# Journey data based arrival forecasting for bicycle hire schemes

Marcel C. Guenther and Jeremy T. Bradley

Imperial College London,
180 Queen's Gate,
SW7 2RH, London, United Kingdom
{mcg05,jb}@doc.ic.ac.uk

**Disclaimer** This document contains corrections for an ealier paper published with Springer for ASMTA'13 under the same title. Due to a bug in our R source code, none of the LRA forecasts presented in the original publication were true out-of-sample forecasts, thus making the IPCTMC forecasts look worse than they really are. This erratum addresses any mistakes that this bug has caused. While the background and method sections 1-4 are mostly the same, there are significant differences in sections 5 and 6.

**Abstract.** The global emergence of city bicycle hire schemes has recently received a lot of attention in the performance and modelling research community. A particularly important challenge is the accurate forecast of future bicycle migration trends, as these assist service providers to ensure availability of bicycles and parking spaces at docking stations, which is vital to match customer expectations. This study looks at how historical information about individual journeys could be used to improve interval arrival forecasts for small groups of docking stations. Specifically, we compare the performance of small area arrival predictions for two types of models, a mean-field analysable time-inhomogeneous population CTMC model (IPCTMC) and a multiple linear regression model with ARIMA error (LRA). The models are validated using historical rush hour journey data from the London Barclays Cycle Hire scheme, which is used to train the models and to test their prediction accuracy.

**Keywords:** IPCTMC, Time-inhomogeneous population models, Mean-field analysis, Multiple linear regression with ARIMA error, Bicycle sharing schemes, Spatial modelling

## 1 Introduction

On the $30^{\text{th}}$ of July 2010, the Barclays Cycle Hire scheme launched in London, after similar schemes had proved to be popular in metropoles such as Paris, Barcelona and Vienna. Bike hire schemes generally feature a number of docking stations where bikes can either be rented or returned. Stations are installed all

over the city so that the maximum distance between neighbouring stations is at most 500 metres. Moreover, stations vary in the numbers of parking slots they provide, the largest stations being close to transport hubs. The schemes are aimed to provide a cost-effective, green solution to the last-mile transport problem in large cities for both tourists and commuters. While tourists can purchase day memberships for the hire scheme, commuters can also opt for a discounted annual membership. Naturally, the growing popularity of cycle hire as well as the abundance of publicly available data from various operational hire schemes has attracted interest in the performance research community.

Much like traditional transport providers, cycle hire operators face classical problems such as infrastructure planning [1], pricing [2] and policy improvement [3–5]. However, our focus in this paper is related to the challenge of being able to forecast the number of available bikes and parking slots at different docking stations [6–8]. Being able to make such forecasts is of vital interest to both operators and customers. With the growing availability of mobile internet access, users of transport systems nowadays expect the availability of real-time transport information. Hence, multi-modal end-to-end routing applications that consider bicycle hire as a possible mode of transport, need to be able to accurately forecast the availability of bikes and parking spaces at suggested origin and destination docking stations [8, 9]. Moreover, operators further require good future estimates as to when stations become empty or full in order to redistribute bikes and antagonise such trends [5, 10]. Aside from purely quantitative performance evaluation, efforts have also been made to visualise migration trends of bicycle schemes. In particular [11] and [12, 13] show that a lot can be learnt about the dynamics of bike sharing systems by using an appropriate visualisation.

This paper has two aims. Firstly we investigate whether journey information will enable us to train models that provide better arrival forecasts than models that solely use departure and arrival data without any information that relates a single departure to a single arrival. This is interesting as the latter data was used in [7, 8]. Moreover, we present a novel time-inhomogeneous Population CTMC (IPCTMC) model that can provide forecasts for the number of interval arrivals for small sets of docking stations at rush hour[1]. Futhermore, we compare the accuracy of the model with a traditional multiple linear regression with ARIMA error (LRA) approach that is more akin to the time series techniques used in [7, 8]. Our results indicate that under realistic assumptions both models produce similar results. However, when provided with perfect information about future journey the IPCTMC model outperforms the regression approach. Irrespective of the modelling technique, our results further indicate that for predictions of 30 minutes or less, models trained on historical origin to destination journey data produce better forecasts than models trained without journey data. This motivates further study on how to combine journey data with station occupancy time series data in order to improve forecasts of the latter. Unfortunately, since

[1] By interval arrivals we mean the total number of future arrivals for a set of neighbouring docking stations over a fixed interval of time.

we do not currently possess matching journey and station occupancy time series data sets, this research is beyond the scope of this paper.

The remainder of this paper is organised as follows; In Section 1.1 we briefly review the literature on bicycle hire schemes. Section 2 subsequently introduces the reader to Population Continuous-Time Markov Chains and presents a time-inhomogeneous extension. In Section 3 and Section 4 we develop an IPCTMC as well as an LRA model for interval arrival forecasts and compare their forecast quality in Section 5. Section 5.1 discusses other aspects of the two different modelling approaches.

### 1.1 Related work

Froehlich *et al.* [6] were the first to propose a station occupancy forecast model based on historical time series data describing the number of bikes docked at a particular station. To make predictions about individual stations, they trained a Bayesian Network model for each station, using time, prediction window and the current proportion of occupied parking slots as regression parameters. Given the current state of the system, their model forecasts future station capacity as either $0-20\%$ full, $20-40\%$ full and so on. However, judging from their error analysis, their model only marginally outperforms a simple historical trend predictor.

Kaltenbrunner *et al.* [7] suggested an ARMA model, which was trained on similar time series data. Their most important contribution was the observation that the prediction error can be vastly reduced by incorporating information about the occupancy of behaviourally similar neighbouring stations. Although their ARMA model produces significantly better results than comparable historical trend models, their decision to use an ARMA process to predict a highly time-inhomogeneous process is possibly sub-optimal. Furthermore, both their model fitting and their error analysis is performed on smoothed time series data, making it harder to judge how accurate their model truly is.

Yoon *et al.* [8] recently addressed some of these shortcomings. By fitting an ARIMA process they were able to capture the inhomogeneous nature of bike hires better. Furthermore, instead of only looking at neighbouring stations for extra exogenous variables for their regression model, the authors computed a detailed time-lag dependent cross-correlation metric for all pairs of stations. This approach improves the long-term forecast quality, as it captures both positively correlated neighbouring stations that experience similar traffic as well as negatively correlated stations that have opposite migration trends. Interestingly, while their ARIMA model provides the best forecast in a benchmark carried out on data from the Dublin cycle hire scheme, it does not appear to outperform Kaltenbrunner's ARMA model by much.

Aside from research on out-of-sample forecasts of future station occupancy, other notable contributions are the cluster analysis provided in [14] and the departure forecasting model discussed in [15]. Furthermore Fricker *et al.* [5] use mean-field analysis to investigate how fleet size and customer behaviour influence the station occupancy problem.

## 2 PCTMCs

Population Continuous-Time Markov Chains (PCTMCs) consist of a finite set of populations $S$, $n = |S|$ and a set $E$ of transition classes [16]. States are represented as an integer vector $\boldsymbol{P}(t) = (P_1(t), \ldots, P_n(t)) \in \mathbb{Z}^n$, with the $i^{th}$ component being the current population level of species $S_i \in S$ at time $t$. A transition class $(r_e, \boldsymbol{c}_e) \in E$ for an event $e$ describes a transition with negatively exponentially distributed delay $D$ at rate $r_e : \mathbb{Z}^n \to \mathbb{R}$ which changes the population vector $\boldsymbol{P}(t+D)$ to $\boldsymbol{P}(t) + \boldsymbol{c}_e$. The analogue to PCTMCs in the systems biology literature are Chemical Reaction Systems, were $\boldsymbol{P}(t)$ describes a molecule count vector and transition classes represent chemical reactions between the molecules with $r_e$ being the reaction rate function and $\boldsymbol{c}_e$ the stoichiometric vector for a specific reaction. For notational convenience we write an event/reaction $e$ as

$$\underbrace{S_* + \cdots + S_*}_{\text{in}} \to \underbrace{S_* + \cdots + S_*}_{\text{out}} \qquad \text{at } r_e(\boldsymbol{P}(t)) \tag{1}$$

where $S_* \in S$ represent different species that are involved in the event. The corresponding change vector is $\boldsymbol{c}_e = (s_1^{\text{out}} - s_1^{\text{in}}, \ldots, s_n^{\text{out}} - s_n^{\text{in}}) \in \mathbb{Z}^n$ where $s_i^{\text{in}}$ represents the number of occurrences of a particular species $S_i \in S$ on the left hand side of the event and $s_i^{\text{out}}$ its number of occurrences on the right hand side. The event rate is

$$\begin{cases} r_e(\boldsymbol{P}(t)) & \text{if } P_i(t) \geq s_i^{\text{in}} \ \forall i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

When used to describe spatially distributed populations, we denote a species $S$ at location $l$ at time $t$ as $S@l(t)$ [17].

An important feature of PCTMC models is that approximations to the evolution of population moments of the underlying stochastic process can be represented by the following system of ODEs [18]

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[T(\boldsymbol{P}(t))] = \sum_{e \in E} \mathbb{E}[(T(\boldsymbol{P}(t) + \boldsymbol{c}_e) - T(\boldsymbol{P}(t)))r_e(\boldsymbol{P}(t))] \tag{3}$$

To obtain the ODE describing the evolution of the mean of a population, all we need to do is to substitute $T(\boldsymbol{P}(t)) = P_i(t)$ in the above equation, where $P_i(t)$ is the random variable representing the population count of species $S_i$ at time $t$. In the literature the resulting ODEs are often referred to as mean-field approximations [18].

### 2.1 Time-inhomogeneous PCTMCs

While the PCTMC formalism has been applied to problems in many application areas, it would be rather inaccurate to describe the model presented in Section 3 using a time-homogeneous CTMC process, since many parameters, such

as departure rates as well as destination of journeys, vary with time. Hence, we present a time-inhomogeneous extension to the PCTMC formalism, which we term IPCTMC, that is going to be released in a future version of the Grouped PEPA Analyser (GPA) [19]. To our knowledge this is the only paper aside from [20], which applies mean-field analysis to IPCTMC models. In IPTMCs we allow deterministic rate and population changes that occur at deterministic times. This implies that any reaction rate $r_e(\boldsymbol{P}(t))$ (cf. Section 2) is now time dependent, i.e. $r_e(\boldsymbol{P}(t), t)$ becomes

$$\begin{cases} r_e(\boldsymbol{P}(t), t_1) & \text{if } P_i(t) \geq s_i^{\text{in}} \ \forall i = 1, \ldots, n \wedge t < t_1 \\ r_e(\boldsymbol{P}(t), t_2) & \text{if } P_i(t) \geq s_i^{\text{in}} \ \forall i = 1, \ldots, n \wedge t_1 \leq t < t_2 \\ \ldots & \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $t_1, t_2, \ldots$ are deterministic times at which reaction rate changes occur. Similarly, we allow deterministically timed events that result in an affine transformation of the population vector $\boldsymbol{P}(t)$. In the following we informally assert that if a deterministic population change occurs at time $t_d$ then no population changes occur due to random PCTMC events between $t_d - \delta t$ and $t_d$. Should no such interval exist, then we assume that the deterministic event is triggered immediately after the random event. Let $\mathcal{D}$ denote the set of all deterministic events, s.t. $(t_d, \mathbf{M}) \in \mathcal{D}$, where

$$\mathbf{M}_{n,n+1} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 & d_1 \\ 0 & \lambda_2 & \cdots & 0 & d_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_n & d_n \end{pmatrix} \tag{5}$$

and the updated population count vector becomes

$$\boldsymbol{P}(t_d) = \mathbf{M} \left( \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \end{pmatrix} \boldsymbol{P}(t_d - \delta t) + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \right) \tag{6}$$

As an example we can now describe the reset of population $p_1$ to $d_1$ and the population jump of population $p_1$ by $d_1$ individuals as

$$\mathbf{Reset}_{n,n+1} = \begin{pmatrix} 0 & 0 & \cdots & 0 & d_1 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \qquad \mathbf{Jump}_{n,n+1} = \begin{pmatrix} 1 & 0 & \cdots & 0 & d_1 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \tag{7}$$

We can also vary the $\lambda_i$, for instance to double a population. In both IPCTMC simulation runs and mean-field analysis, rate changes that are part of a deterministic event, simply involve an update of that rate as the event occurs. Moreover,

when simulating an IPCTMC model, any population change can be immediately applied to the population count vector using Eq. (6). However, mean-field ODEs analysis of IPCTMCs is more complicated with regards to the population vector updates, since we keep track of population moments rather than the values of actual random variables. As a consequence we need to expand the affine transformation inside the expectation expressions. For instance if we have two populations $X$ and $Y$ and that the following deterministic population change occurs at $t_d$; $X(t_d) = \lambda_x X(t_d - \delta t) + d_1$ and $Y(t_d) = Y(t_d - \delta t)$ then

$$\begin{aligned}\mathbb{E}[X(t_d)Y(t_d)] &= \mathbb{E}[(\lambda_x X(t_d - \delta t) + d_1)Y(t_d - \delta t)] \\ &= \lambda_x \mathbb{E}[X(t_d - \delta t)Y(t_d - \delta t)] + d_1 \mathbb{E}[Y(t_d - \delta t)] \end{aligned} \tag{8}$$

and similarly for $\mathbb{E}[X(t_d)^2]$ or any other moments. Fortunately, so long as the transformation of the population vector remains affine, we do not have to use moment closures [16] in order to compute the new values for the moments at time $t_d$. More complex variants of mean-field analysable IPCTMC models, such as ones with non-linear population vector transformations or population transformations that are subject to boundary conditions, are also possible. However, these require further treatment, which are beyond the scope of this work.

## 3 An IPCTMC interval arrival forecasting model

Figure 1 shows the species and parameters of our IPCTMC forecasting model. Our model is estimating the number of future arrivals for a set of neighbouring stations denoted in a small area $A$. In Section 5 we will then look at forecasts for different areas $A$ at different times of the day, e.g. for areas depicted in Figure 2. Since we model arrivals, we are particularly interested in areas around stations that are in danger of running out of available parking spaces.

All states shown in Figure 1 correspond to species in the formal definition of the underlying IPCTMC. Each journey is assumed to start in one of the states $DeparturesCl@1, \ldots, DeparturesCl@n$. Those which end in $A$, will finish in $Arrivals@A$ after experiencing a delay represented by $PhCl@i$. All other journeys are assumed to end $Elsewhere$. Naturally, for the purpose of forecasting, we are most interested in the population of agents that transit to state $Arrivals@A$ during the forecast interval $[t_0, t_{fcast}]$. $DeparturesCl@1, \ldots, DeparturesCl@n$ encapsulate all stations that are starting points for journeys that terminate in $A$, including stations that lie inside $A$. However, while the $Arrivals@A$ state captures arrivals at stations within a specific geographical area, departure cluster membership of stations is based on similarity in journey time distribution. In other words for each station within a cluster, the distribution of time it takes a journey that starts from this station and ends at any station in $A$ has to be similar. Since journey durations are generally not exponentially distributed, we use additional states (cf. $PhCl@i$) to represent the clusters' characteristic journey time distributions as phase-type approximations.

Having described the states of the IPCTMC model, we now need to explain its parameters. In the following we assume that we can initialise our model us-

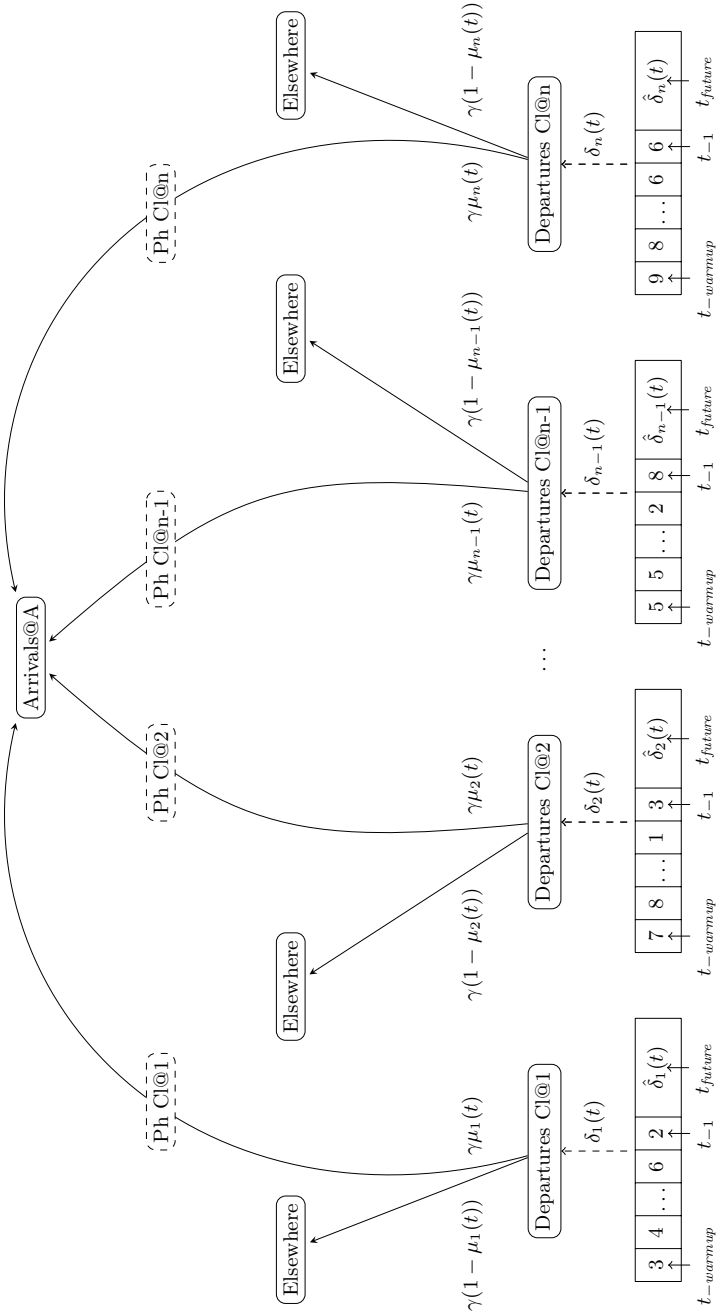



**Fig. 1:** The IPCTMC arrival forecasting model.

**(a)** City of London (Stn175) **(b)** Waterloo (Stn361)

**Fig. 2:** Two examples of areas that lack parking spaces during rush hour (cf. Table 1). Each circle represents a docking station, the larger the circle the more docks it has.

ing historical information for the interval $[t_{-warmup}, \ldots, t_0]$ and that we make a forecast for the number of agents that reach state $Arrivals@A$ during the interval $[t_0, t_{fcast}]$. $t_{-warmup}$ is chosen to be large enough such that it represents the 99% percentile of each cluster's joint journey time distribution, i.e. departures before $t_{-warmup}$ are unlikely to affect our forecast. $\delta_i(t)$ is the time-inhomogeneous component that describes departures from cluster $i$. During the interval $[t_{-warmup}, t_0]$, we know all journeys that have departed from any station in the cluster and therefore we can use population jumps to increase the departure population of a given cluster. This is done every minute, i.e. at $t_{-warmup}$ we increase the population of $DeparturesCl@i$ by the number of departures that were observed for this cluster during the interval $[t_{-warmup}, t_{-warmup+1}]$. To ensure that the additional journeys actually leave state $DeparturesCl@i$ by $t_{-warmup+1}$, we choose a sufficiently large exponential rate $\gamma$. During the actual forecast period $[t_0, t_{fcast}]$, we do no longer use population jumps, but instead treat $\hat{\delta}_i(t)$ as exponential rates at which new journeys start from cluster $i$. The $\mu_i(t)$ parameters reflect the proportion of departures from cluster $i$ that is heading for stations in area $A$.

### 3.1 IPCTMC model parameter estimation

To create clusters we first chose a forecast period and assumed journey times to be iid. Weekday rush hour journeys are particularly suitable for this, since commuters, unlike tourists, are less likely to cycle in groups. To train a model for a particular area $A$, we first measured the journey time samples for all training journeys that ended in $A$. Subsequently we grouped these observations by their start station and computed the 10% and the 90% percentiles of the resulting station journey time distributions. After that, a $k$-means clustering algorithm was

run on the coordinates. To ensure that stations with a larger number of outgoing journeys have a larger impact than less frequently used ones, we created $n$ identical $(10\%, 90\%)$ cluster points for a station with $n$ observations in the training data set. We varied the number of clusters $k$ to get a decent trade-off between cluster compactness and clusters size. Although more clusters should generally produce convoluted cluster journey time distributions that are more similar to the distributions of the individual stations in the clusters, too many clusters generate large state-spaces and make cluster parameter estimation harder. Having computed the clusters we fitted a Hyper-Erlang approximation for each cluster journey time distribution using HyperStar [21]. We discuss the impact of the number of clusters on the forecast accuracy in Section 5. Using between 10 to 100 PCTMC species for each $PhCl@i$, we generally obtained low relative errors of less than 5% up to the $4^{\text{th}}$ uncentred distribution moment. In addition to generating the states and the transitions in the IPCTMC model using the phase-type fits, we also generated the time series data describing the minutely departures from all clusters for all training and test dates. $\hat{\mu}_i(t)$ estimates were computed using 5-minute window daily averages that were then further averaged for identical times on different days, thus producing a single estimate for each time point in the observed time period, e.g. morning or afternoon rush hour. Moreover, for all minutely cluster departures we also recorded the precise $\mu_i(t)$ value, and assumed perfect knowledge of journey destinations up to $t_0$ when computing a forecast. Naturally, this is an idealised assumption, however, as Come *et al.* note in [14], the majority of rush hour journeys is made by subscribed cyclists, which would allow service providers to accurately predict journey destinations from their history, especially when we consider the destination to be an area rather than a specific station. Future $\hat{\delta}_i(t)$ rates are estimated naively, i.e. we assume all future $\hat{\delta}_i(t)$ are the same as $\delta_i(t_{-1})$. Naturally, in practice we would use a more elaborate method to forecast future $\hat{\delta}_i(t)$, potentially using more complex models such as those discussed in [15], however a the comparison with the LRA model this is not necessary. To gain insight into how improved departure forecasts affect area arrival forecast accuracy, we can simply compare forecasts made using naively predicted future cluster departures with forecasts in which we assume future departures and $\mu_i(t)$ to be known.

## 4 Linear Regression ARIMA error forecasting model

Our motivation for comparing the IPCTMC model to a time series model (cf. [7, 8]), was to investigate whether an inhomogeneous process would yield better results than a time series model fitted to arrival and departure data, which was made stationary by means of seasonal normalising. After various experiments with different time series model classes, we found that a simple multiple linear regression model with ARIMA error (LRA) worked best. Like Yoon *et al.* [8] we decided to choose an observation frequency of 5 minutes for departure and

arrival observations. The model can then be expressed as follows

$$\begin{aligned}\hat{Y}(t+5) = \alpha_{11}X_1(t) + \ldots + \alpha_{1m}X_1(t-5m) + \ldots + \\ \alpha_{n1}X_n(t) + \ldots + \alpha_{nm}X_n(t-5m) + N(t)\end{aligned} \tag{9}$$

where $\hat{Y}(t+5)$ is the forecasted # arrivals for a target area during $[t, t+5]$. To forecast arrivals over longer intervals, e.g. $[t, t+5i]$ with $i > 1$, we simply successively forecast $\hat{Y}(t+5), \ldots, \hat{Y}(t+5i)$ and subsequently sum all individual point forecasts $\sum_{n=1}^{i} \hat{Y}(t+5n)$. $N(t)$ is an ARIMA(p,d,q) process fitted to the regression residues to characterise the noise of the system. The exogenous $X_c(t)$ variables represent the # of departures from cluster $c$ during $[t-5, t]$. If we were to choose $m = 4$ in Eq. (9) then we would regress on departures that occurred during $[t-25, t]$. Any out-of-sample variables such as $X_c(t+5i)$, where $i \geq 1$, are forecasted using ARIMA or naive models, so that our $\hat{Y}(t)$ forecasts remain out-of-sample, too. The fitting procedure for the $X_c(t)$ coefficients depends on whether our LRA model is *journey* or *non-journey* based, cf. Section 5. When the model is journey based we assume that we have information about the destination of historical departures, whereas in non-journey based models we assume that the relationship between departures and arrivals is unknown. When training a non-journey LRA, the $X_c(t)$ variables represent all departures from the cluster during $[t-5, t]$, whereas in the journey information based LRA model, the exogenous variables represent only those departures that head for the target area whose arrivals we aim to predict. While it is straightforward to forecast $X_c(t)$ for non-journey LRA models using an ARIMA model for each departure cluster, it turns out that for journey information based LRA models it is best to use the same ARIMA models to predict future departures and subsequently multiply the point forecasts by the same historical $\hat{\mu}_i(t)$ estimates that we use the IPCTMC model. In practice we found that this gives much better results than fitting ARIMA models to cluster departure time series which already implicitly contain the $\mu_i(t)$ (cf. Section 3) information.

We used the *R forecast* package [22] to estimate the $\alpha_{ci}$ parameters from training data. However, since we had to fit a single LRA model to a number of non-consecutive time series traces for the same time period on different days, e.g. an LRA model for the interval arrivals of a specific set of stations during the afternoon rush hour, we had to use the interleaving technique described in [23][2]. Furthermore, instead of fitting any observed pairs of $Y, \boldsymbol{X}$ directly, we first normalised each observation at time $t$ by its corresponding time dependent sample mean and standard deviation, which were estimated from the training data. Normalisation is an important step since it generates more stationary time series. We also tried other common differencing techniques to transform the input data, but normalising appeared to work best.

[2] While the actual error is assumed to be ARIMA(p,d,q), the interleaving technique requires us to fit a SARIMA$(p \cdot \#days, 0, q \cdot \#days)(0, d, 0)_{\#days}$, where $\#days$ is the number of days in the time series training data set that we use to train our model.

## 5 Model analysis

In this section we compare the forecast accuracy of models described in Sections 3.1 and 4 for different areas $A$ during the morning and the afternoon rush hour, 6-9am and 4-7pm respectively. Each area is a $500m \times 500m$ square (cf. Figure 2) centred around the stations listed in Table 1, which usually run out of bikes during one of the two rush hours (cf. *Rush hour* in Table 1) and out of bikes during the other. We consider an area rather than a single station as we assume that cyclists, who arrive at a full station, are likely to seek alternate parking spaces in the surrounding area so long as these lie within $250m$ of their initial destination. From a user perspective area forecasts should therefore be equally useful as single station forecasts, so long as the radius of the area is small enough.

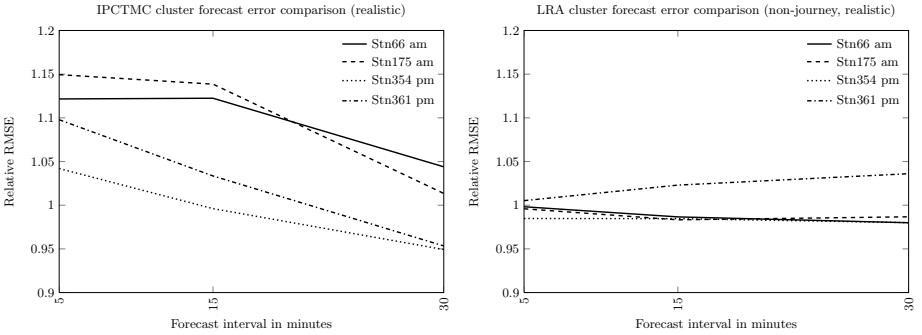
We used May 2012 journey data from the Barclays Cycle Hire scheme to train our models and June 2012 data to test their forecast accuracy. As mentioned earlier, our aim is twofold, namely to investigate the impact of additional origin destination journey information on prediction accuracy and a comparison between the IPCTMC and the LRA model. In the following we distinguish between *journey based* models that use origin destination information and *non-journey* models, which only use information about departures and arrivals without knowing which departure is related to which arrival. Additionally, we discriminate between *realistic* out-of-sample forecasts, which use forecasting techniques for regression variable and interval arrival predictions and *perfect* information models, where all future journeys are assumed to be known at the time a forecast is made. Recall from Section 3.1 that even for realistic forecasts we assume journey destinations of departures up to $t_0$ to be known.

| Ref | Station | Rush hour | #Bike slots (station / ttl area) |
|---|---|---|---|
| Stn66 am | Holborn Circus | morning | 39 / 155 |
| Stn175 am | Appold Street | morning | 26 / 130 |
| Stn354 pm | Northumberland Avenue | afternoon | 36 / 134 |
| Stn361 pm | Waterloo Station 2 | afternoon | 55 / 280 |

**Table 1:** Each $500m \times 500m$ arrival area *Ref* is centred around a single station *Station* that tends to lack free parking slots during *Rush hour*.

We start by looking at different ways to incorporate journey information in our models. One way to achieve this is journey time dependent clustering of departure stations that we described in Section 3.1. Figure 3a shows the Root Mean Squared Error (RMSE) for the interval arrival forecasts of IPCTMC models with one departure cluster divided by the RMSE of the corresponding IPCTMC models with $4 - 6$ clusters for different areas and forecast interval lengths. A similar comparison is made in Figure 3b for the corresponding LRA models. In both cases we compare the accuracy of realistic forecasts. Values $> 1$

imply that the multi-cluster model is more accurate. The IPCTMC graph indicates that more clusters appear to produce significantly better fits for realistic short-term forecasts up to 15 minutes, whereas for longer intervals, extra clusters do not outperform single cluster predictions. For the LRA results on the other hand, single and multi-cluster models seem to give similar levels of performance. However, as Figure 4a shows, IPCTMC models fitted on more departure clusters




**(a)** **(b)**

**Fig. 3:** Comparing realistic forecast error of multi-cluster departure models against single departure cluster models for both IPCTMC and LRA. Values $> 1$ imply that multi-cluster models outperform their single cluster counterparts.

clearly outperform single cluster predictions when perfect information about future departures and journey destinations is available after $t_0$. Again for LRA models (cf. Figure 4b) no such trend is visible, though the single cluster models perform less well in comparison to multi-cluster models than in the realistic forecast scenario. This suggests that the extra regression variables for additional clusters result in a better fit. However, in practice we would prefer single cluster models for they are more parsimonious due to the lower number of exogenous variables (cf. Eq. (9)). In the above examples the IPCTMC model uses journey data for clustering as well as for estimating the $\hat{\mu}_i(t)$ parameters, whereas for the LRA model the only the knowledge infered from journey data is the departure cluster membership of departure stations. To see how LRA models perform when trained using all available journey information we considered a second case where we look at what happens if we train a single cluster LRA model on journey data. To do this, we train the model considering only journeys that are heading to the arrival area which we are monitoring. As before, future LRA departure regression variables are estimated using naive models. The reason we forecast future departures naively rather than using a more accurate ARIMA model is to make the later comparison with the IPCTMC model more fair. In Figure 5 we show the RMSE of non-journey data based, single cluster LRA model forecasts divided by the RMSE of a journey information based single cluster LRA model. Hence values $> 1$ indicate that the journey based LRA model is better than

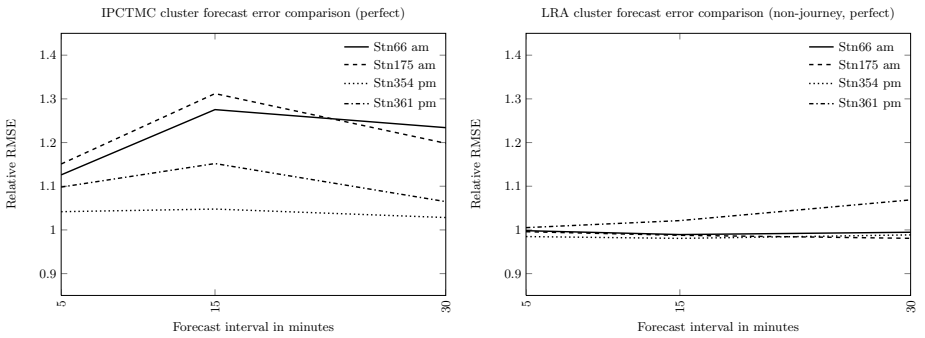



**Fig. 4:** Same error metric as in Figure 3 assuming perfect information. Values $> 1$ imply that multi-cluster models outperform their single cluster counterparts.

the LRA model trained on non-journey data. Under realistic conditions (cf. Figure 5a) as well as in the perfect information scenario the journey based models clearly produce better forecasts. Moreover, Figure 6 indicates that much like the

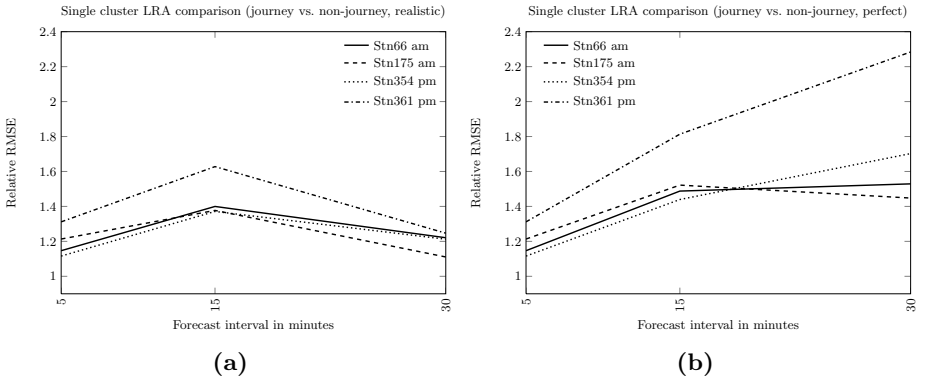



**Fig. 5:** Comparing the accuracy of non-journey single cluster LRA models with that of journey information based single cluster LRA models in realistic and in perfect information scenarios. Values $> 1$ indicate that the journey based model performs better.

for non-journey models compared in Figures 3b, 4b adding additional departure clusters does not improve the LRA model accuracy for 15 to 30 minute forecasts.

Having discussed the impact of journey information on the accuracy of forecasts, we now look at the difference in accuracy between the LRA model and the IPCTMC model by comparing the RMSE of multi-cluster IPCTMC forecasts

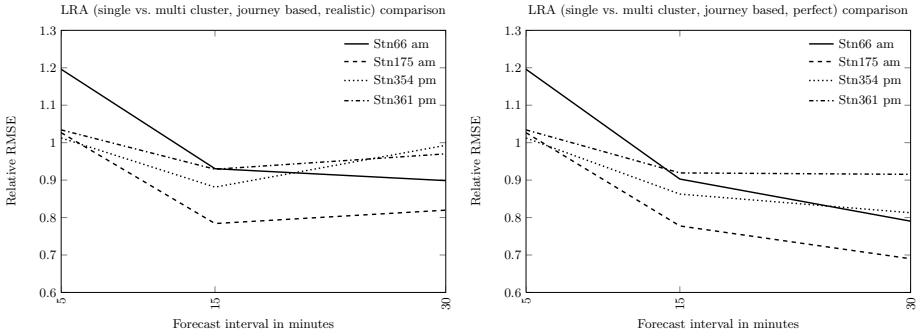



**Fig. 6:** Comparing the accuracy of journey based single cluster LRA models with that of journey information based multi-cluster LRA models in realistic and in perfect information scenarios. Values $> 1$ imply that the multi-cluster LRA is more accurate.

divided by the RMSE of the journey based single cluster LRA forecasts. In the realistic scenario both models use naive forecasts for future cluster departures. Under realistic conditions (cf. Figure 7a), there appears to be no clear winner, although it seems as if the LRA model starts to outperform IPCTMC forecasts as forecast intervals becomes longer. In Figure 7b, where we assume perfect information about future journeys, the IPCTMC model clearly outperforms the LRA model. Hence, the decision as to which model to use might well depend on the accuracy with which we can forecast future journeys, since the IPCTMC model appears to be more sensitive to the quality of the input data.

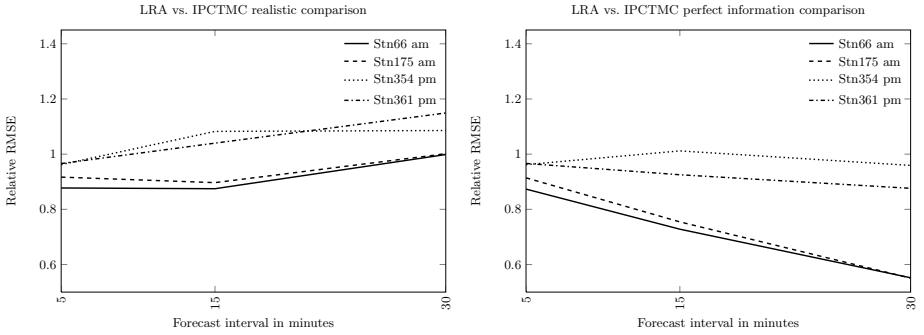


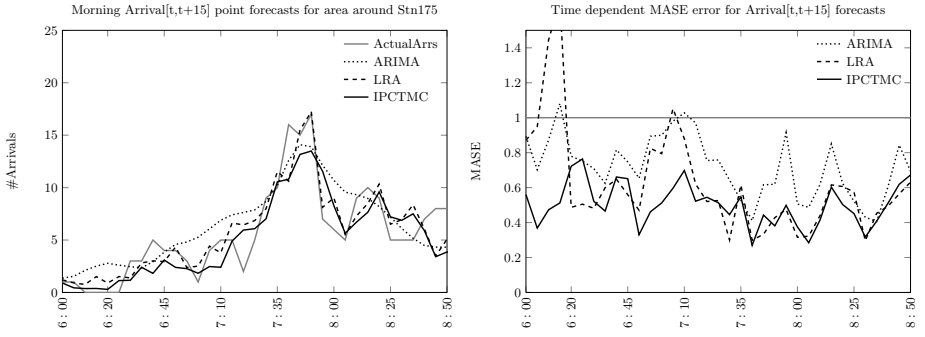

**Fig. 7:** Comparing the accuracy of a multi-cluster IPCTMC models with that of a journey based single cluster LRA models in realistic and in perfect information scenarios. Values $> 1$ indicate that the LRA forecast is better than the IPCTMC forecast.

Figures 8a, 9a show the LRA and the IPCTMC forecast traces for 15/30 minute interval arrival forecasts for areas around Stn175 / Stn361 (cf. Table 1) in the morning / afternoon of the 11/06/2012 & 06/06/2012, respectively. In order to highlight that these multivariate explanatory models produce better forecasts than univariate models, we also fitted an ARIMA model to the normalised arrival time series for comparison. The MASE[3] error illustrated in Figures 8b, 9b visualises time dependent superiority of complex forecasting techniques over the naive method, which assumes that the number of arrivals for the next interval is the same as for the previous one. Unlike [7, 8] we do not show raw RMSE errors, since these vary a lot between different forecast areas. Ultimately, for actual station occupancy forecasts, the percentage of incorrect predictions that would cause customers to wait for available bikes or parking slots would make a good error statistic.

The graphs in Figures 8b, 9b study the time dependent MASE statistic for different times in the morning and the afternoon rush hour for different geographical areas over the entire month of June 2012. These diagrams give another perspective on the forecast quality of the LRA and the IPCTMC models. In particular for the area around Stn361 it can be seen that although the RMSE error of the LRA model depicted in Figure 7a is slightly better than the one of the IPCTMC, it is much harder to see which model performs better when looking at Figure 9b. This illustrates that it can be really hard to decide which model gives better forecasts. Needless to say though that the time-dependent MASE error is a relative statistic that only tells us how our models fare against the error of the naive model at different times. In practice it might even make sense to investigate the time-dependent RMSE error, to see if a time-dependent mixed LRA-IPCTMC model could be build that chooses the model that is likely to produce the better forecast at a given moment in time according to historical performance. Another observation that can be made by comparing Figures 8b, 9b is that under realistic circumstances the gap between the ARIMA model and the LRA and IPCTMC models is much smaller for the 30 minute interval than for the 15 minute one. This is not unexpected since the uncertainty regarding future departures increases as we make the arrival forecast window longer. Hence, as future research it would be interesting to see whether there is a cut-off forecast length at which the univariate ARIMA actually starts to outperform our multivariate models.

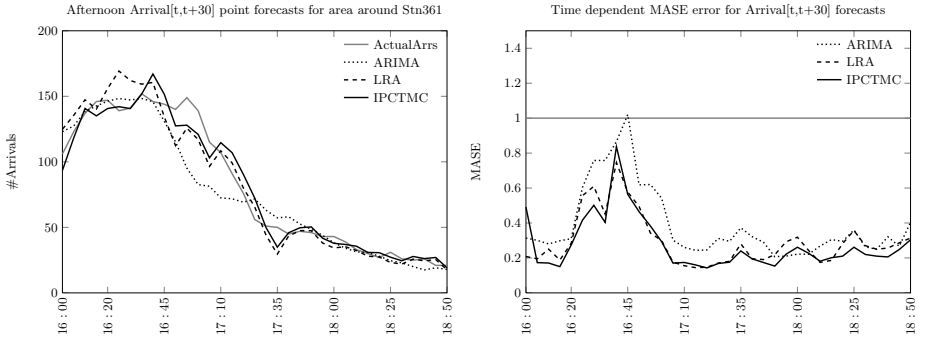
Finally Figures 10a, 10b suggest that some of the IPCTMC errors might be systematic. In the 15-minute forecast (cf. Figure 10a) significant correlations with lag $> 3$ and in the 30-minute forecast (cf. Figure 10b) correlations with lag $> 6$ indicate that there is some potential for correcting the errors as they lie outside the forecast period. Both models seem to have such significant correlations, but whether these can be corrected requires further investigation.

[3] The MASE [22] statistic is the average model forecast error divided by the average naive forecast error and thus provides a good benchmark for any prediction model.




**(a)** Point forecasts 11/06/2012 **(b)** Point forecast MASE (June 2012)

**Fig. 8:** Comparing different morning arrival forecasts for area around Stn175.

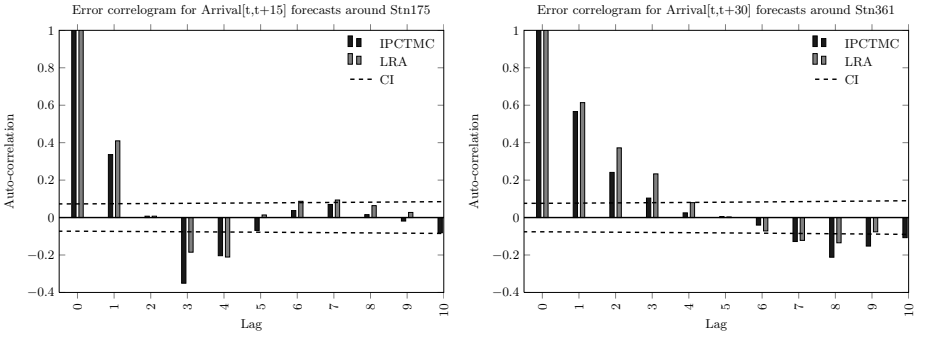



**(a)** Point forecasts 06/06/2012 **(b)** Point forecast MASE (June 2012)

**Fig. 9:** Comparing different afternoon arrival forecasts for area around Stn361.

### 5.1 Model discussion

The examples above indicate that LRA and IPCMTC forecasts both perform well under realistic conditions. It seems, however, that LRA models have a slight edge when forecasting 30 minute intervals, though we certainly require further empirical evidence to support this thesis. While the forecast accuracy is of great importance to many practical applications, we also need to consider other aspects of the models, such as the explanatory value of the model. While our IPCTMC models are fully transparent, some aspects of the LRA model remain somewhat obscure. As for many other time series models the hardest part in the LRA model is to transform the evidently time inhomogeneous data to make it stationary (cf. Section 4). While normalisation based on historical statistics seems an intuitive way of achieving this, we had to try various transformations before we decided that normalisation gave sufficiently good fits for our training data. Also in spite of yielding good results, it is hard to judge whether the resulting series are truly time-homogeneous. Furthermore, while the IPCTMC




**(a)** Stn175 - Forecast (June 2012) **(b)** Stn361 - Forecast (June 2012)

**Fig. 10:** Out-of-sample forecast error correlograms for different areas and forecas methods. Confidence intervals (CI) widen with increasing lag.

model transparently describes the relationship between departures and arrivals, it is harder to interpret the relationship between journey time, origin and destination from LRA parameters in the same manner. This could potentially be a problem if we were to extended the model for other purposes such as model provisioning or policy testing. For example, for future work we would like investigate whether sensors on bicycles can be used to collect enough data to provide sufficiently timely information about different geographical areas, using similar mechanisms as described in [24]. Clearly, in this case IPCTMC models are likely to produce better, easier to understand results than regression models.

Aside from modelling challenges, two other important factors are the time required to train the model and to compute a forecast. Parameter fitting is significantly faster for the IPCTMC model than for the LRA model. Although the fitting procedure described in Section 3.1 sounds work intensive, clustering and phase-type fitting can generally be done quickly. For the purpose of this study we fitted phase-type distributions by hand, but there is no reason why this cannot be automated. The time required to fit the LRA model in R on the other hand depends on the amount of training data available, the number of ARIMA configurations that we are prepared to try and on how large we choose $m$ (cf. Section 4) to be. Yet, in practice LRA fitting can be done within a few minutes.

The subsequent forecast speed for either LRA and IPCTMC models is negligably small, e.g. point forecasts can be computed in less than 10 seconds using either model. Moreover, due to the linear nature of the model, it is also feasible to use simulation in order to analyse the IPCTMC models. However, if we were to add any non-linear rates to our model, for instance in order to investigate gossip like behaviour like in [24], then mean-field analysis would become significantly faster than simulation. In addition to this, non-linear relationships would make regression model fitting harder and also more computationally expensive.

## 6 Conclusions

In this paper we investigated whether the availability of journey data can help to produce more accurate bike migration forecast models and we compared the accuracy of a traditional time series model with those of a strictly time-inhomogeneous PCTMC model. Although our models are not immediately comparable to the station occupancy models proposed by Kaltenbrunner [7] and Yoon [8], we have shown that journey information has the potential to improve such forecasts. In the future we would thus like to investigate, whether our small area arrival predictions can be successfully applied to enhance existing station and area occupancy prediction models

As for the comparison between LRA and IPCTMC, it was surprising to see that both models produce similar results under realistic conditions, since we initially assumed that the time-inhomogeneous nature of the process would make it hard to fit a time series model. Moreover, in addition to being a decent model, the IPCTMC model further retains a much more intuitive representation of the system that is being modelled. Further, we showed that given perfect information regarding future journeys, the IPCTMC model outperforms the LRA model. Thus we can conclude that the passage time information encapsulated in the IPCTMC model is an important feature for this kind of system and difficult to embed in the exogenous variables of a regression model. These findings encourage further research into IPCTMC models, both for forecasting as well as for model based provisioning tasks. In particular the fact that IPCTMC models are also mean-field analysable should become especially advantageous in situations where we encounter non-linear effects such as mass-action kinetics, which cannot easily be approximated by linear models and become expensive to simulate as populations increase. In the future we therefore plan to apply IPCTMC models to investigate phenomena such as geographical crowd data sourcing in a system of mobile agents.

## Acknowledgements

Jeremy Bradley is supported in part by EPSRC on the AMPS project, the Analysis of Massively Parallel Stochastic Systems, ref. EP/G011737/1.

## References

1. A. Stannard and G. Wolfenden, "Putting in Place a New Public Transport System in London," in *18th ITS World Congress*, (Orlando Florida), 2011.
2. P. Le Masurier, F. Shore, and J. Hiett, "Cycle-hire-The New Travel Option for Central London," in *European Transport Conference*, (Glasgow), 2010.
3. J.-R. Lin and T.-H. Yang, "Strategic design of public bicycle sharing systems with service level constraints," *Transportation Research Part E: Logistics and Transportation Review*, vol. 47, pp. 284–294, Mar. 2011.

4. N. Lathia, S. Ahmed, and L. Capra, "Measuring the impact of opening the London shared bicycle scheme to casual users," *Transportation Research Part C: Emerging Technologies*, vol. 22, pp. 88–102, June 2012.
5. C. Fricker and N. Gast, "Incentives and redistribution in bike-sharing systems with stations of finite capacity," Jan. 2012.
6. J. Froehlich, J. Neumann, and N. Oliver, "Sensing and Predicting the Pulse of the City through Shared Bicycling," in *Twenty-First International Joint Conference on Artificial Intelligence*, (Pasadena), 2009.
7. A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs, "Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system," *Pervasive and Mobile Computing*, vol. 6, no. 4, pp. 455–466, 2010.
8. J. W. Yoon, F. Pinelli, and F. Calabrese, "Cityride: A Predictive Bike Sharing Journey Advisor," in *2012 IEEE 13th International Conference on Mobile Data Management*, pp. 306–311, IEEE, July 2012.
9. R. Kaleta, *An Integrated London Journey Planner*. Master's thesis, Imperial College London, 2012.
10. J. Li, C. Ren, B. Shao, Q. Wang, M. He, J. Dong, and F. Chu, "A solution for reallocating public bike among bike stations," in *Proceedings of 2012 9th IEEE International Conference on Networking, Sensing and Control*, pp. 352–355, IEEE, Apr. 2012.
11. A. Slingsby, J. Dykes, and J. Wood, "Visualizing the dynamics of Londons bicycle hire scheme," *Cartographica The International Journal for Geographic Information and Geovisualization*, vol. 46, no. 4, pp. 239 – 251, 2011.
12. O. O'Brien, "Bike Share Map (http://bikes.oobrien.com)," 2010.
13. O. OBrien, J. Cheshire, and M. Batty, "Mining bicycle sharing data for generating insights into sustainable transport systems," *Journal of Transport Geography*, July 2013.
14. E. Côme and L. Oukhellou, "Model-based count series clustering for Bike-sharing system usage mining, a case study with the V'lib' system of Paris," *Submitted to ACM TIST*, 2012.
15. P. Borgnat, P. Abry, P. Flandrin, C. Robardet, J.-B. Rouquier, and E. Fleury, "SHARED BICYCLES IN A CITY: A SIGNAL PROCESSING AND DATA ANALYSIS PERSPECTIVE," *Advances in Complex Systems*, vol. 14, pp. 415–438, June 2011.
16. M. C. Guenther, A. Stefanek, and J. T. Bradley, "Moment closures for performance models with highly non-linear rates," in *9th European Performance Engineering Workshop (EPEW)*, (Munich), 2012.
17. V. Galpin, "Towards a spatial stochastic process algebra," in *Proceedings of the 7th Workshop on Process Algebra and Stochastically Timed Activities (PASTA)*, (Edinburgh), 2008.
18. R. A. Hayden and J. T. Bradley, "A fluid analysis framework for a Markovian process algebra," *Theoretical Computer Science*, vol. 411, no. 22-24, pp. 2260–2297, 2010.
19. A. Stefanek, R. A. Hayden, and J. T. Bradley, "A new tool for the performance analysis of massively parallel computer systems," *Eighth Workshop on Quantitative Aspects of Programming Languages QAPL 2010 March 2728 2010 Paphos Cyprus*, 2010.
20. A. Stefanek, R. A. Hayden, and J. T. Bradley, "Mean-field Analysis of Large Scale Markov Fluid Models with Fluid Dependent and Time-Inhomogeneous Rates," *Technical report*, 2013.

21. P. Reinecke, T. Krauss, and K. Wolter, "HyperStar: Phase-Type Fitting Made Easy," in *2012 Ninth International Conference on Quantitative Evaluation of Systems*, pp. 201–202, IEEE, Sept. 2012.
22. R. J. Hyndman, "Automatic Time Series Forecasting : The forecast Package for R," *Journal Of Statistical Software*, vol. 27, no. 3, pp. 1–22, 2008.
23. R. S. Bowden and B. R. Clarke, "A single series representation of multiple independent ARMA processes," *Journal of Time Series Analysis*, vol. 33, pp. 304–311, Mar. 2012.
24. A. Chaintreau, J. Y. Le Boudec, and N. Ristanovic, "The age of gossip: spatial mean field regime," *Evolution*, pp. 109–120, 2009.