Gaussian Process-Based Predictive Modeling for Bus Ridership

Sourav Bhattacharya
Helsinki Institute for
Information Technology HIIT
Department of Computer
Science
University of Helsinki, Finland
sourav.bhattacharya@
cs.helsinki.fi

Arto Klami
Helsinki Institute for
Information Technology HIIT
Department of Computer
Science
University of Helsinki, Finland
arto.klami@cs.helsinki.fi

Santi Phithakkitnukoon
Computing Department
The Open University, UK
santi.phi@open.ac.uk

Marco Veloso
Centro de Informática e
Sistemas
Universidade de Coimbra,
Portugal
mveloso@dei.uc.pt

Petteri Nurmi
Helsinki Institute for
Information Technology HIIT
Department of Computer
Science
University of Helsinki, Finland
petteri.nurmi@cs.helsinki.fi

Carlos Bento
Centro de Informática e
Sistemas
Universidade de Coimbra,
Portugal
bento@dei.uc.pt

UbiComp '13 Adjunct, Sept 8-12, 2013, Zurich, Switzerland.
ACM 978-1-4503-2139-6/13/09...$15.00.

Abstract

The dynamics of a city are characterized, among others, by the traveling patterns of its dwellers. Accurate knowledge of human mobility patterns would have applications, e.g., in urban design, in the optimization of public transportation operating costs, and in the improvement of public transportation services. The present paper combines a large scale bus transportation dataset with publicly available data sources to predict bus usage. We propose a Gaussian process-based approach for modeling and predicting bus ridership. To validate our approach we perform experiments on data collected from Lisbon, Portugal. The results demonstrate significant improvements in prediction accuracy compared to a probabilistic baseline predictor.

Author Keywords

Gaussian process, modeling and urban computing.

ACM Classification Keywords

Theory of computation [Machine learning approaches]: Gaussian processes.

Introduction

It is well known that optimized public transit can reduce congestion, gasoline consumption, and overall carbon footprint [5, 13]. From a customer perspective, however, a mobility choice is only a choice if it is fast, comfortable, and reliable. Therefore, strategies for increasing public transit ridership and improving user satisfaction are an active and ongoing area of research.

Transit authorities are increasingly taking advantage of pervasive sensing technologies to provide new types of intelligent transportation services and to understand transit usage. As an example, services that enable travelers to access real-time transit vehicle location, arrival time, connection, and other related pieces of information are emerging (e.g., [6, 8, 17]). From a public transport management perspective, this real-time information can help to improve service and resource management effectiveness, and hence also ridership and user satisfaction.

The present paper investigates how the combination of data collected from pervasive sensors (ticketing and bus arrival information) and public, open data sources (weather and transit network information) can be used to enhance public transportation systems, e.g., for service and resource management such as dispatching and scheduling. Specifically, we propose a Gaussian process-based technique for modeling and predicting bus ridership, i.e., the number of passengers using a specific bus stop at a given period of time. The proposed model allows efficient estimation of the ridership rates in various contexts. Besides modeling the weekly and daily fluctuations in bus usage, the model allows incorporating additional predictors, such as the weather conditions or demographic information.

Most of present day public transport systems, such as buses, are operating with fixed schedules, respectively for weekdays, weekends, night hours, and holidays. Making use of historical logs and contextual information, such as weather and time contexts, can give more insights into the dynamics of ridership demands and user behavior, paving way towards adaptive public transport systems that are responsive to user need and demands. This approach also contributes to the visions of adaptive transportation systems [3, 4] and the 'swam' concept introduced in the 2006 UK Government's Foresight program (i.e., on-demand public transport systems).

We validate our approach using data collected from Lisbon, Portugal, demonstrating that our Gaussian Process-based predictor achieves significantly better performance than a probabilistic baseline predictor.

Related Work

The idea of using pervasive sensing in traffic engineering was first introduced by Zito et al. [18] who investigated the use of GPS for intelligent highway services. Recent advances in information and communication technologies have led to development of several services for transport systems. Camacho et al. [2] presented an overview of IT-based services offered in public transport, and discussed how passenger-centric services can improve public transport systems, particularly service quality and passenger satisfaction.

Availability of public transportation data such as train and bus arrival times opens up new directions for researchers to get a better understanding of the current systems and seek ways to improve upon them. Ferrari et al. [7] described a methodology for measuring accessibility of public transportation system-based on analysis of underground train data, whereas Patnaik et al. [10] developed a predictive model for bus arrival times. Pinel et al. [11] presented a method to measure accessibility of a city using bus probe data. Uno et al. [15] proposed a methodology to evaluate road network-based on travel time stability and reliability using bus probe data. Bejan et al. [1] developed a model for bus journey time estimation and examined influential factors such as time of the day and day of the week. Sun et al. [14] analyzed encounter patterns of people using bus GPS and ticketing data and found regularity in the patterns, which was shown as an empirical evidence of the 'familiar strangers' concept in social network studies.

Data Description

We consider a dataset containing bus information collected over a period of two months (April-May 2010) by one of the largest bus operators in Lisbon, Portugal. The dataset consists of two parts: 1) *bus probe data* and 2) *ticketing data*. The bus probe data contains arrival information of buses at various stops along their predefined route. The ticketing data, on the other hand, contains information about passengers getting on a bus. Below we describe the datasets in greater detail and include description of publicly available weather data and applied data pre-processing.

Bus Probe Data

The bus probe dataset contains information of bus arrival times at $2,104$ bus stops along various bus routes in the city of Lisbon. Overall, the dataset contains measurements from 96 different bus lines comprising of 187 bus routes[1]. During the period of two months, over $650,000$ bus trips were recorded. The mobility context of a bus was captured by recording bus ID, heading/direction, bus stop ID, arrival time at a bus stop and bus stop location (*latitude, longitude*). The locations of the bus stops considered in this study are shown in Figure 1 overlaid on the road network of Lisbon (courtesy of OpenStreetMap[2]).

Ticketing Data

The second part of the dataset contains bus ridership information collected during passenger ticketing. Nearly all bus riders in Lisbon use a transit pass, which is a pre-purchased card that allows the user to use a bus service. When the user uses the transit pass to enter the bus, the card ID is recorded along with a timestamp and bus specific information such as vehicle ID and bus route ID. During the study period of two months, $812,170$ anonymized unique passengers were recorded in the ticketing data.

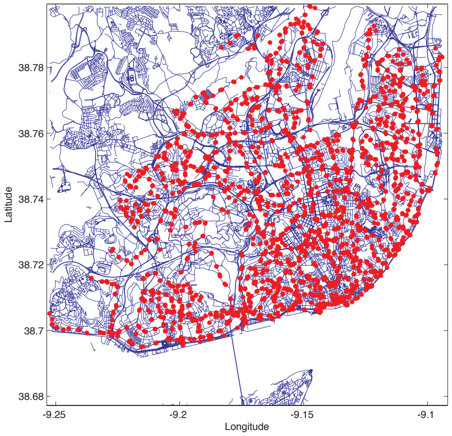

Figure 1: Spatial distribution of all bus stops on top of the road network in the city of Lisbon, Portugal.

Weather data

We augment the Lisbon transportation data with weather information, e.g., temperature, humidity, rain and thunder

[1] Bus routes are calculated by considering different directions of journey by a bus-line, e.g., 'Up', 'Down' or 'Circular'

[2] http://www.openstreetmap.org/

extracted from a publicly available weather data source (courtesy of Weather Underground[3]) that provides weather indicators every half an hour.

Data Pre-processing

Before analysis, we process the dataset by removing all bus lines that are not present in both datasets. The bus probe and ticketing datasets were collected independently, but can be combined to generate *ridership* information by matching all passengers' boarding times recorded in the ticketing dataset to a specific bus arrival time present in the bus probe dataset. More specifically, given the boarding time of a passenger and the bus line ID, we search all records of the same bus line for the day and identify the passenger's bus stop as the one having smallest positive time difference to the arrival times of a bus at various stops along the bus route.The resulting integrated data is shown in Figure 2, which shows the ridership per hour recorded at two bus stops for a period of one week. The plot indicates a clear difference in bus ridership during weekdays and weekends, additionally it clearly shows the existence of a daily cycle.

Modeling Bus Ridership

For predicting the number of people getting on a bus we build separate predictive models for each of the bus stops. The task then is to create a model that takes as input a set of variables describing the current context and provides as an output the expected number of passengers entering a bus in that context. In the simplest case, the context is defined as the time of day and the day of week, but in the more general case the context variables should provide all the information that is needed for making the prediction, such as the weather conditions or the demographics of the destinations that can be reached by the busline.

In the present work we first introduce models that use time as the only context. In particular, we model the daily and hourly fluctuations, depicted in Figure 2, by describing the context with two variables: the day of the week d and the time of the day h (discretized into one-hour intervals). The task then is to estimate $7 \times 24 = 168$ parameters $\boldsymbol{\mu}_{d,h}$ that describe the expected number of passengers getting on a bus at that given time. We propose using Gaussian process (GP) regression [12] for solving the problem, but also present a baseline model that simply counts the passengers in the training data for comparative analysis.

After addressing the task of modeling the weekly and hourly cycles in passenger volume, we demonstrate how the proposed GP model directly extends to richer contexts that enable more accurate predictions. In particular, we incorporate the current weather conditions into the model.

Baseline Algorithm

To model the passenger counts we define $\boldsymbol{\mu}_{d,h}$ as the expected number of passengers getting on any bus that arrives at the stop during day d and hour h. Given a reasonably large collection of training data, we can directly estimate these parameters as the average number of passengers seen entering in that temporal context. We use $\boldsymbol{p}^t_{d,h}$ to denote the number of passengers using the stop during day d and hour h for a given training index (week) t, and $n^t_{d,h}$ to denote the number of buses using the stop. Then the expected number of passengers entering a bus at those conditions can be estimated as:

$$\boldsymbol{\mu}_{d,h} = \frac{\sum_{t=1}^{T} \boldsymbol{p}^t_{d,h}}{\sum_{t=1}^{T} n^t_{d,h}}, \qquad (1)$$

where T denotes the number of weeks present in the training data. Besides intuitive reasoning, the model can

[3] http://www.wunderground.com/

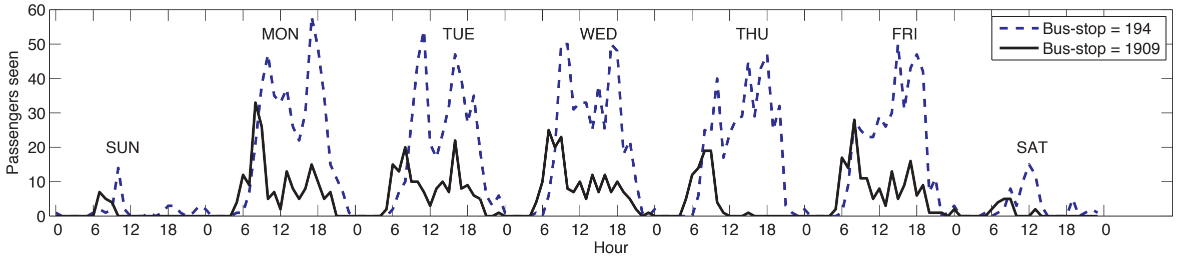

Figure 2: Variation of number of passengers using two bus stops recorded over a period of one week.

be justified as the maximum likelihood estimator for observations that are assumed to follow a normal distribution. For predicting the ridership for a future test case, we simply multiply the expected count with the observed number of buses:

$$\hat{\boldsymbol{p}}_{d,h} = n_{d,h} \cdot \boldsymbol{\mu}_{d,h}. \tag{2}$$

This baseline model makes reasonably accurate predictions in well-defined discrete contexts. It effectively just looks at the history and predicts that the number of passengers will equal that of the mean count in the history, which is often a reasonable prediction. Even a model with $T = 1$ which just predicts that the number of passengers will be the same as it was during the same time last week is useful; we will later demonstrate this as the comparison model. However, the model breaks down if we attempt to extend it to richer contexts. It estimates the ratio separately for each possible context, and hence needs discrete context variables. The richer the context, the more parameters are needed in the model and, more crucially, the more training data is needed to accurately estimate the parameters. It cannot make reasonable predictions for contexts that did not occur in the training data, and for rich context this will necessarily be the case for several if not majority of the contexts.

Gaussian Process Model

As described above, the straightforward approach of looking back at historical data cannot work for rich contexts. Instead, we need to learn models that generalize over nearby contexts. Even if the training data had no examples of the passenger count at Wednesday 3PM when it is raining, the model should provide a reasonable estimate by borrowing information about rainy afternoons in other days and the usual count of passengers near Wednesday 3PM. This task can in general be solved by learning the parameters for different contexts together and by regularizing the solutions towards each other.

In this article we propose to solve the contextual ridership problem by a nonparametric Bayesian approach called Gaussian process regression [12], a powerful probabilistic inference technique that has gained popularity in recent years. It is a regression technique that predicts $\boldsymbol{\mu}$, the expected number of passengers, for arbitrary contexts using a nonlinear mapping from the context variables to the output while constraining the predictions of similar contexts to be close. The GP regression model is nonparametric, which means that we do not need to specify the functional form of the mapping in advance, and in particular need not restrict to modeling the outputs with any simple family such as linear functions. Instead, we only need to define a way of computing similarity between the contexts in the form of a kernel function. The kernel function then implicitly defines a family of smooth functions between the contexts and the outputs.

We denote the context by $\boldsymbol{x}$, which is a vector over the context variables, for example $\boldsymbol{x} = [h, d]$ in the case of temporal context only. The output variable in the training data is computed as $y = \boldsymbol{p}_{d,h}/n_{d,h}$. The GP regression model then specifies that the independent variable $\boldsymbol{y}$ are noisy versions of an arbitrary non-linear function $f(\boldsymbol{x})$, where $f(\boldsymbol{x})$ has a Gaussian process prior. This prior, which is on the functions themselves, specifies that the joint distribution of $f(\boldsymbol{x})$ and $f(\boldsymbol{x'})$ for any $\boldsymbol{x}$ and $\boldsymbol{x}'$ is Gaussian, and that the covariance of these two is given by a similarity kernel $k(\boldsymbol{x}, \boldsymbol{x}')$. This similarity kernel uniquely defines the properties of the GP prior space.

The actual model is defined as

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{f}, \phi &\sim \Pi_{i=1}^{n} p(y_i|f_i, \phi), && (3)\\
f(\boldsymbol{x}|\theta) &\sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'})), && (4)\\
\theta, \phi &\sim p(\theta)p(\phi), && (5)
\end{aligned}$$

where the observations $\boldsymbol{y} = [y_i, \ldots, y_n]^T$ are assumed to be conditionally independent given the function values $f(\boldsymbol{x})$, $\mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'}))$ is the GP prior, and θ and ϕ are hyperparameters of the likelihood and kernel functions with some suitable prior distributions. To fully specify the model we need to fix the likelihood function, the kernel function, and the prior distributions for the hyperparameters.

We start with the likelihood, choosing the Student t-distribution given by:

$$\boldsymbol{y}|\boldsymbol{f}, \nu, \sigma_s \sim \Pi_{i=1}^{n} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma_s} \left(1 + \frac{(y_i - f_i)^2}{\nu\sigma_s^2}\right)^{-(\nu+1)/2},$$

where ν is the degrees of freedom and σ_s is the scale of the observation noise, dictating how close to the function values $f(\pm x)$ the observations are likely to fall. We chose the t-distribution instead of the computationally easier normal distribution due to its robustness to outliers [9], which might be prevalent in our data especially for low-volume stops; even if the average rate is very low, a party of several people might occasionally get on a bus, creating a strong outlier.

The choice of the kernel function determines how the functions $f(\boldsymbol{x})$ behave in the space of all possible contexts, controlling the smoothness of the predictions. We adopt the common choice of Gaussian kernel function

$$k(\boldsymbol{x}, \boldsymbol{x'}) = \sigma_a^2 \exp(-\frac{\sum_{k=1}^{d}(x_k - x_k')^2}{2l_k^2}), \qquad (6)$$

where σ_a^2 controls the amplitude of the functions and the l_k parameters are the length-scales for individual context variables. We could also use other kernels to encode different kind of relationships between the contexts, but the Gaussian kernel is a flexible choice for modeling continuous and ordinal context parameters.

Learning the GP model consists of specifying the values for the hyperparameters, in our case $\phi = (\nu, \sigma_s)$ for the likelihood function and $\theta = (\sigma_a, \{l_k\}_{k=1}^{d})$ for the kernel function. The values for these parameters uniquely define the predictions of the model, and hence there is no need to learn anything else; the model has no actual parameters for the mappings themselves, but instead it defines the predictions indirectly as the posterior distribution of the function values $f(\boldsymbol{x})$ for unseen test contexts. The most important hyperparameters are the length scales l_k, which control the smoothness of the predictions; the larger l_k is for a specific context variable the smoother the predictions are because of borrowing information from a broader range along that variable.

For learning the parameters we use Bayesian inference, namely the Markov chain Monte Carlo (MCMC) technique. Instead of learning specific values for the parameters, we compute the predictions by integrating over the posterior distribution of the hyperparameter values, using the GPstuff package [16] and the default choices for the prior parameters as provided in the package. These predictions cannot be expressed in close form analytical solution, but they are computationally efficient for reasonably sized training collections. That is, for any given context $\boldsymbol{x}$ observed at the test time we can compute the expected count $\boldsymbol{\mu}$. Furthermore, we directly get confidence intervals for the predictions.

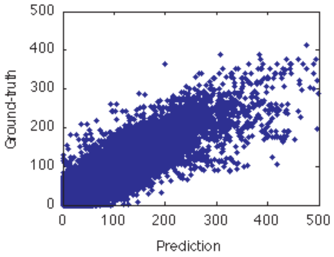

Figure 3: Scatter plot of observed bus ridership with its prediction obtained from GP model.

Results

To illustrate the proposed modeling framework, we conduct a number of experiments on the Lisbon transportation dataset. After the pre-processing and data integration we split the data into individual weeks and performed a 5-fold cross validation to illustrate the predictions and to compare the GP model with the baseline. We always train the models with one week of training data and apply it for predicting the passenger count for the remaining weeks, separately for each of the bus stops. To measure the quality of the prediction we use root mean-square error (RMSE) measure.

We start with the simple temporal context of the day of week and the hour of day, learning both the GP model and the baseline model. Figure 3 shows the crossplot of the observed counts and the predictions made by the GP model, showing that the model does reasonably well. It somewhat overestimates the highest counts and the predictions are noisy, but the correlation between the predictions and the true values is high $(R = 0.93)$. The corresponding plot for the baseline model would be similar, but noisier.

To compare the two models in quantitative terms, we compute the RMSE for each of the bus stops, resulting in an average error of 5.1 for GP and 5.7 for the baseline model. A McNemar test indicated that the resulting difference is statistically significant $(p \ll 0.01.)$ Figure 4 illustrates the difference by plotting the cumulative average RMSE over all the bus stops.

As an exemplary illustration of GP model, we choose one bus stop and plot the bus ridership for 'Wednesday' in Figure 6. The plot includes the mean prediction and the confidence intervals provided by the MCMC integration. The actual observations covering 8 weeks of test data typically fall within the confidence bounds. Figure 5 shows the posterior distribution of the length scale parameter l_k for the variable h. The mode of the distribution is around 3, indicating that the learned kernel for the hour influences the bus ridership prediction significantly for a duration of 3 neighboring hours.

Finally, to illustrate the way the GP model can handle richer contexts, we apply the model on data that include a binary indicator telling whether it rains or

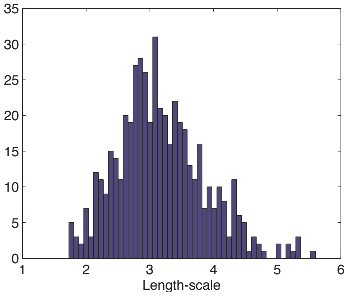

Figure 5: Histogram of the sampled length-scale parameter for the hour variable obtained during MCMC sampling.

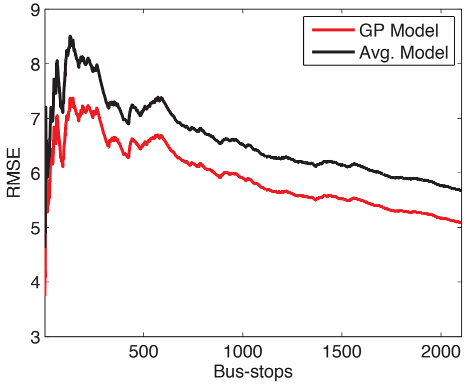

Figure 4: Cumulative RMSE over all bus stops with a 5-fold cross validation.

not. We train the model using one week of data, chosen so that there is instances of rain during the week, and then apply it on the remaining weeks. This results in small improvement, i.e., average RMSE from 4.95 to 4.93, in the prediction accuracy (with $p = 0.02$ i.e., significant at the 95% level).

Conclusion and Future Work

Identifying factors influencing bus ridership is an important step, both towards understanding the dynamics of a city and towards enabling on-demand intelligent transportation services. In this work we proposed a Gaussian process-based predictive model that uses contexts, e.g., time day of the week, hour and rain information to predict bus ridership. We evaluated our approach using two months of bus data collected from Lisbon, Portugal. The GP model results in accurate bus ridership predictions even with a small amount of training data, and is capable of generalizing across contexts.

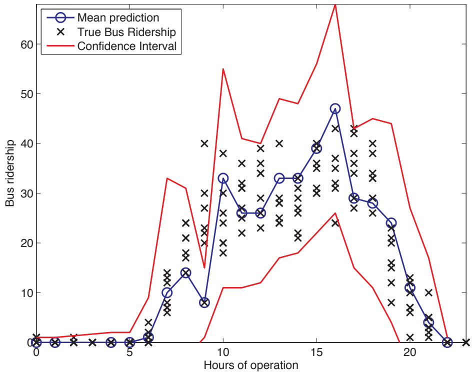

Figure 6: Illustration of the GP prediction for 'Wednesday' at one of the stops. The observed test samples fall usually within the confidence interval, and the mean prediction closely approximates them.

Through experimental analysis we showed that the performance of GP outperforms a simple probabilistic baseline. By complementing the temporal context with a simple weather indicator, we also demonstrated that richer contextual descriptions can further improve predictive accuracy. Even though the improvement in accuracy was only marginal, the experiment acts as a proof of concept that the modeling framework can be extended to include more complex contexts, e.g., temperature, demographic information, bus line popularity, proximity to touristic attraction, etc. We plan to include rich context modeling using GP as our future work.

Acknowledgement

S. Bhattacharya has received funding from the Future Internet Graduate School (FIGS) and the Foundation of Nokia Corporation. A. Klami and P. Nurmi acknowledge TEKES (Tivit Data to Intelligence program) for funding.

References

[1] Bejan, A., Gibbens, R., Evans, D., Beresford, A., Bacon, J., and Friday, A. Statistical modelling and analysis of sparse bus probe data in urban areas. In *13th IEEE Intelligent Transportation Systems Conference*, IEEE (September 2010).

[2] Camacho, T., Foth, M., and Rakotonirainy, A. Pervasive technology and public transport : opportunities beyond telematics. *IEEE Pervasive Computing 12*, 1 (2013), 18–25.

[3] Cervero, R., and Beutler, J. Adaptive transit: Enhancing suburban transit services. University of california transportation center, working papers, University of California Transportation Center, 2000.

[4] Crainic, T. G., Errico, F., Malucelli, F., and Nonato, M. Designing the master schedule for demand-adaptive transit systems. *Annals OR 194*, 1 (2012), 151–166.

[5] Davis, T., and Hale, M. Public transportation's contribution to u.s. greenhouse gas reduction. Tech. rep., 2007.

[6] Dziekan, K., and Kottenhoff, K. Dynamic at-stop real-time information displays for public transport: effects on customers. *Transportation Research Part A: Policy and Practice 41* (2007), 489–501.

[7] Ferrari, L., Berlingerio, M., Calabrese, F., and Curtis-Davidson, B. Measuring public-transport accessibility using pervasive mobility data. *IEEE Pervasive Computing 12*, 1 (2013), 26–33.

[8] Ferris, B., Watkins, K., and Borning, A. Onebusaway: A transit traveller information system, 2010.

[9] Jylänki, P., Vanhatalo, J., and Vehtari, A. Robust gaussian process regression with a student-t likelihood. *J. Mach. Learn. Res. 12* (Nov. 2011), 3227–3257.

[10] Patnaik, J., Chien, S., and Bladikas, A. Estimation of bus arrival times using apc data. *Journal of Public Transportation 7*, 1 (2004), 1–20.

[11] Pinel, F., Hou, A., Calabrese, F., Nanni, M., Zegras, C., and Ratti, C. Space and time-dependant bus accessibility: A case study in rome. In *12th IEEE Intelligent Transportation Systems Conference*, IEEE (September 2009), 1–6.

[12] Rasmussen, C. E., and Williams, C. K. I. *Gaussian processes for machine learning*. The MIT Press, 2006.

[13] Schrank, D., and Lomax, T. Urban mobility report texas transportation institute. Tech. rep., 2009.

[14] Sun, L., Axhausen, K., Lee, D.-H., and Huang, X. Understanding metropolitan collective encounter patterns. *arXiv 1301*, 5979 (January 2013).

[15] Uno, N., Kurauchi, F., Tamura, H., and Iida, Y. Using bus probe data for analysis of travel time variability. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations 13*, 1 (2009), 2–15.

[16] Vanhatalo, J., Riihimaki, J., Hartikainen, J., Jylanki, P., Tolvanen, V., and Vehtari, A. Bayesian modeling with gaussian processes using the gpstuff toolbox, 2013. http://mloss.org/software/view/451/.

[17] Zimmerman, J., Tomasic, A., Garrod, C., Yoo, D., Hiruncharoenvate, C., Aziz, R., Thiruvengadam, N. R., Huang, Y., and Steinfeld, A. Field trial of tiramisu: crowd-sourcing bus arrival times to spur co-design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, ACM (New York, NY, USA, 2011), 1677–1686.

[18] Zito, R., D'Este, G., and Taylor, M. Global positioning systems in the time domain: How useful a tool for intelligent vehicle-highway systems? *Transportation Research C 3*, 4 (1995), 193–2009.