JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2022.2026896

Taylor & Francis
Taylor & Francis Group

Bayesian hierarchical models for the prediction of the driver flow and passenger waiting times in a stochastic carpooling service

Panayotis Papoutsis [a,b,c], Tarn Duong[c], Bertrand Michel[a,b] and Anne Philippe[b]

[a]Department of Computing and Mathematics, Nantes Central Engineering School, Nantes, France; [b]Jean Leray Mathematics Laboratory, University of Nantes, Nantes, France; [c]Department of Data Science-GIS, Ecov, Nantes, France

ABSTRACT

Carpooling is an integral component in smart carbon-neutral cities, in particular to facilitate home-work commuting. We study an innovative carpooling service which offers stochastic passenger-driver matching. Stochastic matching is when a passenger makes a carpooling request, and then waits for the first driver from a population of drivers who are already en route. Crucially a designated driver is not assigned as in a traditional carpooling service. For this new form of stochastic carpooling, we propose a two-stage Bayesian hierarchical model to predict the driver flow and the passenger waiting times. The first stage focuses on prediction of the aggregated daily driver flows, and the second stage processes these daily driver flow into hourly predictions of the passenger waiting times. We demonstrate, for an operational carpooling service, that the predictions from our Bayesian hierarchical model outperform the predictions from a frequentist model and a Bayesian non-hierarchical model. The inferences from our proposed model provide insights for the service operator in their evidence-based decision making.

ARTICLE HISTORY
Received 17 July 2020
Accepted 1 January 2022

KEYWORDS
Hierarchical modelling; Gamma regression; GPS traces; MCMC; multi-level moving average

1. Introduction

Providing ecologically sustainable transportation that is accessible for all is one of the major challenges in the transition to post-carbon societies. A key component of the solution is carpooling services which cater to the mobility requirements in marginalised peri-urban regions with sparser population and physical/digital infrastructure. These carpooling lines, which closely resemble traditional bus lines, connect the physical meeting points for drivers and passengers. The meeting points are placed strategically in highly frequented areas, which take into account various factors such as aggregated traffic flow, socioeconomic characteristics, pedestrian accessibility, local government regulations, etc. This concentrates the demand and the supply of carpooling so that they can reach a critical mass more quickly and more sustainably. These meeting points are where the passenger makes a carpooling

CONTACT Panayotis Papoutsis papoutsispanayotis@gmail.com Department of Computing and Mathematics, Nantes Central Engineering School, Nantes F-44300, France Jean Leray Mathematics Laboratory, University of Nantes, Nantes F-44300, France Department of Data Science-GIS, Ecov, Nantes F-44200, France

request on an electronic console. Since a driver is not allocated in advance, this request is then displayed on a electronic sign on the roadside which informs all passing drivers of a passenger request to the specified destination. This is a real-time, stochastic matching between a passenger and a flow of potential drivers. This new type of passenger-driver matching, along with the aggregating effects of the highly frequented physical meeting points, enables carpooling to reach economical feasibility in peri-urban regions.

From a mathematical and technological point-of-view, it is vastly more difficult to provide a reliable waiting time of a driver arrival in stochastic matching than in deterministic matching. In the latter, a reliable waiting time requires only the tracking of a single assigned driver, whereas stochastic matching requires a more comprehensive understanding of the general driver flow. To assist in the construction of this understanding, the service operator can incentivise drivers to share their GPS locations in real-time. We focus on how to predict the driver flows and passenger waiting times from these GPS traces, which can be considered to be a form of crowd-sourced data collection [13]. Due to the novelty of stochastic carpooling, there is a scarcity of empirically verified predictive models, apart from simple frequentist approaches [16,17]. We propose a more sophisticated Bayesian hierarchical approach where we first build predictive models of the potential driver flow from the observed GPS driver traces, and which are subsequently employed to predict the passenger waiting times. At the time when a passenger request is made, we model this instantaneous driver flow as a moving average of previous driver flows. Then we model the passenger waiting time as a regression with covariates based on the driver traffic flow from the first stage. Our objective is to construct a Bayesian two-stage hierarchical model which is able to predict well the daily driver flows and the hourly passenger waiting times.

The empirical data in this paper are extracted from the 'Lane' stochastic carpooling service (www.lanemove.com), operated by the carpooling provider Ecov (www.ecov.fr), in conjunction with Instant System (www.instant-system.com), in a peri-urban region in south-eastern France. See Papoutsis *et al.* [16] for more details on its set-up. The data collection period is the 382 days from 2018-05-15 (service launch) to 2019-05-31 (beginning of the following year's summer holiday season in France). The daily driver flows in the Lane network are presented in Figure 1, where we enumerate each driver GPS trace, rather than each unique driver. The ordinary weekdays (ORD) are in orange, the school holidays (SCH) in green, and the public holidays/weekends (PWE) in blue. The classic temporal cycles of driver flow data indicate that a moving average is a relevant approach for prediction.

The passenger waiting times cover the period from 2019-07-25 to 2020-02-17. This range of dates is different to those for the driver GPS traces above since, due to operational technical difficulties, the passenger waiting times were not reliably recorded from 2018-05-15 until 2019-10-21, so these dates are excluded from the analysis. In Figure 2 are the (approximately) 1500 observed passenger waiting times aggregated for each day. The Lane service is guaranteed only for ordinary weekdays, and whilst the passengers and drivers are not prevented from using the service on other days, there are far fewer carpooling requests on school holiday weekdays and no requests on public holidays/weekends.

In Section 2, we describe the two stages of the Bayesian hierarchical model for the daily driver traffic flow and the hourly passenger waiting times. In Section 3, we carry out a validation of the proposed model with simulated data. In Section 4 we apply it to empirical data drawn from an operational carpooling service, and compare the predictions of driver and

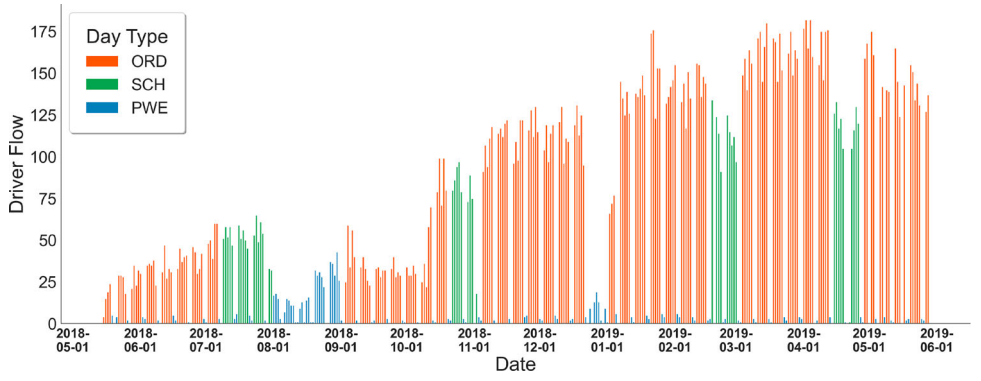

Figure 1. Daily driver flow in the Lane carpooling service, from 2018-05-15 to 2019-05-31. The ordinary weekdays (ORD) are in orange, the school holidays (SCH) in green and the public holidays/weekend days (PWE) in blue.

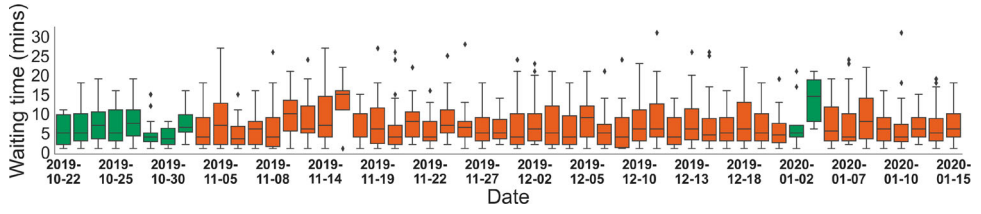

Figure 2. Passenger waiting times (in minutes) in the Lane carpooling service from 2019-10-22 to 2020-01-15. The ordinary work days (ORD) are in orange, and the school holidays (SCH) in green.

passenger behaviour with those from frequentist and Bayesian non-hierarchical models. We end with some concluding remarks.

2. Bayesian hierarchical modelling of driver flow and passenger waiting times

As the driver flow and the passenger waiting time are fundamental quantities in transportation research, their estimation and prediction are the subject of a vast field of active research so we cite only a few references. Historically the simplest models for the driver flow are the moving window averages [18]. More advanced methods draw from time series analysis, within a frequentist [5] or a Bayesian framework [9] have been posited. Established methods for waiting time prediction for stochastic carpooling tend to be frequentist approaches [16,17].

Due to the hierarchical relationship between the driver flows and the passenger waiting times in a stochastic carpooling service, it is natural to consider nested hierarchical models. A general introduction to hierarchical models is provided in Gelman [7] and Gelman and Hill [8]. Hierarchical models can be implemented with frequentist approaches, though we cite only industrial applications using Bayesian approaches here, e.g. image analysis learning [14], football results prediction [1] and electricity load forecasting [20]. Within the transport sector, examples include traffic accident prediction [4] and traffic flow modelling

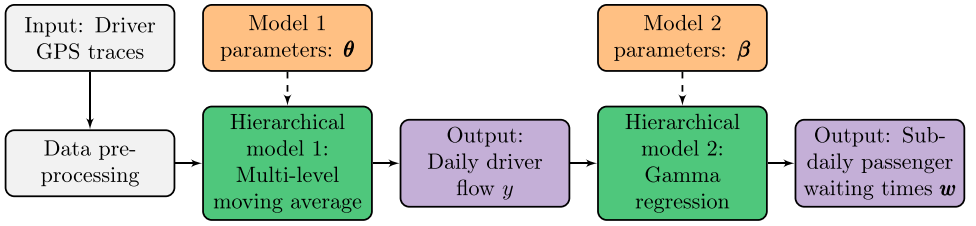

Figure 3. Flowchart of Bayesian hierarchical model for driver flow and passenger waiting time prediction. The input data (driver GPS traces) are in grey, the hierarchical models in green, the model parameters in orange, and the model outputs in purple.

[21]. These latter approaches do not combine the driver traffic flow and the passenger waiting times and do not analyse data with differing time scales in the different stages in the hierarchical model, as we propose.

Our proposed Bayesian multi-level hierarchical model is composed of two nested stages, as illustrated in the flowchart in Figure 3. The input data (driver GPS traces) are preprocessed, as outlined in Appendix A.1, so that they are suitable as input into the hierarchical models. The first model is a multi-level moving average model. It combines the robustness and simplicity of moving averages with the targeted adjustments of the multi-level coefficients [9,18]. The coefficient $\boldsymbol{\theta}$ in this moving average model depends on the day types [2,12], and so the number of components of $\boldsymbol{\theta}$ depends on the number of different day types considered. The output from the first hierarchical model is the daily driver flow, which is the immediate input into the second hierarchical model. The latter is a Gamma regression, whose regression coefficient $\boldsymbol{\beta}$ has S components, with $\beta_s \in (0, 1)$ for $s = 1, \ldots, S$, for each of the S time intervals of a 24-hour period. The role of $\boldsymbol{\beta}$ is to assign the daily traffic flow to these sub-daily time intervals. The output of this second hierarchical model is the temporal profile of the passenger waiting times $\boldsymbol{w} \in \mathbb{R}_+^S$ for these sub-daily time intervals. The scarcity of the driver GPS traces allows us to model the driver flow robustly only at a daily level, whereas a higher temporal resolution of the output passenger waiting times is required for a carpooling service. Bayesian hierarchical models offer an intuitive treatment of these differing temporal resolutions within a single workflow.

2.1. Multi-level moving average for driver flows

From a visual inspection of the daily driver flows in Figure 1, a standard moving average which ignores the day types would be unable to account for the abrupt differences in the driver flow when consecutive days are of different day types. Let the day type function of day i be

$$\mathrm{DT}(i) = \begin{cases} \mathrm{ORD} & \text{if day } i \text{ is an ordinary workday} \\ \mathrm{SCH} & \text{if day } i \text{ is a school holiday} \\ \mathrm{PWE} & \text{if day } i \text{ is a public holiday or a weekend} \end{cases} \quad (1)$$

where $i = 1, \ldots, N$. Thus a suitable K^{th} order recurrence relation of the daily driver flow y_i, for $i \geq K \geq 1$, satisfies

$$y_i = \alpha_{\mathrm{DT}(i)} \sum_{k=1}^{K} \eta_{\mathrm{DT}(i-k)} y_{i-k} + \varepsilon_i \tag{2}$$

where $\alpha_{\mathrm{DT}(\cdot)}$ is the coefficient for the current day i, $\eta_{\mathrm{DT}(\cdot)} = \mathbf{1}\{\mathrm{DT}(\cdot) = \mathrm{ORD}\} + \eta_{\mathrm{SCH}}\mathbf{1}\{\mathrm{DT}(\cdot) = \mathrm{SCH}\} + \eta_{\mathrm{PWE}}\mathbf{1}\{\mathrm{DT}(\cdot) = \mathrm{PWE}\}$ are the coefficients for the past K driver flows, and $\{\varepsilon_i\}$ are a sequence of independent normal random variables $\mathcal{N}(0, \sigma_\varepsilon^2)$. To ensure the identifiability of $\eta_{\mathrm{DT}(\cdot)}$, without loss of generality, we set $\eta_{\mathrm{ORD}} = 1$ for all days.

The model in Equation (2) has a moving average structure of order K, but with two additional multi-level coefficients that make the average adaptive to the day types for the current day i and the previous K days. The multi-level coefficients $\eta_{\mathrm{DT}(\cdot)}$ allows us to model the current driver conditioned on the previous day types, whereas the multi-level coefficients $\alpha_{\mathrm{DT}(\cdot)}$ re-scale these flows conditioned on the current day type. For example, if day i is a school holiday, then the right hand side of Equation (2) is

$$\begin{aligned} \alpha_{\mathrm{SCH}} \sum_{k=1}^{K} [&\mathbf{1}\{\mathrm{DT}(i-k) = \mathrm{ORD}\} + \eta_{\mathrm{SCH}}\mathbf{1}\{\mathrm{DT}(i-k) = \mathrm{SCH}\} \\ &+ \eta_{\mathrm{PWE}}\mathbf{1}\{\mathrm{DT}(i-k) = \mathrm{PWE}\}] y_{i-k}. \end{aligned} \tag{3}$$

In the summand of Equation (3), the day type functions allow us to sum over the K previous days, even if they are of different types. If a previous day is a work day, then its contribution to the current driver flow is $\alpha_{\mathrm{SCH}} y_{i-k}$; if a previous day is a school holiday then it is $\alpha_{\mathrm{SCH}}\, \eta_{\mathrm{SCH}} y_{i-k}$; if a previous day is a public holiday/weekend then it is $\alpha_{\mathrm{SCH}} \eta_{\mathrm{PWE}} y_{i-k}$.

Our model in Equation (2) possesses a similar structure to an autoregressive model, though it does not strictly satisfy the definition of the latter. It cannot be defined with a back shift operator due to the action of the multi-level coefficients $\alpha_{\mathrm{DT}(\cdot)}$ and $\eta_{\mathrm{DT}(\cdot)}$, and the process $\{y_i \in [0, \infty), i = 1, 2, \ldots\}$ is non-stationary due to the drift in the driver participation rate after the launch of the carpooling service.

Let $\boldsymbol{\theta} = (\alpha_{\mathrm{ORD}}, \alpha_{\mathrm{SCH}}, \alpha_{\mathrm{PWE}}, \eta_{\mathrm{ORD}}, \eta_{\mathrm{SCH}}, \eta_{\mathrm{PWE}}, \sigma_\varepsilon^2)$, where we fix $\eta_{\mathrm{ORD}} = 1$ identically, and have $\alpha_{\mathrm{ORD}}, \alpha_{\mathrm{SCH}}, \alpha_{\mathrm{PWE}}, \eta_{\mathrm{SCH}}, \eta_{\mathrm{PWE}} \in (0, 1)$ and $\sigma_\varepsilon^2 \in \mathbb{R}_+$. Let the N days of observed daily driver flows be $y_i, i = 1, \ldots, N$, where $N > K$. Since the error variables are independent Gaussian random variables, then the conditional likelihood of $\boldsymbol{y} = (y_K, y_{K+1}, \ldots, y_N)$ is

$$L(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{(N-K+1)/2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2} \sum_{i=K}^{N} (y_i - g_i(\boldsymbol{\theta}))^2 \right]$$

where $g_i(\boldsymbol{\theta}) = \alpha_{\mathrm{DT}(i)} \sum_{k=1}^{K} \eta_{\mathrm{DT}(i-k)} y_{i-k}$. This conditional likelihood is formed by the product of the conditional densities of y_i, given $y_{i-K}, \ldots, y_{i-1}$, for $i = K+1, \ldots, N$.

In Bayesian analysis, the parameter of interest $\boldsymbol{\theta}$ is a random variable, and its prior distribution π represents our belief in its uncertainty. The posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ represents an update of the prior distribution by taking into account the observed data. In our case, we do not have access to existing knowledge that would provide an informative prior and

thus we form a non-informative prior on $\boldsymbol{\theta}$, i.e. $\pi(\boldsymbol{\theta}) \propto \sigma_\varepsilon^{-2}$ [3, Chapter 1]. This leads to the following posterior distribution

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto L(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \propto \frac{1}{\sigma_\varepsilon^{N-K+1}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=K}^{N}(y_i - g_i(\boldsymbol{\theta}))^2\right]. \tag{4}$$

For the inference on $\boldsymbol{\theta}$, Monte Carlo approximations are required since the posterior distribution (and its moments, quantiles etc.) cannot be calculated explicitly. The most widely used family of methods is the Monte Carlo Markov Chain (MCMC) which generates a Markov Chain $\{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots\}$ whose equilibrium distribution converges to the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

The next stage is to predict a driver flow $\tilde{y}$ in the future from the observed past driver flows $\boldsymbol{y}$. Bayesian prediction is based on the posterior predictive distribution of $\tilde{y}|\boldsymbol{y}$. Its density $p(\tilde{y}|\boldsymbol{y})$ is given by

$$p(\tilde{y}|\boldsymbol{y}) = \int_\Theta p(\tilde{y}|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\theta} \tag{5}$$

where $\Theta = (0,1)^5 \times \mathbb{R}_+$. Since $p(\tilde{y}|\boldsymbol{y})$ is a compound probability distribution, we can simulate samples from this predictive distribution.

For the choice of an MCMC sampler, we use the NUT sampler [11], which is the default in the pyStan Python package (https://pystan.readthedocs.io). This package is an interface to the state-of-art platform for Bayesian computations Stan (https://mc-stan.org). To carry out the integration and then a random draw from the posterior predictive distribution of daily driver flows $p(\tilde{y}|\boldsymbol{y})$ in Equation (4), we are only required to input the prior $\pi(\boldsymbol{\theta})$, the likelihood $L(\boldsymbol{y}|\boldsymbol{\theta})$ and the recurrence relation which generates the vector of simulated driver flows $\boldsymbol{y}$ (Algorithm 2 in Appendix A.2) into pyStan. The latter automatically simulates for the j^{th} iteration, $j = 1, \dots, J$,

$$\tilde{\boldsymbol{y}}^{(j)} = \begin{bmatrix} \tilde{y}^{(j,1)} \\ \vdots \\ \tilde{y}^{(j,N)} \end{bmatrix} \sim \begin{bmatrix} p(\tilde{y}^{(j,1)}|\boldsymbol{y}) \\ \vdots \\ p(\tilde{y}^{(j,N)}|\boldsymbol{y}) \end{bmatrix}, \tag{6}$$

and its final output is the sequence of posterior prediction vectors $\tilde{\mathbb{Y}} = \{\tilde{\boldsymbol{y}}^{(1)}, \dots, \tilde{\boldsymbol{y}}^{(J)}\}$.

2.2. Gamma regression for passenger waiting times

For simplicity, we assume that a passenger can only make one request at a time for themselves only at a carpooling meeting point, and the drivers can embark only one passenger in their vehicle in the order that the passenger requests are made. For day i, let y_i be the daily traffic flow, and that n_i passengers make carpooling requests at times $t_{i,1} < \cdots < t_{i,n_i}$. Let $t'_{i,j}$ be the driver arrival time for the passenger request at time $t_{i,j}, i = 1, \dots, N$ and $j = 1, \dots, n_i$. The perceived waiting time $w^*_{i,j}$ and the pseudo waiting time $w_{i,j}$ for the

passenger request at time $t_{i,j}$ are

$$w^*_{i,j} = t'_{i,j} - t_{i,j}$$
$$w_{i,j} = t'_{i,j} - \max(t_{i,j}, t'_{i,j-1})$$

with the convention $t'_{i,0} = t_{i,1}$ for the first passenger on day i. Figure 4 illustrates the difference between the perceived and the pseudo waiting times for two passengers A, B who are both not the first passenger of the day. Passenger A arrives first and is the j^{th} passenger of day i, and makes a carpooling request at time $t_{i,j}$. Passenger B arrives immediately afterwards and is the $(j+1)^{th}$ passenger with request time $t_{i,j+1}$. Suppose that there are at least two drivers en route to embark these passengers, and they have not received any passenger requests before passenger A's request. The first driver arrives at $t'_{i,j} > t_{i,j+1}$ (i.e. after passenger B's request time) and the second driver at $t'_{i,j+1} > t'_{i,j}$. The perceived waiting time for the passenger A is $w^*_{i,j} = t'_{i,j} - t_{i,j}$ (the blue brace in Figure 4) and for the passenger B it is $w^*_{i,j+1} = t'_{i,j+1} - t_{i,j+1}$ (the green brace). The pseudo waiting time for passenger A is $w_{i,j} = w^*_{i,j}$ (the blue brace) since they are at the front of the queue, and for passenger B it is $w_{i,j+1} = t'_{i,j+1} - t'_{i,j}$ (the grey brace). The pseudo waiting time $w_{i,j+1}$ for passenger B is the difference between their departure time and the departure time of the previous passenger A, and this is shorter than the perceived waiting time $w^*_{i,j+1}$.

From Figure 4, we observe that the perceived the waiting times $w^*_{i,j}$ and $w^*_{i,j+1}$ for passengers A and B overlap, whereas the pseudo waiting times $w_{i,j}$ and $w_{i,j+1}$ do not overlap by construction. The overlapping nature of the interval processes that determine the perceived waiting times renders the problem of their unconditional prediction to be non-identifiable. Thus we focus on the pseudo waiting times, and we wish to formulate sub-daily predictions of them.

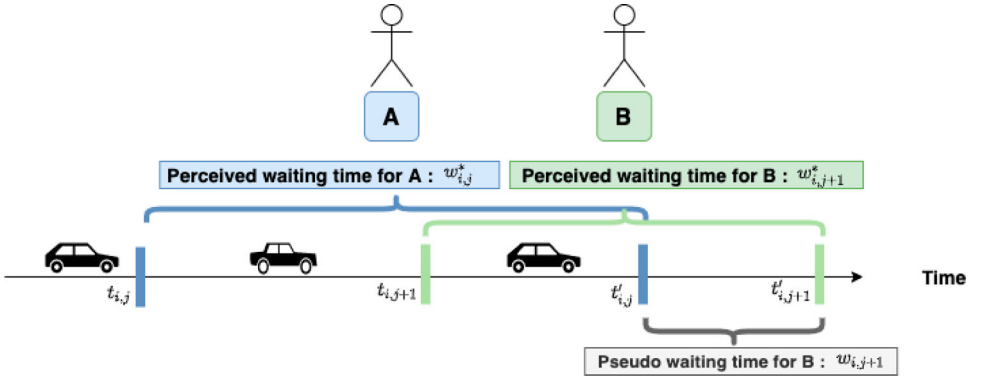

Figure 4. Perceived and pseudo waiting times for the case of two passengers at a carpooling meeting point. Passenger A is at the head of the queue so their perceived waiting time (blue brace) coincides with their pseudo waiting time. For passenger B, their pseudo waiting time (grey brace) is the difference between their departure time and the departure time of passenger A, which is shorter than their perceived waiting time (green brace).

2.2.1. Modelling for pseudo waiting times

Let the 24 hour period of a day be divided into S equal intervals $I_1 < \cdots < I_S$. The fraction of the daily driver flow y_i on each interval $I_s, s = 1, \ldots, S$ is $y_i\beta_s$, where $\beta_s \geq 0$ and $\sum_{s=1}^{S} \beta_s = 1$. Conditional on the traffic flow y_i and the passenger request times $t_{i,j} \in I_s$, we suppose that the pseudo waiting times $w_{i,j}$ are independent Gamma random variables with parameters ν and $\beta_s y_i$, i.e.

$$w_{i,j}|(y_i, \boldsymbol{\beta}, t_{i,j} \in I_s) \sim \Gamma(\nu, \beta_s y_i) \tag{7}$$

for $i = 1 \ldots N$ and $j = 1, \ldots, n_i$. This Gamma regression model assumes that the pseudo waiting times depend on the time of day and on the daily driver flow. Furthermore, it ensures that the conditional mean pseudo waiting time is

$$\mathbb{E}[w_{i,j}|(y_i, \boldsymbol{\beta}, t_{i,j} \in I_s)] = \frac{\nu}{\beta_s y_i}$$

which is consistent with our intuition of the inverse relationship between the driver flow and the waiting time. Since $\boldsymbol{\beta}$ is constant for all i, then the model assumes that the relative proportions of the driver flow in the intervals $I_1, \ldots, I_S$ remain unchanged for all aggregate daily driver flows.

A Dirichlet distribution is a natural choice as a prior distribution on the coefficients $\boldsymbol{\beta}$: $\boldsymbol{\beta} \sim \text{Dir}(S, \boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_S)$ are the concentration parameters, since it imposes the constraint $\sum_{s=1}^{S} \beta_s = 1$ on the coefficients. The corresponding Dirichlet density is $p(\boldsymbol{\beta}) = [\prod_{s=1}^{S} \beta_s^{\alpha_s - 1}]/\text{B}(\boldsymbol{\alpha})$ where $B(\boldsymbol{\alpha}) = \prod_{s=1}^{S} \Gamma(\alpha_s) \big/ \Gamma(\sum_{s=1}^{S} \alpha_s)$ and $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u}\, du$. The $\boldsymbol{\beta}$ vector allows us to rebuild the temporal distribution of the sub-daily traffic flow from an aggregated daily driver flow.

Let $\boldsymbol{t}_i = (t_{i,1}, \ldots, t_{i,n_i})$ be the vector of the n_i observed passenger carpooling request times for the day $i \in \{1, \ldots, N\}$, and $\boldsymbol{t} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_N)$ be all observed passenger carpooling request times. Likewise for the passenger pseudo waiting times $\boldsymbol{w}_i$ for day i, and $\boldsymbol{w}$ for all days. Let $\boldsymbol{y} = (y_1, \ldots, y_N)$ be the observed driver flows for all days. It is reasonable to assume that the waiting times are mutually independent given $(\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{t})$. The conditional likelihood of the pseudo waiting times is thus given by the joint density of $\boldsymbol{w}$ given $(\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{t})$

$$L(\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{t}) = \prod_{i=1}^{N} p(\boldsymbol{w}_i|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{t}) = \prod_{i=1}^{N} p(\boldsymbol{w}_i|\boldsymbol{\beta}, y_i, \boldsymbol{t}_i)$$

since $p(\boldsymbol{w}_i|\boldsymbol{\beta}, y_i, \boldsymbol{t}_i) = \prod_{s=1}^{S} \prod_{\{j:t_{i,j} \in I_s\}} (\beta_s y_i)^{\nu} w_{i,j}^{\nu-1} \exp(-\beta_s y_i w_{i,j}) / \Gamma(\nu)$. Then we obtain the posterior density of $\boldsymbol{\beta}$, using a non-informative prior on $\boldsymbol{\beta}$, as

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{w}) \propto \prod_{i=1}^{N} \prod_{s=1}^{S} \prod_{\{j:t_{i,j} \in I_s\}} \frac{(\beta_s y_i)^{\nu}}{\Gamma(\nu)} w_{i,j}^{\nu-1} \exp(-\beta_s y_i w_{i,j}) \mathbf{1}\{\boldsymbol{\beta} \in \mathbb{R}_+^S\}.$$

Let $\tilde{w}_s$ be the pseudo waiting time for a future day for a passenger who makes a carpooling request in the time interval I_s. If we observe a new daily driver flow $\tilde{y}$ on this future day,

then the posterior predictive distribution of the waiting time $\tilde{w}_s$ is

$$p(\tilde{w}_s|\tilde{y},\boldsymbol{y},\boldsymbol{w}) = \int_0^1 p(\tilde{w}_s|\tilde{y},\beta_s)\pi(\beta_s|\boldsymbol{y},\boldsymbol{w},\boldsymbol{t})\,\mathrm{d}\beta_s, \tag{8}$$

which is then collated into an S-vector $(p(\tilde{w}_1|\tilde{y},\boldsymbol{y},\boldsymbol{w}),\ldots,p(\tilde{w}_S|\tilde{y},\boldsymbol{y},\boldsymbol{w}))$ for all time intervals $I_1,\ldots,I_S$.

To carry out the integration and then a random draw from the posterior predictive distribution of $p(\tilde{w}_s|\tilde{y},\boldsymbol{y},\boldsymbol{w})$ in Equation (8), we are only required to input the posterior predicted value of the driver flow $\tilde{y}$ Equation (5), the recurrence relation which generates the vector of simulated driver flows $\boldsymbol{y}$ (Algorithm 2 in Appendix A.2), and the recurrence relation which generates the vector of simulated passenger pseudo waiting times $\boldsymbol{w}$ (Algorithm 3 in Appendix A.2) into pyStan. The latter automatically simulates this $N \times S$ matrix distribution, for the j^{th} iteration, $j = 1,\ldots,J$,

$$\tilde{\boldsymbol{W}}^{(j)} = \begin{bmatrix} \tilde{w}_{1,1}^{(j)} & \ldots & \tilde{w}_{1,S}^{(j)} \\ \vdots & & \vdots \\ \tilde{w}_{1,N}^{(j)} & \ldots & \tilde{w}_{N,S}^{(j)} \end{bmatrix} \sim \begin{bmatrix} p(\tilde{w}_1^{(j,1)}|\tilde{y},\boldsymbol{y},\boldsymbol{w}) & \ldots & p(\tilde{w}_S^{(j,1)}|\tilde{y},\boldsymbol{y},\boldsymbol{w}) \\ \vdots & & \vdots \\ p(\tilde{w}_1^{(j,N)}|\tilde{y},\boldsymbol{y},\boldsymbol{w}) & \ldots & p(\tilde{w}_S^{(j,N)}|\tilde{y},\boldsymbol{y},\boldsymbol{w}) \end{bmatrix}$$

and its final output is the sequence of posterior prediction matrices $\tilde{\mathbb{W}} = \{\tilde{\boldsymbol{W}}^{(1)},\ldots,\tilde{\boldsymbol{W}}^{(J)})\}$.

2.2.2. Modelling for perceived waiting times

As for now, we have focused on the pseudo waiting times $w_{i,j}$, though for the applications the perceived waiting times $w_{i,j}^*$ are more pertinent, since the latter are the true waiting times from the passenger point-of-view. The current set-up of the Lane carpooling service is not able to collect reliable perceived waiting times since the passenger arrivals are not reliably tracked. Also, as alluded to earlier, the unconditional prediction of the perceived waiting times is a non-identifiable problem.

However, we can provide a framework for their analysis based on conditioning on simulated passenger arrival processes. Since the complete posterior predictive distribution in Equation (8) is available, we can integrate it with respect to the passenger arrival distribution to obtain the pseudo waiting time distribution. Then we are able to reconstruct the perceived waiting times using $w_{i,j+1}^* = w_{i,j+1} + [w_{i,j} - \zeta_i|(w_{i,j} > \zeta_i)]$ with $\zeta_i = t_{i,j+1} - t_{i,j}$. Let the passenger arrivals be a Poisson process, for $t > 0$, $N(t) = \max\{n : \sum_{k=0}^{n} A_k \leq t\}$ where $N(0) = 0$, $A_0 = 0$, and $A_k \sim \mathcal{E}(\lambda)$ and $\lambda > 0$ are independent exponential random variables. The parameter λ is the rate of the passenger arrivals, and it measures the mean number of arrivals over a unit of time. An application of this methodology is given in Section 4.3.

3. Model validation with simulated pseudo waiting times

We choose parameter values to produce simulated data which are comparable to those observed in the Lane carpooling service. We set the initial day $i = 1$ to be 2018-01-01, and the weekdays (ORD), school holidays (SCH) and public holidays/weekends (PWE) to be those observed in south-eastern France. For the simulation algorithms, the number of

days is $N = 365$, the day types coefficients are $\boldsymbol{\theta} = (0.333, 0.33, 0.331, 1, 1, 1)$, the autoregression order is $K = 3$, the error variance is $\sigma_\varepsilon^2 = 5$, the 24 hour period is divided in $S = 8$ equal intervals of 3 hours, the first Gamma shape parameter is $\nu = 7$, the Gamma regression parameters are $\boldsymbol{\beta} = (0.012, 0.01, 0.011, 0.013, 0.018, 0.016, 0.017, 0.019)$, and the number of replicates (waiting times per time interval) is $J = 10$.

We generate one simulated data set of $N = 365$ days, each with one daily driver flow $y_i, i = 1, \ldots, N$ (Algorithm 2), and $J = 10$ passenger pseudo waiting time $N \times S$ matrices $\mathbb{W} = \{\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(J)}\}$ (Algorithm 3), and the corresponding $N \times S$ posterior prediction matrices $\tilde{\mathbb{W}} = \{\tilde{\boldsymbol{W}}^{(1)}, \ldots, \tilde{\boldsymbol{W}}^{(J)}\}$. The data from these $N = 365$ days from 2018-01-01 to 2018-12-31 form the reference training data set. With the same parameters, we simulate a further $\tilde{N} = 5$ days (2019-01-01 to 2019-01-05) as the oracle test data set of $\tilde{N} \times S$ matrices $\mathbb{W}_{\text{test}} = \{\boldsymbol{W}_{\text{test}}^{(1)}, \ldots, \boldsymbol{W}_{\text{test}}^{(J)}\}$. Furthermore, from the training data only, for these same extra $\tilde{N}$ days, we generate the corresponding $\tilde{N} \times S$ posterior prediction matrices $\tilde{\mathbb{W}}_{\text{test}} = \{\tilde{\boldsymbol{W}}_{\text{test}}^{(1)}, \ldots, \tilde{\boldsymbol{W}}_{\text{test}}^{(J)}\}$. For brevity we have omitted the comparison of the driver flows and focus on the passenger waiting times for these simulated data: we make a more thorough comparison of both driver flows and passenger waiting times for the empirical data in the sequel.

From a passenger point of view, whilst the magnitude of the waiting times are important as a perception of the service quality, it is equally important that these posterior predicted waiting times be as close to the observed ones, whatever their magnitude. For example, suppose that a driver arrives 12 minutes after a passenger makes a carpooling request. In this case, a prediction of 15 minutes is better than 5 minutes since the former is closer to the observed waiting time than the latter (which is too optimistic). Therefore we propose the following metric to measure these discrepancies for a given threshold δ:

$$\text{PE}(\mathbb{W}, \tilde{\mathbb{W}}; \delta) = \frac{1}{J\tilde{N}S} \sum_{i=1}^{\tilde{N}} \sum_{s=1}^{S} \sum_{j=1}^{J} \mathbf{1}\{|\bar{\tilde{w}}_{i,s} - w_{i,s}^{(j)}| < \delta\} \tag{9}$$

where $\bar{\tilde{w}}_{i,s} = \frac{1}{J} \sum_{j=1}^{J} \tilde{w}_s^{(j,i)}$ is the mean of the posterior predicted waiting times distribution for day i, and time interval I_s.

We focus on the temporal profiles, over the $S = 8$ periods of a day, of the waiting times. In Figure 5 are the quantiles of the waiting times for all time intervals $I_s, s = 1, \ldots, 8$, for all days $i = 1, \ldots, \tilde{N}$ in the test phase. The grey box plots are of the observations $\mathbb{W}_{\text{test},i,s}$ and the light, medium and dark purple circles superimposed over the box plots are the 50%, 75%, 95% quantiles of the posterior predictions $\tilde{\mathbb{W}}_{\text{test},i,s}$. For operational purposes, short term prediction for the coming week is sufficient, and this is verified by the close agreement of the quantiles of the posterior predicted pseudo waiting times with their observed values for all $\tilde{N} = 5$ prediction days. The advantage of an MCMC approach here is that we are able to reproduce the entire sampling distribution of the predicted waiting times, which is more comprehensive than point or interval predictions.

The PE metric from Equation (9), as a function of δ, illustrated in Figure 6, during both the training phase $\text{PE}(\mathbb{W}, \tilde{\mathbb{W}}; \delta)$ (blue curve) and the test phase $\text{PE}(\mathbb{W}_{\text{test}}, \tilde{\mathbb{W}}_{\text{test}}; \delta)$ (red curve). The test predictions are more accurate than the training predictions for small values of $\delta < 4$ minutes since red PE curve is above the blue PE curve in this interval. This reverses for δ greater than 4 minutes, and after 12 minutes, both curves level off at 1. Thus

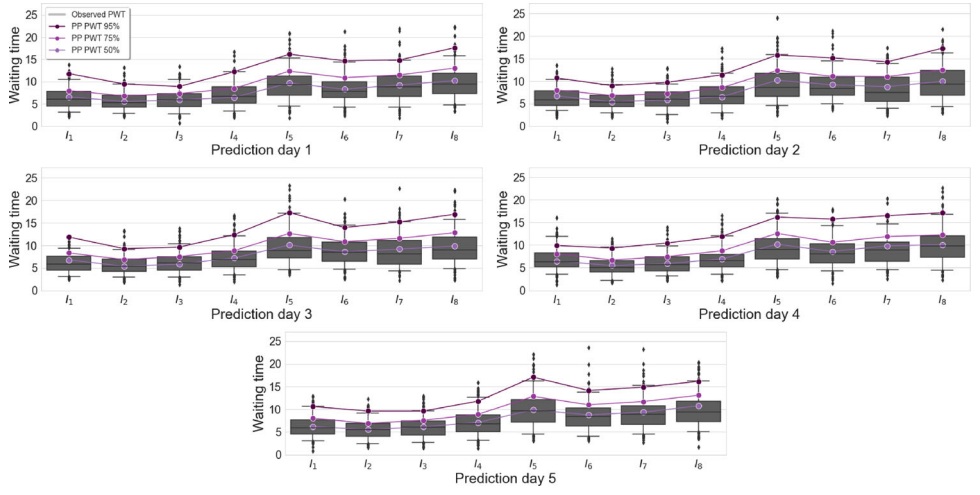

Figure 5. Predictions of simulated pseudo waiting times (in minutes) for 3-hourly intervals, for all $\tilde{N}$ prediction days. The observed waiting times are the grey box plots, and the 50%, 75%, 95% quantiles of the posterior predicted waiting times are the light, medium and dark purple circles.

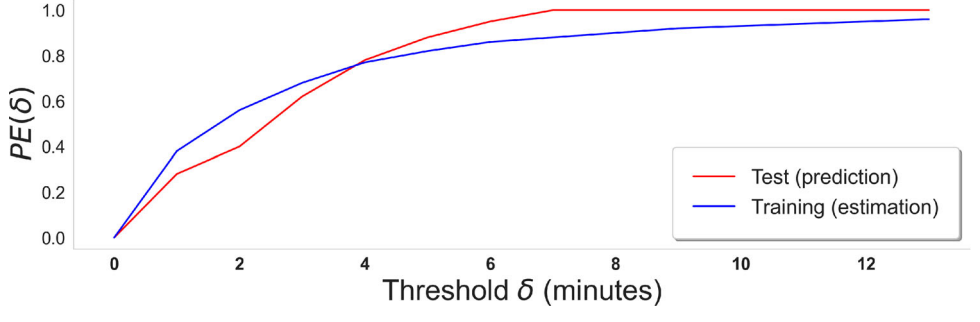

Figure 6. Evolution of the PE metric of simulated and posterior predicted pseudo waiting times, as a function of the threshold δ. The blue curve is for the training phase, and the red curve for the test phase.

the posterior predictions from our proposed Bayesian hierarchical model can have robust prediction performance according to this PE metric.

4. Model validation with the lane carpooling service

Our objective is to employ the two-stage Bayesian hierarchical model to predict the daily driver flow distribution and the passenger pseudo waiting time distribution for the hourly intervals $I_s, s = 1, \ldots, S$, with $S = 24$ for the upcoming week. These are then compared to the observed driver flows and the pseudo waiting times from the same period.

4.1. Daily driver flows

We have approximately 5000 GPS traces for the 382 days from 2018-05-15 to 2019-05-31. We first apply the preprocessing, as outlined in Appendix A.1, to convert the driver GPS

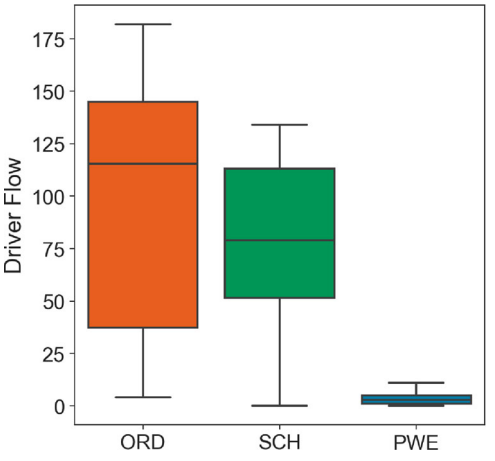

Figure 7. Aggregate driver flow by day type, for the Lane carpooling service from 2018-05-15 to 2019-05-31. The ordinary weekdays (ORD) are in orange, the school holidays (SCH) in green and the public holidays/weekend days (PWE) in blue.

traces into a format suitable for computing the daily driver flows y_i. For the driver flow moving average model in Equation (2), the $\boldsymbol{\theta}$ coefficient has a different value for each day type, since these day types are a key determinant of home-work daily commutes. This is verified empirically by the box plots of the daily driver flow in Figure 7. The daily driver flow for an ordinary weekday (ORD) approaches 150 trajectories, which is about double the driver flow on school holidays (SCH), and more than 20 times larger than on the public holiday/weekends (PWE).

We divide the observed driver flows y_i into 6 different pairs of training phases, starting from 2018-05-15 and with varying N, and test phases with $\tilde{N} = 7$. In each case, we select a test week with certain characteristics as outlined in Table 1. The first column are the dates of the test week, the second column are the day types in the test week, the third column are the dates of the training weeks, and the fourth column is the number of training days (N). For these training-test scenarios, in addition to our proposed Bayesian hierarchical multi-level (BHML) predictions, we compute predictions from a baseline frequentist model (BASE), and a Bayesian Prophet model (PROP). The details of these competing models are described in Appendix A.3. We input the daily driver flows into the first hierarchical model from the Bayesian hierarchical multi-level model BHML to produce the posterior predicted daily driver flows $\tilde{y}_i$, as well the corresponding predictions/estimations from the frequentist baseline model BASE and the Bayesian Prophet model PROP.

For the Test scenario # 6, the training phase covers the dates 2018-05-15 to 2019-05-19. In Figure 8 is the evolution of the goodness-of-fit of the three different models for daily driver flow estimation (leaving out the first week 2018-05-15 to 2018-05-21 which serves as the 'burn-in' period). The goodness-of-fit is measured by the MSE of the estimated and the observed daily driver flows, aggregated per week. Visually the BHML tends to have the best goodness-of-fit (smallest MSE) for most weeks. The sum of these weekly MSEs are BASE: 421.9, PROP: 816.9, BHML: 297.2, which confirms our visual impression that the BHML achieves the best overall estimation accuracy.

Table 1. Training-test scenarios for daily driver flows. The first column are the dates of the test week ($\tilde{N} = 7$), the second is the day types in the test week, the third are the dates of the training weeks and the fourth column is the number of training days N.

	Test week	Test week day types	Training weeks	#training days (N)
#1	2019-01-14 – 2019-01-20	[I]All ORD after holiday period (PWE/SCH)	2018-05-15 – 2019-01-13	244
#2	2019-02-25 – 2019-03-03	All SCH	2018-05-15 – 2019-02-24	286
#3	2019-04-29 – 2019-05-05	[I]All ORD except 1 PWE (2019-05-01)	2018-05-15 – 2019-04-28	349
#4	2019-05-06 – 2019-05-12	[I]All ORD except 1 PWE (2019-05-08)	2018-05-15 – 2019-05-05	356
#5	2019-05-13 – 2019-05-19	[I]All ORD except 1 PWE (transport strike 2019-05-16)	2018-05-15 – 2019-05-12	363
#6	2019-05-20 – 2019-05-26	All ORD	2018-05-15 – 2019-05-19	370

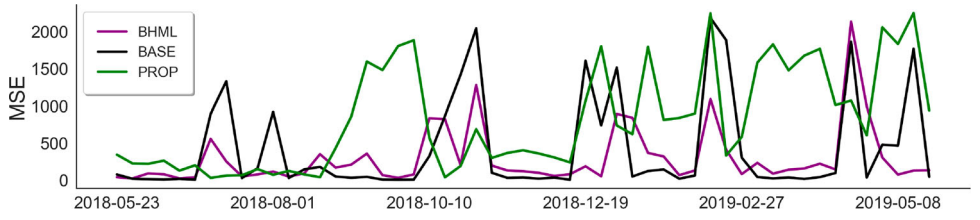

Figure 8. Evolution of the goodness-of-fit of the daily driver flow estimations over the training period (2018-05-15 to 2019-05-19, test scenario # 6). Goodness-of-fit is measured by the weekly aggregated estimation MSE. Bayesian hierarchical multi-level BHML is in purple, frequentist baseline BASE in black, and Bayesian Prophet PROP in green.

Therefore we can be confident that the Bayesian hierarchical multi-level moving average model has good estimation accuracy/goodness-of-fit, but this good performance does not necessarily translate to prediction [15]. So for each scenario described in Table 1, we compute the BHML, BASE and PROP models for the training phase, and then the days of the test phase are input into each these training models to yield the daily driver flow predictions. In Figure 9 are the MSEs between the observed and predicted daily driver flows: the frequentist baseline model BASE in black, the Bayesian Prophet PROP in green, and the Bayesian hierarchical multi-level BHML in purple. The predicted driver flows themselves are presented in Figure A1 in Appendix A.4. Overall the BHML has the best prediction accuracy for all test week scenarios. PROP is the uniformly the worst of these three models for all test weeks. BASE is the best for the test scenario #1 (all ORD after PWE/SCH period) and #6 (all ORD) with almost zero prediction MSE, though the difference with BHML is not so large. These two test scenarios are where all days in the test week are the same day type. For the other test week scenarios #1, #3, #4, #6, BHML has the smallest prediction MSE, some times by a large margin. These test week scenarios include a day which is a different day type to the other days within the test week, which the BHML handles the best.

For the service operator, the sharp differences in the driver flow for different day types within the same week has operational repercussions. For example, since the driver flow is consistently low for all public holidays, the service operator must communicate to passengers that the service quality on a public holiday is not the same as that for an ordinary weekday. This is analogous to a public holiday schedule in lieu of a usual weekday schedule

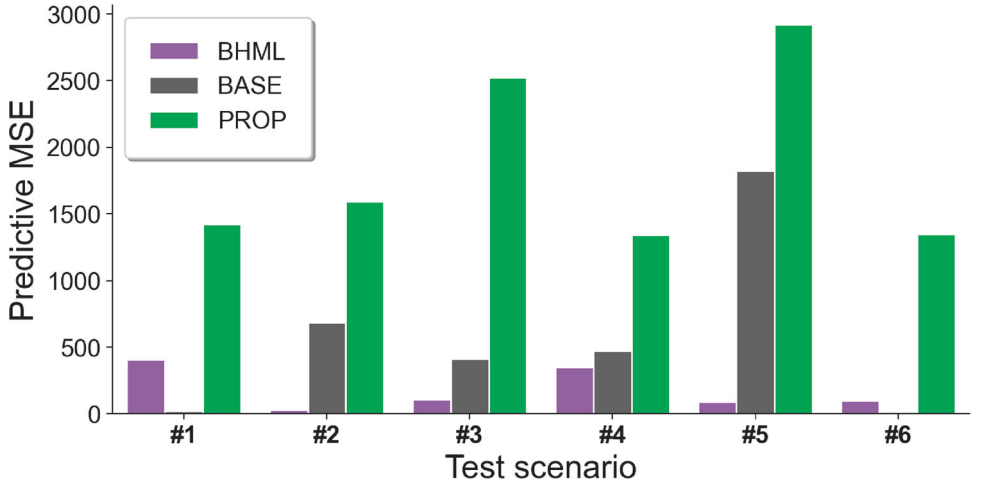

Figure 9. Prediction MSE of the daily driver flow predictions for the six test week scenarios. Bayesian hierarchical multi-level BHML are in purple, frequentist baseline BASE in black, and Bayesian Prophet PROP in green.

provided by a bus operator. On a more positive note, since the driver flow for weekdays on either side of the public holiday is similar to other weekdays further away, then the service operator can also communicate that this temporary reduction in service quality is limited to the public holiday itself and the usual weekday service level can be assured on the preceding and following weekdays.

4.2. Temporal profiles of passenger pseudo waiting times

For the passenger pseudo waiting time Gamma regression, the $\boldsymbol{\beta}$ coefficient, which determines the intra-day distribution of the waiting times, is considered to be constant for all days (see Figure A2 in Appendix A.4 for an illustration of the mean observed daily traffic flows for each weekday from the Lane carpooling service). We observe that each week day has a similar shape so this gives some empirical justification for supposing a constant $\boldsymbol{\beta}$ for all days. For the service operator, this means that it can treat all non-public holiday weekdays as similar to each other.

Before we examine the predictions from this Gamma regression model, we provide some heuristic justification of the model itself, namely concerning the choice of the Gamma distribution and the conditioning of the waiting times with respect to the driver flow (see Figure A3 in Appendix A.4 for a visual justification of the fits of these empirical waiting times to Gamma distributions). As expressed in Equation (7), the pseudo waiting times are represented by different Gamma distributions for each hourly interval, which implies that the mean pseudo waiting times are decreasing functions of the driver flows. In Figure 10, we have divided the observed daily driver flows into three categories: low (< 100 vehicles per day), medium (100–200 vehicles per day), and high (> 200 vehicles per day). Overall we observe that the mean waiting time is inversely proportional to the driver flow level.

Now that we have verified that a Gamma regression model is suitable for the data observed in the Lane carpooling service, we proceed with the BHML to form predictions

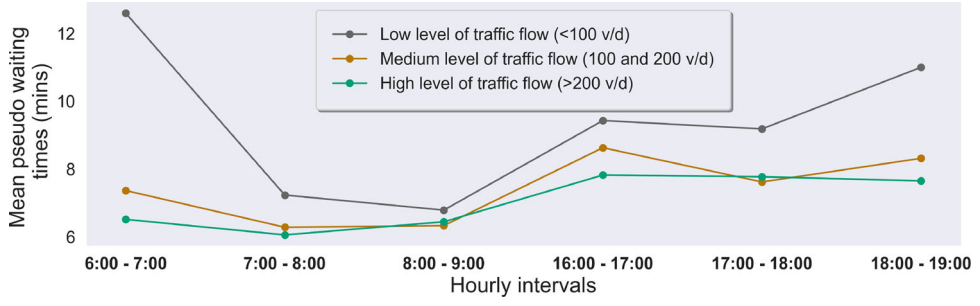

Figure 10. Mean pseudo waiting times per hourly intervals as a function of driver flow from 2019-07-25 to 2020-02-17. Grey: low ($<$ 100 vehicles per day), brown: medium (100–200 vehicles per day), and green: high ($>$ 200 vehicles per day).

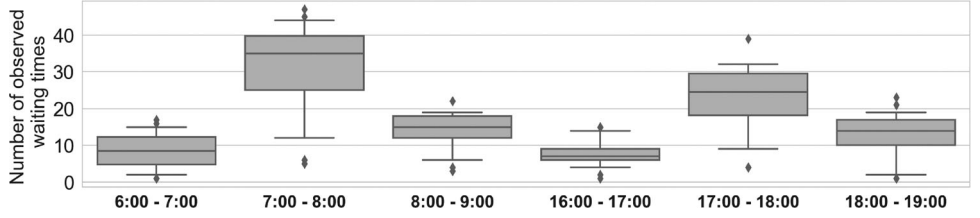

Figure 11. Box plots of the weekly number of observed pseudo waiting times for each hourly interval for weekdays from 2019-07-25 to 2020-02-17.

of the passenger pseudo waiting times. Since there are insufficient passenger carpooling requests to robustly compute observed hourly waiting time profiles over an entire day for the school holidays (SCH) and the public holidays/weekends (PWE), we restrict ourselves to forming predictions for the weekdays (ORD). In Figure 11 are the box plots of the weekly number of observed pseudo waiting times for each hourly interval for the weekdays from 2019-07-25 to 2020-02-17. Although there are $S = 24$ hourly intervals, only those 6 which correspond to the service operating hours (06:00–09:00 and 16:00–19:00) contain any observed passenger waiting times.

There are a maximum of around 40 observed waiting times per hourly interval per week, which are not sufficient to infer robustly their distribution within each interval. To remedy this data sparsity, we aggregate a moving window of test data so for time interval I_s on day i, we combine its observed pseudo waiting times $w_{\text{test},i,s}$ with those for the same time interval from the previous 5 weeks with the same day of week and same day type, i.e. $\{w_{\text{test},i-k,s} : \text{DT}(i-k) = \text{DT}(i), \text{DN}(i-k) = \text{DN}(i), k = 1, \ldots, 35\}$. These days added to the test data are correspondingly removed from the training data. We aggregate the final 5 weeks to be a single test phase, so the Test scenario # 7 is composed of training weeks (2019-07-25 – 2020-01-12) with 1289 observed training waiting times, and test weeks (2020-01-13 – 2020-02-17) with 520 observed test waiting times. We make predictions for only the last test week (2020-02-10 – 2020-02-17), so the number of prediction weekdays remains $\tilde{N} = 5$.

In Figure 12 are the box plots of the observed pseudo waiting times and the quantiles for the posterior predictions, for the hourly intervals for the Test scenario # 7. The observed pseudo waiting times are displayed as the grey box plots, and the 50%, 75%, 95%

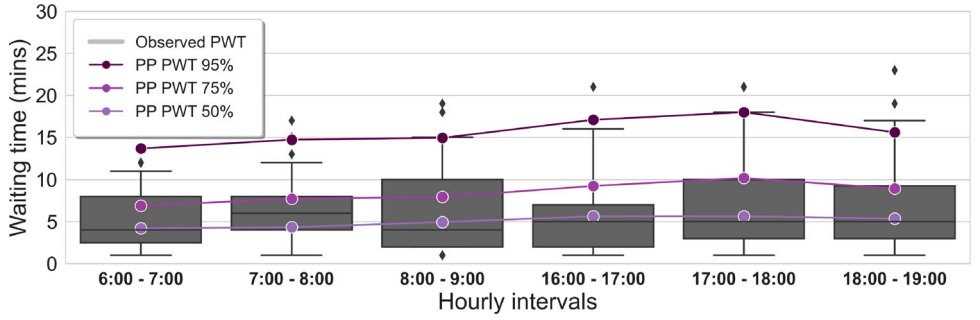

Figure 12. Predictions of passenger pseudo waiting times (PP PWT) for hourly time intervals for the Test scenario # 7. The observed waiting times are the grey box plots, and the 50%, 75%, 95% quantiles of the posterior predicted waiting times are the light, medium and dark purple circles.

quantiles of the posterior predicted waiting times are the light, medium and dark purple circles. The advantage of the BHML is that we have the entire sampling distribution of the predicted waiting times, which is more comprehensive than point or interval predictions of the usual regression models. The median and upper quartile of the predicted pseudo waiting times tend to track those for the observed waiting times, especially for the 06:00–07:00, 17:00–18:00 and 18:00–19:00 intervals. From anecdotal evidence provided by Ecov, 15 minutes corresponds roughly to the maximum time that passengers are willing to wait for a driver to arrive if a pre-arranged meeting time has not been made. With the BHML predictions, we can assert that 95% of the waiting times for passenger requests do not exceed this 15 minutes threshold during most of the operating hours. Whilst this information could also be established with the empirical quantiles of the observed waiting times, the advantage of the BHML is that it gives a more solid basis that this performance will continue into the future.

Lastly we consider our custom PE metric from Equation (9) on the BHML posterior predictions. This metric is illustrated in Figure 13, for both the training phase $\mathrm{PE}(\mathbb{W}, \tilde{\mathbb{W}}; \delta)$ (blue curve) and the test phase $\mathrm{PE}(\mathbb{W}_{\text{test}}, \tilde{\mathbb{W}}_{\text{test}}; \delta)$ (red curve). The blue curve dominates the red curve for most values of δ. This implies that the posterior predictions are more accurate during the test phase than in the training phase. This gives us confidence that the BHML posterior predictions are robust and are not based on over-fitting on the training data.

For the service operator, the BHML implies that the key factor in determining the passenger waiting time (i.e. the output from the second stage) is the driver flow (i.e. the output from the first stage). It is thus imperative that a consistent level of driver participation in the carpooling service is maintained so that consistent waiting times can be provided to passengers. Since they are non-professional drivers, then non-monetary incentives are crucial in maintaining their participation in a stochastic carpooling service [22,23].

In contrast to the comparison of the predicted driver flows from the BHML to those from competing models in the previous section, the comparison of the Gamma regression with other possibilities in the second stage of the BHML is not considered here. According to Papoutsis *et al.* [16], the main predictor for the passenger waiting times is the driver flow, and so we conjecture that the choice of the passenger waiting times prediction model is of secondary importance with respect to the choice of the driver flow prediction model.

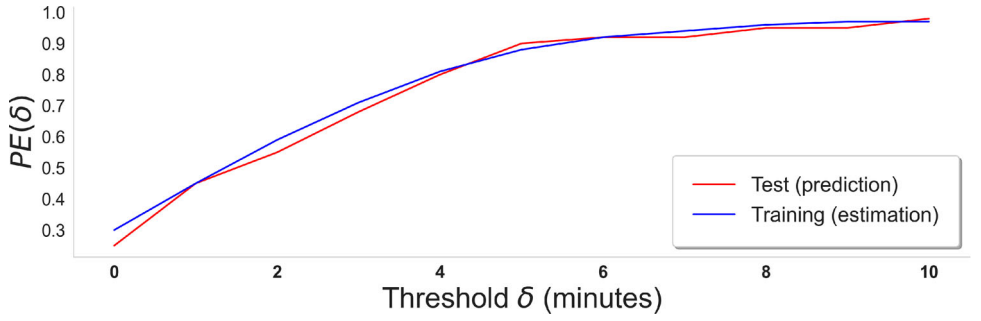

Figure 13. Evolution of the PE metric of observed and BHML posterior predicted passenger pseudo waiting times, as a function of the threshold δ. The blue curve is for the training phase, and the red curve for the test phase.

4.3. Temporal profiles for passenger perceived waiting times

Following the methodology described in Section 2.2.2 we propose the method below in order to reconstruct the perceived waiting times. In our case, the unit of time is one hour and we focus on the opening hours of the service (I_1 = 6:00–7:00, I_2 = 7:00–8:00, I_3 = 8:00–9:00, I_4 = 16:00–17:00, I_5 = 17:00–18:00, I_6 = 18:00–19:00). The base passenger arrival rate is denoted by λ_1, as shown in Table 2. The two other scenarios involve λ_2, an increase by 50%, and λ_3, an increase of 100%.

For the base passenger arrival scenario, there are few situations where the passenger requests overlap each other, and so there is little difference between the pseudo and the perceived waiting times. The second and third scenarios with increased rates of passenger arrivals lead to increased overlapping passenger requests. This is of intense interest to the service operator because it informs them how the passenger perceived waiting times respond to the increased passenger requests whilst maintaining the current driver flow. In Figure 14 are the box plots for the pseudo and perceived waiting times for the three passenger arrival scenarios from Table 2 with a constant driver flow. The pseudo waiting times of the left panel are similar for the three scenarios since they do not account for overlapping passenger arrivals. On the other hand, we observe that the perceived waiting times on the right panel tend to increase as the number of passengers arriving increases. For λ_2 with a 50% increase in the passenger arrivals, the perceived waiting times remain acceptable for a stochastic carpooling service (median less than 15 minutes). However for λ_3 with a 100% increase in the passenger arrivals, the perceived waiting times exceed 15 minutes for many passengers. For the service operator to reduce the waiting times, the driver flow must be increased by increasing the driver participation rate.

Table 2. Poisson passenger arrival rates per hourly intervals. λ_1 is the base passenger arrival rate, λ_2 is an increase by 50% and λ_3 is an increase of 100%.

	Morning intervals			Evening intervals		
	06:00–07:00	07:00–08:00	08:00–09:00	16:00–17:00	17:00–18:00	18:00–19:00
λ_1	8	6	4	6	4	4
λ_2	12	9	6	9	6	6
λ_3	16	12	8	12	8	8

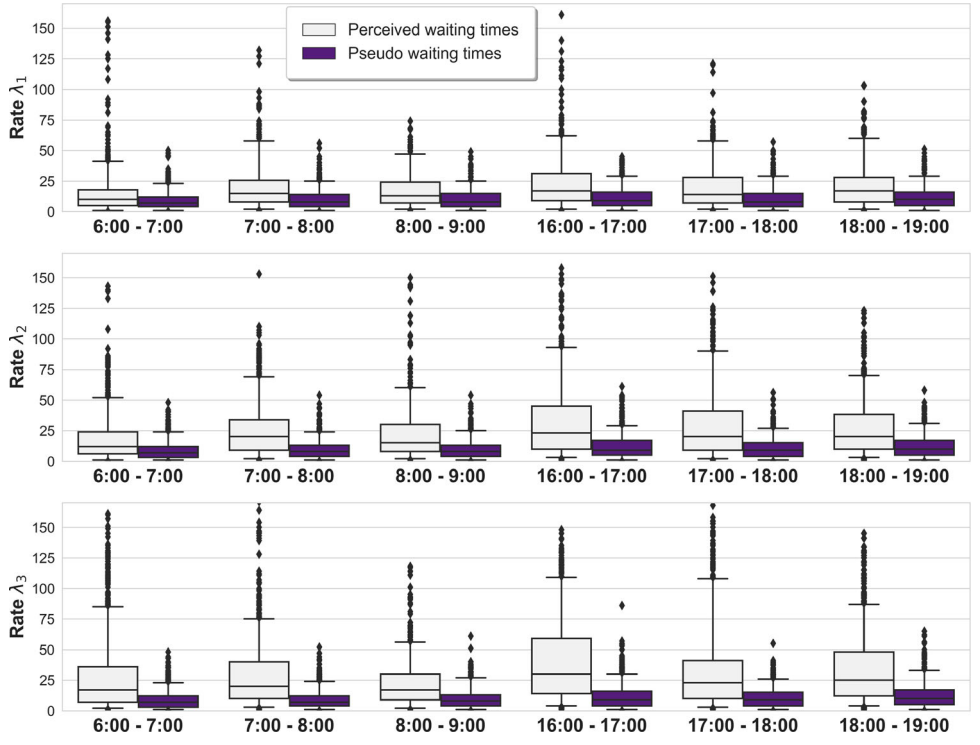

Figure 14. Evolution of passenger waiting times as a function of increased passenger arrival rates. Perceived waiting times (light grey), pseudo waiting times (purple). λ_1 is the base passenger arrival rate, λ_2 is an increase by 50% and λ_3 is an increase of 100%.

5. Conclusions

The main contribution of this paper is the prediction of the daily driver flows and the hourly passenger waiting times using a nested two-stage Bayesian hierarchical model. The first stage is a multi-level moving average model of the daily driver flows, where the multi-level coefficient depends on if the current day is a work day, a school holiday or a public holiday/weekend. The second stage is a Gamma regression where the covariates are the daily driver flows from the first stage, and the response variables are the hourly passenger waiting times. The predicted driver flows and passenger waiting times are robust going into the future, since we demonstrated that they are not due to over-fitting. Furthermore, since we analyse the data from an operational carpooling service, we are able to provide operational advice. For the service operator, the baseline frequentist model is the simplest to implement, and so may be sufficient under certain cost-benefit scenarios. However only the more complex BHML can be utilised for more in-depth data analysis, such as quantiles and confidence regions of passenger waiting times, and for forward planning with the effect of increased passenger requests on waiting times.

We focused on modelling the driver arrival processes and assumed the passenger arrivals to be non-random in the first stage, and on pseudo waiting times in the second stage. One main advantage of the Bayesian hierarchical framework is that it is straightforward to generalise any of the models in the constituent stages (i) to allow the passenger

arrivals to also be a random process, and (ii) to predict both the pseudo and perceived passenger waiting times. These perceived waiting times are of intense operational interest to stochastic carpooling service providers.

Acknowledgements

The authors thank Ecov for providing the data sets of the driver GPS traces and the passenger waiting times. The authors also thank Safa Fennia, Madeleine Zuber, Flavien Sindou and Constant Bridon from Ecov, and Gérard Biau from Sorbonne University for their feedback.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Panayotis Papoutsis http://orcid.org/0000-0001-8470-6024

References

[1] G. Baio and M. Blangiardo, *Bayesian hierarchical model for the prediction of football results*, J. Appl. Stat. 37 (2010), pp. 253–264.
[2] Y. Bao, F. Xiao, Z. Gao, and Z. Gao, *Investigation of the traffic congestion during public holiday and the impact of the toll-exemption policy*, Transp. Res. Part B 104 (2017), pp. 58–81.
[3] P. Congdon, *Applied Bayesian Modelling*, John Wiley & Sons, 2014.
[4] M. Deublein, M. Schubert, B.T. Adey, J. Köhler, and M.H. Faber, *Prediction of road accidents: A Bayesian hierarchical approach*, Accid. Anal. Prev. 51 (2013), pp. 274–291.
[5] A. Ding, X. Zhao, and L. Jiao, *Traffic flow time series prediction based on statistics learning theory*, *Proceedings of the IEEE 5th International Conference on Intelligent Transportation Systems*, 2002, pp. 727–730.
[6] Facebook Core Data Science Group, *Forecasting at Scale*, Facebook. Python package version 0.5. 2019, Available at https://facebook.github.io/prophet. Stan model file Available at https://github.com/facebook/prophet/blob/master/R/inst/stan/prophet.stan.
[7] A. Gelman, *Multilevel (hierarchical) modeling: What it can and cannot do*, Technometrics 48 (2006), pp. 432–435.
[8] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, 2006.
[9] B. Ghosh, B. Basu, and M. O'Mahony, *Bayesian time-series model for short-term traffic flow forecasting*, J. Transp. Eng. 133 (2007), pp. 180–189.
[10] P.G. Gould, A.B. Koehler, J.K. Ord, R.D. Snyder, R.J. Hyndman, and F. Vahid-Araghi, *Forecasting time series with multiple seasonal patterns*, Eur. J. Oper. Res. 191 (2008), pp. 207–222.
[11] M.D. Hoffman and A. Gelman, *The no-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo*, J. Mach. Learn. Res. 15 (2014), pp. 1593–1623.
[12] K.S. Kung, K. Greco, S. Sobolevsky, and C. Ratti, *Exploring universal patterns in human homework commuting from mobile phone data*, PLoS One 9 (2014), pp. e96180.
[13] D.W. Lee and S.H.L. Liang, *Crowd-sourced carpool recommendation based on simple and efficient trajectory grouping*, *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, 2011, pp. 12–17.
[14] F.-F. Li and P. Perona, *A Bayesian hierarchical model for learning natural scene categories*, *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2, 2005, pp. 524–531.
[15] S. Makridakis, R.J. Hyndman, and F. Petropoulos, *Forecasting in social settings: The state of the art*, Int. J. Forecast. 36 (2020), pp. 15–28.

[16] P. Papoutsis, S. Fennia, C. Bridon, and T. Duong, *Relaxing door-to-door matching reduces passenger waiting times: A workflow for the analysis of driver GPS traces in a stochastic carpooling service*, Transp. Eng. 4 (2021), pp. 100061.
[17] J.-B Ray, *Planning a real-time ridesharing network: Critical mass and role of transfers*, *5th Transport Research Arena (TRA) Conference: Transport Solutions from Research to Deployment*. IFSTTAR, 2014.
[18] Y. Stephanedes, P.G. Michalopoulos, and R.A. Plum, *Improved estimation of traffic flow for real-time control*, Transp. Res. Rec. 795 (1980), pp. 28–39.
[19] S.J. Taylor and B. Letham, *Forecasting at scale*, Am. Stat. 72 (2018), pp. 37–45.
[20] S. Wang, X. Sun, and U. Lall, *A hierarchical Bayesian regression model for predicting summer residential electricity demand across the USA*, Energy 140 (2017), pp. 601–611.
[21] L.C. Zammit, M. Attard, and K. Scerri, *Bayesian hierarchical modelling of traffic flow – with application to Malta's road network*, *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, 2013, pp. 1376–1381.
[22] D. Zhu, *More generous for small favour? exploring the role of monetary and pro-social incentives of daily ride sharing using a field experiment in rural Île-de-France*, DigiWorld Econ. J. 108 (2017), pp. 77–97.
[23] D. Zhu, *The limits of money in daily ridesharing: Evidence from a field experiment in rural France*, Revue D'Économie Industrielle 173 (2021), pp. 161–202.

Appendix 1

A.1 GPS traces pre-processing

A GPS trace is an ℓ-sequence of triplets $\boldsymbol{X} = \{(X_i, Y_i, T_i)\}_{i=1}^{\ell}$ where (X_i, Y_i) are the longitude, latitude coordinates of the GPS sensor at the i^{th} timestamp T_i. The m pick-up/drop-off locations in the carpooling network are represented by their GPS coordinates $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_m$. Around each of the m pick-up/drop-off locations $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_m$, a ball of 1 km radius is drawn to obtain $B(\boldsymbol{M}_1), \ldots, B(\boldsymbol{M}_m)$. The intersection of these balls and the GPS trace, $\boldsymbol{X} \cap B(\boldsymbol{M}_1), \ldots, \boldsymbol{X} \cap B(\boldsymbol{M}_m)$, is m sub-sequences of the GPS points of $\boldsymbol{X}$. For those pick-up/drop-off locations with non-empty intersections, we consider that the driver is able to collect a passenger at these points without onerous detours.

This only considers the spatial proximity of the driver to a passenger at a pick-up/drop-off location. For the carpooling to succeed, they also need to be also in temporal proximity. Among the spatial intersections $\boldsymbol{X} \cap B(\boldsymbol{M}_1), \ldots, \boldsymbol{X} \cap B(\boldsymbol{M}_m)$, we examine the corresponding timestamps and retain only those in a suitably restrained time interval. If this set of spatio-temporal intersections is non-empty then we proceed to compute the closest GPS points in $\boldsymbol{X}$ to the pick-up/drop-off locations $\boldsymbol{M}_j$, as defined by $\boldsymbol{X}_{\boldsymbol{M}_j} = \{(X_k, Y_k, T_k) : k = \text{argmin}_{1 \le i \le \ell} \|((X_i, Y_i) - \boldsymbol{M}_j\|\}, j = 1, \ldots, m$. From this closest point $\boldsymbol{X}_{\boldsymbol{M}_j}$, we extract the corresponding timestamp T_k to be an estimate of the driver arrival time at $\boldsymbol{M}_j$.

As an example, suppose that there two pick-up/drop-off points $\boldsymbol{M}_1, \boldsymbol{M}_2$ at which the GPS trace $\boldsymbol{X}$ has well-defined estimated arrival times. Then the ℓ points of $\boldsymbol{X}$ can be reduced to the sequence of 4 points $\tilde{\boldsymbol{X}} = \{(X_1, Y_1, T_1) > \boldsymbol{X}_{\boldsymbol{M}_1} > \boldsymbol{X}_{\boldsymbol{M}_2} > (X_\ell, Y_\ell, T_\ell)\}$ where (X_1, Y_1, T_1) is the driver origin and (X_ℓ, Y_ℓ, T_ℓ) is the driver destination. With this simplified trace $\tilde{\boldsymbol{X}}$, we are still able to determine if the driver can fulfil a passenger request at $\boldsymbol{M}_1$ for a trip going to $\boldsymbol{M}_2$ at time t. The complex topology of $\boldsymbol{X}$ is simplified by retaining a small number of key derived indicators [13].

A.2 Simulation algorithms

Algorithm 1 simulates a driver flow for a single day with no day types. Equation (2) with no day types simplifies to $y_i = \alpha \sum_{k=1}^{K} y_{i-k} + \varepsilon_i$, which is a true autoregressive model. The inputs are the day i, the coefficient α, the autoregression order K, and the error variance σ_ε^2. The output is a single driver flow for day i. The repeat loop ensures that the simulated driver flow is strictly positive. To simulate a sequence of N driver flows, we initialise the values generated by Algorithm 1 for $i = 1, \ldots, K$ days, and then iterate Algorithm 1 sequentially for $i = K + 1, \ldots, N$.

Algorithm 1: Daily driver flow without day types

procedure TrafficFlow($i, \alpha, K, \sigma_\varepsilon^2$)
if $i <= K$ **then**
  initialise $y \longleftarrow \mathcal{N}(30, \sigma_\varepsilon^2)$
else
  repeat
    $y \longleftarrow \mathcal{N}(\alpha \sum_{k=1}^{K}$ TrafficFlow$(i-k, \alpha, K, \sigma_\varepsilon^2), \sigma_\varepsilon^2)$
  until $y > 0$;
end
return: y driver flow for day i

With Algorithm 1 defined, it is straightforward to define one with day types (i.e. Equation (2)) in Algorithm 2. The latter has similar inputs: the day i, the day type coefficients $\boldsymbol{\theta}$, the autoregression order K, the error variance σ_ε^2, and the vector coefficients $\boldsymbol{\theta}$. The output is the daily driver flow for day i, accounting for the day types before day i.

Algorithm 2: Daily driver flow with day types

procedure TrafficFlowDT$(i, \boldsymbol{\theta}, K, \sigma_\varepsilon^2)$
if DT(i) == ORD **then**
  $y \longleftarrow$ TrafficFlow$(i, \alpha_{\text{ORD}}, K, \sigma_\varepsilon^2)$
else
  if DT(i) == SCH **then**
    $y \longleftarrow$ TrafficFlow$(i, \alpha_{\text{SCH}}\eta_{\text{SCH}}, K, \sigma_\varepsilon^2)$
  else
    if DT(i) == PWE **then**
      $y \longleftarrow$ TrafficFlow$(i, \alpha_{\text{PWE}}\eta_{\text{PWE}}, K, \sigma_\varepsilon^2)$
    end
  end
end

Algorithm 3 simulates the passenger pseudo waiting times in Equation (7) for a sequence of days. The inputs are the number of days N, the day type coefficients $\boldsymbol{\theta}$, the autoregression order K, the error variance σ_ε^2, the first shape parameter for the Gamma distribution ν, the S regression parameters $\boldsymbol{\beta}$, and the number of replicates of the waiting times J. The output are J replicates of a pseudo waiting time for each time interval $I_s, s = 1, \ldots, S$, for each day $i = 1, \ldots, N$. The TrafficFlowDT procedure (Algorithm 2) is called outside of the replicates loop since all waiting times on a given day are simulated from the same daily driver flow.

An iteration of the nested loop in Algorithm 3 in the Appendix results in a single $N \times S$ matrix of pseudo waiting times drawn from the appropriate Gamma distributions

$$\boldsymbol{W}^{(j)} \sim \begin{bmatrix} \Gamma(\nu, \beta_1 y_1) & \ldots & \Gamma(\nu, \beta_S y_1) \\ \vdots & & \vdots \\ \Gamma(\nu, \beta_1 y_N) & \ldots & \Gamma(\nu, \beta_S y_N) \end{bmatrix}.$$

Algorithm 3: Passenger pseudo waiting times

procedure WaitingTime($N, \boldsymbol{\theta}, K, \sigma_\varepsilon^2, \nu, \boldsymbol{\beta}, J$)
  $S \longleftarrow$ Len($\boldsymbol{\beta}$)
  for i **in** *1:N* **do**
    $Y[i] \longleftarrow$ TrafficFlowDT($i, \boldsymbol{\theta}, K, \sigma_\varepsilon^2$)
  end
  for j **in** *1:J* **do**
    for i **in** *1:N* **do**
      for s **in** *1:S* **do**
        $\boldsymbol{W}^{(j)}[i,s] \longleftarrow \Gamma(\nu, \beta_s Y[i])$
      end
    end
  end
  return: $\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(J)}$ sequence of waiting time matrices

for $j = 1, \ldots, J$. These are collated into the sequence

$$\mathbb{W} = \{\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(J)}\} = \left\{ \begin{bmatrix} w_{1,1}^{(1)} & \cdots & w_{1,S}^{(1)} \\ \vdots & & \vdots \\ w_{N,1}^{(1)} & \cdots & w_{N,S}^{(1)} \end{bmatrix}, \ldots, \begin{bmatrix} w_{1,1}^{(J)} & \cdots & w_{1,S}^{(J)} \\ \vdots & & \vdots \\ w_{N,1}^{(J)} & \cdots & w_{N,S}^{(J)} \end{bmatrix} \right\}.$$

As Equation (8) generates only a single posterior prediction $\tilde{w}_s$ for a time interval I_s, we collate these $\tilde{w}_s$ for $s = 1, \ldots S$ into an S-vector, and in turn collate N of these S-vectors of posterior prediction distributions row-wise into a $N \times S$ matrix.

A.3 Competing models for driver flows

In addition to the multi-level moving average model for driver flows, we consider a baseline frequentist model and a Bayesian Prophet model. The baseline frequentist model has multi-levels like our model, but without the Bayesian moving average structure. To account for the school/public holidays, as proposed by [10], if day i is not a school/public holiday then the average is calculated over all previous days with the same day of week as day i; and if day i is a school/public holiday, then the average is over all previous school/public holidays. That is,

$$y_i = \frac{1}{|T_d(i)|} \sum_{k \in T_d(i)} y_{i-k} \mathbf{1}\{\mathrm{DT}'(i) \neq \mathrm{HOL}\} + \frac{1}{|T_{\mathrm{HOL}}(i)|} \sum_{k \in T_{\mathrm{HOL}}(i)} y_{i-k} \mathbf{1}\{\mathrm{DT}'(i) = \mathrm{HOL}\} + \varepsilon_i \quad \text{(A1)}$$

where the day type function is

$$\mathrm{DT}'(i) = \begin{cases} \mathrm{ORW} & \text{if day } i \text{ is an ordinary workday or a weekend day} \\ \mathrm{HOL} & \text{if day } i \text{ is a school or a public holiday;} \end{cases}$$

$T_d(i) = \{k : k < i, \mathrm{DN}(i-k) = d\}$ is the set of days with the same day of week before day i; $T_{\mathrm{HOL}}(i) = \{k : k < i, \mathrm{DT}'(i-k) = \mathrm{HOL}\}$ is the set of school/public holidays before day i; and DN is the day of week number function, $\mathrm{DN}(i) = 1$ if day i is a Monday, $\mathrm{DN}(i) = 2$ if day i is a Tuesday etc.

The Bayesian Prophet model, devised by Taylor and Letham [19]; Facebook Core Data Science Group [6], is an additive model with three components:

$$y_i = g(i) + s(i) + h(i) + \varepsilon_i \quad \text{(A2)}$$

where $g(i)$ is the trend, $s(i)$ is the seasonality, and $h(i)$ is the holiday effect. The linear trend is $g(i) = (k + \boldsymbol{a}(i)^\top \boldsymbol{\delta})i + (m + \boldsymbol{a}(i)^\top \boldsymbol{\gamma})$ where k is the growth rate, m is the offset, $\boldsymbol{a}$ is the change point indicator, $\boldsymbol{\delta}$ is the growth rate adjustment, and $\boldsymbol{\gamma}$ is the piece-wise continuity adjustment to ensure that g is continuous. The seasonality component is a Fourier decomposition $s(i) = \sum_{\ell=1}^{L} [\alpha_\ell \cos(2\pi \ell i/P) + \beta_\ell \sin(2\pi \ell i/P)]$ where $(\alpha_\ell, \beta_\ell)$ are the Fourier coefficients, L is the number of Fourier coefficients and P is the period (in days). The holiday effect is $h(i) = \boldsymbol{h}(i)^\top \boldsymbol{\kappa}$ where, say, $\boldsymbol{h}(i) = (\mathbf{1}\{\mathrm{DT}(i) = \mathrm{SCH}\}, \mathbf{1}\{\mathrm{DT}(i) = \mathrm{PWE}\})$ is the vector of indicator variables of the type of holiday of day i, and $\boldsymbol{\kappa}$ is the weight vector, usually equal to the all-ones vector. Taylor and Letham [19] provide the details for the construction of the change point function $\boldsymbol{a}(t)$ and the continuity adjustment parameter $\boldsymbol{\gamma}$. These authors set the number of Fourier coefficients to be $L = 10$ for yearly cycles and $L = 3$ for weekly cycles. What remains is to estimate the trend growth rate k, the offset m, the growth rate adjustments $\boldsymbol{\delta}$ and the Fourier coefficients $\boldsymbol{\alpha}$.

A.4 Supplementary figures for the model validation of the lane carpooling service

Figure A1 illustrates the predictions for each scenario described in Table 1: the Bayesian hierarchical multi-level BHML in purple, the frequentist baseline model BASE in black, the Bayesian Prophet PROP in green, and as well as the observed daily driver flows in blue. The PROP predictions are mostly too low on week days and too high on weekends for all six test weeks in comparison to the observed driver flows, whilst the BHML appears to have marginally better prediction performance than the BASE.

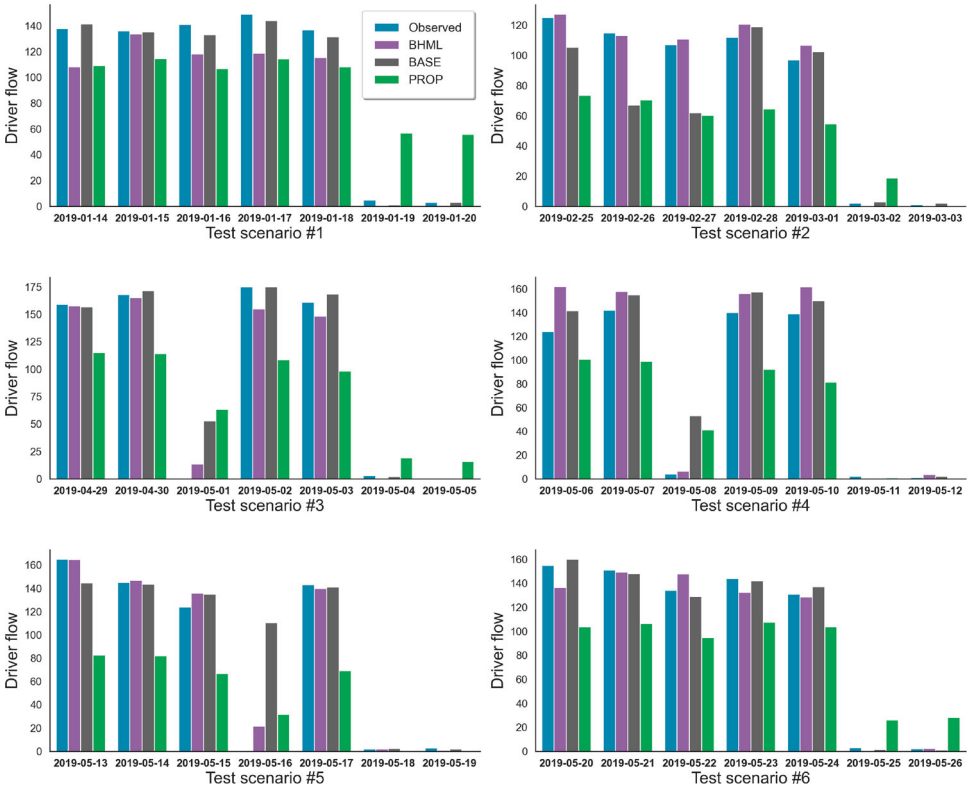

Figure A1. Predictions of daily driver flows for the six test week scenarios. Observed daily driver flows are in blue. Bayesian hierarchical multi-level BHML are in purple, frequentist baseline BASE in black, and Bayesian Prophet PROP in green.

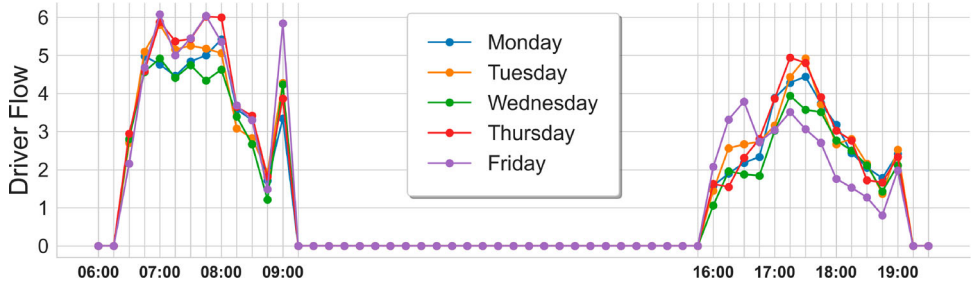

Figure A2. Mean driver flows for 15 minute intervals for each weekday for the Lane carpooling service, from 2018-05-15 to 2019-05-31. Monday is in blue, Tuesday in orange, Wednesday in green, Thursday in pink, Friday in purple.

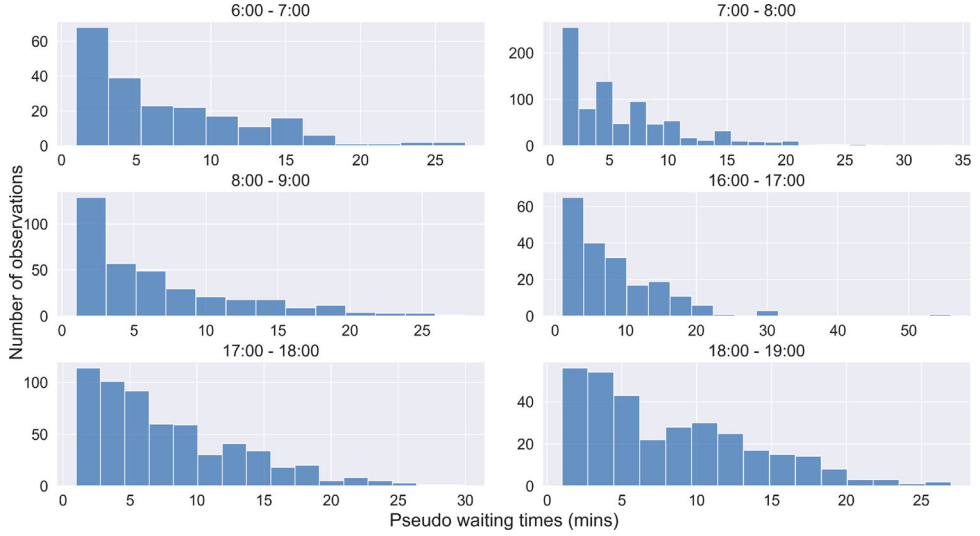

Figure A3. Histograms of observed pseudo waiting times for each hourly interval for weekdays from 2019-07-25 to 2020-02-17.

Figure A2 displays the mean observed daily traffic flows for each weekday from the Lane carpooling service, where the day is divided into 15 minute intervals. Since the service operating hours are 06:00–09:00 and 16:00–19:00, there are few drivers outside them. Each dot in the figure is the mean number of drivers for each 15 minute interval for each week day from 2018-05-15 to 2019-05-31.

In Figure A3 are the histograms of the observed pseudo waiting times for each hourly interval for weekdays from 2019-07-25 to 2020-02-17. We observe that these empirical distributions resemble Gamma distributions and have a different rate parameter β_s within each different hourly interval.