuCC)*. 232–237.
[2] Mark Berman and T. Rolf Turner. 1992. Approximating Point Process Likelihoods with GLIM. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 41, 1 (1992), 31–38.
[3] Christopher M. Bishop. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., New York, NY, USA.
[4] BITRE. 2013. Public transport use in Australia's capital cities, Modelling and forecasting. https://bitre.gov.au/publications/2013/files/report_129.pdf. (2013).
[5] D. J. Daley and D. Vere-Jones. 2008. *An introduction to the theory of point processes. Vol. II* (second ed.).
[6] Philip J. Davis and Philip Rabinowitz. 1984. *Methods of Numerical Integration* (second ed.). Academic Press.
[7] Sami Demiroluk and Kaan Ozbay. 2015. A Doubly Stochastic Point Process Model for Modeling Crashes along a Corridor. In *Transportation Research Board 94th Annual Meeting*.
[8] Ludwik Drazek. 2013. Intensity Estimation for Poisson Processes. (2013). MSc thesis, University of Leeds, Department of Statistics.
[9] Seth Flaxman, Yee Whye Teh, and Dino Sejdinovic. 2017. Poisson intensity estimation with reproducing kernels. In *International Conference on Artificial Intelligence and Statistics*, Vol. 54. 270–279.
[10] Liping Fu and Xuhui Yang. 2002. Design and Implementation of Bus-Holding Control Strategies with Real-Time Information. *Transportation Research Record: Journal of the Transportation Research Board* 1791 (2002), 6–12.
[11] G. Liu and S. C. Wirasinghe. 2001. A simulation model of reliable schedule design for a fixed transit route. *Journal of Advanced Transportation* 35, 2 (2001).
[12] P. McCullagh and J.A. Nelder. 1989. *Generalized Linear Models* (second ed.). Chapman & Hall.
[13] Hongyuan Mei and Jason Eisner. 2016. The Neural Hawkes Process: A Neurally Self-Modulating Multivariate Point Process. *CoRR* abs/1612.09328 (2016).
[14] Luis Moreira-Matias, Joao Gama, Michel Ferreira, Joao Mendes-Moreira, and Luis Damas. 2013. Predicting Taxi Passenger Demand Using Streaming Data. *Trans. Intell. Transport. Sys.* 14, 3 (Sept. 2013), 1393–1402.
[15] Ali Rahimi and Benjamin Recht. 2007. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems (NIPS)*.
[16] Tomer Toledo, Oded Cats, Wilco Burghout, and Haris N. Koutsopoulos. 2010. Mesoscopic simulation for transit operations. *Transportation Research Part C: Emerging Technologies* 18, 6 (2010), 896 – 908.
[17] C. H. Tsai, C. Mulley, and G. Clifton. 2013. Forecasting public transport demand for the Sydney Greater Metropolitan Area: a comparison of univariate and multivariate methods. In *2013 Australasian Transport Research Forum*.
[18] UN-Habitat. 2013. Planning and Design for Sustainable Urban Mobility: Global Report on Human Settlements 2013. (2013).
[19] Shuai Xiao, Junchi Yan, Xiaokang Yang, Hongyuan Zha, and Stephen M. Chu. 2017. Modeling the Intensity Function of Point Process Via Recurrent Neural Networks. In *AAAI Conference on Artificial Intelligence*. 1597–1603.
[20] Zhengyi Zhou, David S. Matteson, Dawn B. Woodard, Shane G. Henderson, and Athanasios C. Micheas. 2015. A Spatio-Temporal Point Process Model for Ambulance Demand. *J. Amer. Statist. Assoc.* 110, 509 (2015), 6–15.