Working paper

2023-12

Statistics and Econometrics
ISSN 2387-0303

Deep Learning and Bayesian Calibration Approach to Hourly Passenger Occupancy Prediction in Beijing Metro: A Study Exploiting Cellular Data and Metro Conditions

He Sun, Stefano Cabras

uc3m | Universidad Carlos III de Madrid

Deep Learning and Bayesian Calibration Approach to Hourly Passenger Occupancy Prediction in Beijing Metro: A Study Exploiting Cellular Data and Metro Conditions

He Sun,
University Carlos III of Madrid (Spain)
Beijing Metro Group Ltd (China)
and
Stefano Cabras
University Carlos III of Madrid (Spain)

November 3, 2023

Abstract

In burgeoning urban landscapes, the proliferation of the populace necessitates swift and accurate urban transit solutions to cater to the citizens' commuting requirements. A pivotal aspect of fostering optimized traffic management and ensuring resilient responses to unanticipated passenger surges is precisely forecasting hourly occupancy levels within urban subway systems. This study embarks on delineating a two-tiered model designed to address this imperative adeptly:

1. **Preliminary Phase - Employing a Feed Forward Neural Network (FFNN)**: In the initial phase, a Feed Forward Neural Network (FFNN) is employed to gauge the occupancy levels across various subway stations. The FFNN, a class of artificial neural networks, is well-suited for this task because it can learn from the data and make predictions or decisions without being explicitly programmed to perform the task. Through a series of interconnected nodes, known as neurons, arranged in layers, the FFNN processes the input data, adjusts its weights based on the error of its predictions, and optimizes the network for accurate forecasting. For the random process of occupation levels in time and space, this phase encapsulates the so-called *process filtration*, wherein the underlying patterns and dynamics of subway occupancy are captured and represented in a structured format, ready for subsequent analysis. The estimates garnered from this phase are pivotal and form the foundation for the subsequent modelling stage.
2. **Subsequent Phase - Implementing a Bayesian Proportional-Odds Model with Hourly Random Effects**: With the estimates from the FFNN at disposal, the study transitions to the subsequent phase wherein a Bayesian Proportional-Odds Model is utilized. This model is particularly adept for scenarios where the response variable is ordinal, as in the case of occupancy levels (Low, Medium, High). The Bayesian framework, underpinned by the principles of probability, facilitates the incorporation of prior probabilities on model parameters and updates this knowledge with observed data to make informed predictions. The unique feature of this model is the incorporation of a random effect for hours, which acknowledges the inherent

variability across different hours of the day. This is paramount in urban transit systems where passenger influx varies significantly with the hour.

The synergy of these two models facilitates calibrated estimations of occupancy levels, both conditionally (relative to the sample) and unconditionally (on a detached test set). *This dual-phase methodology furnishes analysts with a robust and reliable insight into the quality of predictions propounded by this model.* This, in turn, avails a data-driven foundation for making informed decisions in real-time traffic management, emergency response planning, and overall operational optimization of urban subway systems.

The model expounded in this study is presently under scrutiny for potential deployment by the Beijing Metro Group Ltd. This initiative reflects a practical stride towards embracing sophisticated analytical models to ameliorate urban transit management, thereby contributing to the broader objective of fostering sustainable and efficient urban living environments amidst the surging urban populace.

Keywords: Bayesian model calibration; Deep Learning; Integrated Nested Laplace Approximation; Proportional odds model; Spatial-temporal modelling.

1 Introduction to the problem and relevant statistical concepts

The escalating demand for urban transit has led to recurrent congestion and overcrowding within transportation networks, especially during peak traffic hours. Numerous metropolitan regions face persistent traffic challenges despite significant financial investments in transportation infrastructure. Overcrowding at stations poses safety risks and affects the regularity and reliability of public transport services. This, in turn, influences passengers' travel behaviours and path choices, thereby raising the need for adept traffic management solutions. Transport planners and researchers have been propelled to develop transit assignment models in response to these challenges. These models aim to predict passenger volumes across different services connecting various origin-destination pairs and appraise operational services' effectiveness within transit systems (Fu et al., 2012). Accurate and reliable real-time metro passenger flow data is indispensable for developing effective traffic plans, which is crucial for managing the occupancy levels at stations within a metro network.

In contrast to the relative stability observed in daily occupation levels, which reflects the aggregate sum of hourly occupations as per the Central Limit Theorem (CLT) – the fundamental theorem in probability theory that describes how the distribution of the passenger aggregation from hours to the day, becomes Gaussian distributed – hourly passenger counts exhibit pronounced variability across stations and hours. Past studies have underscored the relative regularity of daily passenger flows across workweeks (Jie et al., 2018). Most existing Traffic Control Centers utilize time series models for analysis and forecasting (Eric Lin, 2017). While recent works Han et al. (2019); Roos et al. (2016) have employed Convolutional Neural Networks (CNN) – a type of Deep Learning model primarily used for image and video recognition but also applied in other domains – to model passenger traffic, these methodologies fall short in providing a probabilistic approach to elucidate prediction uncertainty. The extant literature predominantly focuses on modelling time series of counts Davis et al. (2021) or transitions of passengers among stations Zhao & Ma (2022); Fu et al. (2014), with Bayesian approaches to passenger predictions being relatively scant. The Kalman filter used by Jiao et al. (2016) is deemed unsuitable for counting data, and the Multivariate Dynamic Linear Model (DLM) emerges as a potential alternative. However, it necessitates further refinement to account for spatial relations across stations (Petris, 2010; Petris et al., 2009; Rodriguez et al., 2020). No-

tably, Cabras & He (2023) implemented a Bayesian spatiotemporal model for predicting daily passenger flows.

Recent advancements have seen the application of Deep Learning (DL) techniques rooted in Neural Networks (NN) – a type of machine learning model inspired by the human brain's structure and function, capable of learning from data – to predict hourly passenger flows in metro or train networks Xue et al. (2023); Yang et al. (2020); Yin et al. (2023); Zhu et al. (2019); Fu et al. (2022). The model proposed in this work aligns with this trajectory, leveraging multiple data sources, including mobile phone and smart card data for the metro network data. The occupancy level at any given station and time is contingent upon its preceding occupancy levels and the occupancy patterns at adjacent stations. These spatial correlations are intricate and nonlinear, necessitating a bespoke approach to model the traffic dynamics. Functional characteristics of stations and the day of the week also influence passenger flow, alongside other factors like weather or local events. However, these are not accounted for in the available data.

Drawing inspiration from Cabras (2021), the proposed hybrid model in this work melds a DL model to elicit prior distributions – initial beliefs or assumptions about the probabilities of unseen occupation levels – on future occupancy levels, representing a groundbreaking approach to predicting passenger flows. Unlike extant models, our approach endeavours to reconcile all sources of uncertainty encircling predictions, striving for a calibrated prediction model for metro station occupancy levels. Calibration, defined in terms of accuracy, seeks to align the nominal level of accuracy with both the conditional (to the observed sample) and unconditional accuracy (on a test set). Although DL literature primarily focuses on unconditional accuracy, our model underscores the Bayesian aspect, striving to harmonize all three facets of accuracy depending on the data and statistical model employed, and acknowledged for Bayesian models from Hartigan (1966). Figure 7 encapsulates this harmonization, potentially extending to any model approach amalgamating DL and Bayesian frameworks, and this is our main contribution to the statistical methods. This could also be handled by considering the posterior distribution of NN weights (Gawlikowski et al., 2023) with the so-called Bayesian NN. However, Bayesian NNs – neural networks capable of providing a measure of uncertainty for their predictions – are feasible to estimate for too many simple architectures concerning the one required for the predicting problem considered in this work. Nevertheless, the uncertainty surrounding weights is also implicitly considered in the proposed modelling approach using Dropout – a regularization technique for reducing overfitting by randomly setting to zero some weights during training – in our DL algorithm Gal & Ghahramani (2016); Kingma et al. (2015). Hence, with the required complex architecture, the output can be construed as obtained from a Bayesian DL model.

The subsequent discourse in this paper is organized as follows: Section 2 delineates the available data, Section 3 expounds on the Bayesian spatial-temporal model, and Section 3.3 discusses the results and model interpretation. The use of the model for predicting traffic under unexpected situations is illustrated in Section 4. Final remarks and comparisons with alternative approaches are reserved for Section 5, which also discusses the results of some comparisons. The Appendix reports a sketch of the relevant `R` code used in the paper.

2 The available data and relevant terminologies

This excerpt introduces us to a statistical modelling scenario to understand and predict occupancy levels within urban transit systems, particularly at metro stations, during various hours of the day. Let's break down the technical terminologies and concepts used in this text:

1. **Ordered Categorical Random Variable**: The term $Y_{st} = k \in \{1, \ldots, K\}$ represents an ordered categorical random variable. Here, Y_{st} denotes the occupancy level k at a station s at a specific hour t. Ordered categorical variables are types of categorical

variables where the categories have a meaningful order, but the distances between the categories are not defined. Examples include rating scales such as low, medium, and high.

2. **Indices and Sets**: The indices $s \in \{1, \ldots, S\}$ and $t \in \{1, \ldots, T\}$ indicate that the analysis is being performed across multiple stations and various hours within a specified period T. Each station is uniquely identified by index s, and each hour by index t starting from midnight of September 3, 2021 ($t = 0$).

3. **Quantiles (in particular Terciles)**: Quantiles are points taken at regular intervals from the cumulative distribution function (CDF) of a random variable. In this context, terciles ($K = 3$) are used, which divide the data into three equal parts. This allows for a simpler (although less informative) representation of the data by categorizing the occupancy levels into Low ($k = 1$), Medium ($k = 2$), and High ($k = 3$) based on the distribution of people counts. Models with levels $K > 3$ (or with original counts with no categorization) were explored but unsatisfactory in Calibration, as explained below. This underscores the importance of selecting an appropriate level of categorization for the occupancy levels to ensure the model provides reliable predictions.

4. **Source of People Count**: The source of people count is the data source used to obtain the counts of individuals at a station, which directly defines the variable Y_{st}. This could also encompass the tally of individuals entering and departing from the station or the count of individuals connecting, disconnecting, or residing near a cellular phone antenna in the city.

5. **Dataset Period**: The dataset covers all observed hours from September 3, 2021, to October 30, 2021, for every hour of each day. This comprehensive data collection enables a thorough analysis of occupancy levels across different times and stations.

6. **Calibration**: Calibration in this context refers to adjusting the model to ensure the prediction accuracy required by the analyst.

This excerpt sets up a statistical framework to analyze and predict metro station occupancy levels using an ordered categorical variable representation. By categorizing occupancy levels into three quantiles (terciles), the model aims to provide a simplified yet effective way to predict occupancy levels, making it a potentially valuable tool for urban transit management and planning.

2.1 Descriptive statistics for metro occupancy levels

The dataset under consideration comprises records from a total of $S = 273$ metro stations. This extensive data collection provides a granular insight into the occupancy levels across various metro stations at different times. A visual representation of this data can significantly help in understanding the daily patterns of metro usage. For instance, Figure 1 illustrates the occupancy levels for a typical workday, September 8 2021, at specific hours.

The figure elucidates the prevalent usage of the metro system for commuting purposes, particularly during the peak hours of 8 and 18. This pattern reflects the typical workday commute where individuals travel to and from their workplaces. On the contrary, around lunchtime at hour 14, a notable decrease in metro occupancy is observed, especially in central Beijing. This dip could be attributed to various factors, including a break in work schedules or perhaps the proximity of workplaces to eateries, negating the need for metro travel.

As the night advances, an interesting pattern emerges on the city's outskirts, where a noticeable amount of traffic is observed. This presumably indicates individuals utilizing the metro

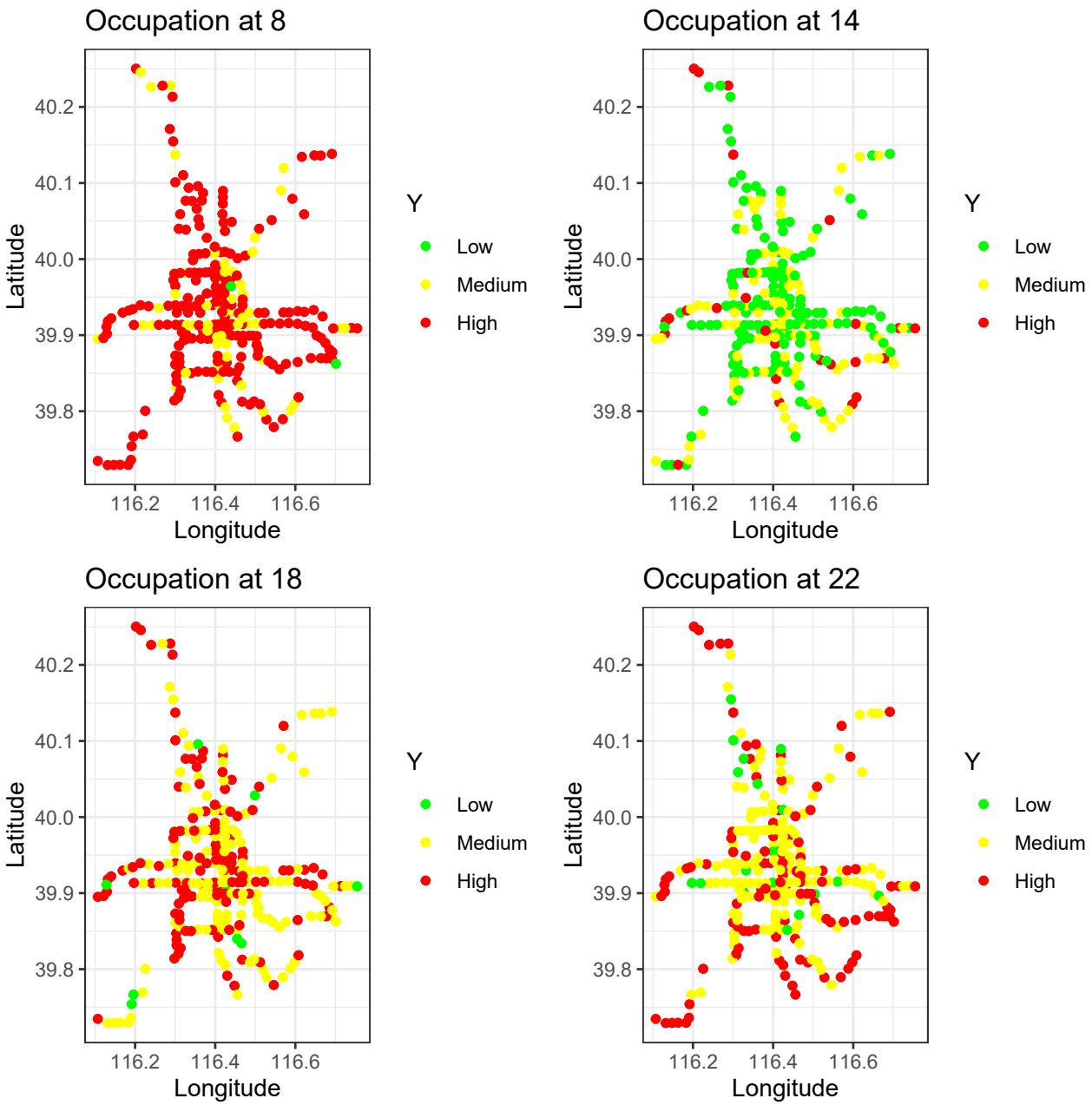

Figure 1: Illustration of occupancy levels Y at different hours on a general working day.

system to return to their residences. The metro system thus serves as a vital conduit facilitating daily commuting, especially in a bustling urban landscape like Beijing. Understanding the patterns of metro occupancy at different hours can aid in better transit planning, ensuring the availability and accessibility of metro services in alignment with the demand, thereby contributing to a smoother urban transit experience.

These patterns and trends gleaned from the data are instrumental in devising effective strategies for managing and accommodating the ebb and flow of metro occupancy, which is crucial for maintaining a reliable and efficient urban transit system. Moreover, the insights derived from this data visualization could serve as a foundation for further analysis and modelling to predict metro occupancy levels, paramount for proactive transit management and enhancing the metro system's overall efficiency and user satisfaction.

2.2 Descriptive statistics for cellular phone antennas levels and their clusters

In the modern era, cellular phone antennas are significant indicators of human activity within a geographical region. This study delves into three particular events associated with each antenna: the number of connections, disconnections, and resident devices. These events are crucial as they reflect the movement and density of individuals near the antenna at a particular time t. Initially, the raw count data for each event, at each antenna at time t, have been encoded into $K = 3$ levels, simplifying the granularity while retaining the essential information regarding human activity. The database encompasses 2265 antennas, with their geographical positions illustrated in Figure 2.

Given the computational constraints of this study, the sheer number of antennas posed a challenge, particularly due to memory overflow issues on the available GPUs. A hierarchical clustering approach was employed using the complete linkage method on the Euclidean distances among antennas to mitigate this. This method grouped the antennas into 50 distinct clusters, reducing the computational load while preserving the spatial relationships among antennas. As shown in Figure 2, the centres of these clusters serve as representative points, encapsulating the data of all antennas within the respective cluster. The data within each cluster was aggregated by summing the number of connected, disconnected, and resident cellular terminals and subsequently calculating the K levels over these sums.

On the same workday, as depicted in Figure 1, the levels of disconnections one hour prior are illustrated in Figure 3.

These levels are insightful as they reflect the dynamism of individuals' movement within and out of each cluster. As Figure 3 shows, the disconnections exhibit a consistent pattern, albeit higher during morning and noon hours. This pattern indicates the typical workday commute, with a surge in disconnections in the morning as individuals depart for work and around noon, possibly for lunch breaks. While the correlation with metro occupancy might be low—since only a subset of individuals use the Metro for city transit—the past values of these levels are not directly utilizable for predicting passenger counts. However, the levels associated with cellular activity (connected/disconnected/resident) shed light on metro occupancy levels, especially at stations close to clusters with higher movement activity. This indirect insight could be instrumental in understanding and predicting metro occupancy levels, thereby contributing to more effective transit management and planning in urban settings.

2.3 split of the dataset in train and testing

The dataset at our disposal spans two months, equating to 60 days or 1440 hours. In the context of machine learning and statistical modelling, the size of the dataset is a critical factor that influences the model's ability to learn and generalize well to unseen data. However, the adequacy of the dataset size is relative to the complexity of the task at hand and the nature of the models being employed.

Deep Learning (DL) is renowned for its capacity to model complex relationships in data, especially when the available dataset is large. However, although seemingly substantial, our dataset of 1,440 hours does not constitute a large sample size for a one-shot Deep Learning analysis. The term "one-shot" here refers to training the model in a single iteration over the entire dataset. This is especially pertinent as some observations must be set aside for model validation, a standard practice to evaluate the model's performance and generalization ability on unseen data. Conversely, when viewed from the lens of Bayesian modelling, which often operates on the principles of probability and uncertainty, this dataset size is considerably large and complex for a parametric or semi-parametric Bayesian model like that in Cabras & He

(2023). Bayesian models are known for working efficiently with smaller datasets compared to their DL counterparts, as they incorporate prior knowledge and uncertainty in their framework.

For modelling, a structured approach is adopted to ensure a robust evaluation of the models. The initial 50 days of data are designated as the training set for the DL model (refer to Section 3.1), which is a common practice to allocate a majority of the data for training to allow the model to learn the underlying patterns in the data. Subsequently, the following 10 days are allocated for testing the model to evaluate its performance on new, unseen data.

A further breakdown is done for the Bayesian calibration model, where days 51-55 (comprising 5 days) are utilized for estimating the Bayesian calibration model (refer to Section 3.2), with the remaining 5 days earmarked for testing the calibration model. This arrangement is meticulous to ensure that the final 120 hours (last 5 days) remain untouched by any training process, serving as the actual testing set for model evaluation. This segregation is crucial to ascertain the models' ability to generalize to new data, a critical aspect of machine learning and statistical modelling.

The Calibration mentioned above graph, a pivotal part of evaluating the models is reported using the data from these final days, as illustrated in Figure 7. This figure encapsulates the essence of Calibration, which measures how well the predicted probabilities align with the actual outcomes, providing a nuanced understanding of the models' performance and reliability.

3 Description of the model used to estimate Metro's occupancy levels.

We introduce the lagged matrix denoted by X_l, characterized as a binary matrix comprised solely of entries being 0s and 1s. The dimensions of this matrix will be specified in the subsequent discussion. The essence of X_l is to encapsulate lagged observations at the preceding hours $t-1, t-2, \ldots, t-l$, where $l = 4$. This matrix is conceived to embody the filtration $\mathcal{F}_l$ of the process Y_{st} up to $l = 4$ hours before, with our focal point estimator of the posterior distribution of $\Pr(Y_{st} = y|\mathcal{F}_l) \approx \Pr(Y_{st} = y|X_l)$.

All the variables in this analysis have been encoded into dummy variables corresponding to their original categories. This encoding lacks constraints, akin to constructing a typical design matrix in regression analysis. The matrix X_l is an exemplar of such encoding, encompassing past occupancy levels Y for all stations, levels of individuals entering/connecting and departing/disconnecting from stations and cellular phone antennas, along with other covariates such as the specific metro station s, the day of the week (ranging from Monday to Sunday), the hour of the day t (ranging from 0 to 23 and encoded into 24 columns of X_l), and the current operational status of the station at time t: open or closed. These variables are encapsulated within the columns of X_l. For instance, envisioning a scenario where the aim is to predict Y_{st} utilizing all available information four hours prior ($l = 4$), the matrix X_l would incorporate, for each hour up to four hours before t: the levels of departures, entries, and occupancy of all stations including station s, alongside analogous data for mobile phone antennas, the hour of the day, and the day of the week and so on. This conglomeration culminates in a sizable matrix, thereby furnishing a substantial dataset to be employed in the Deep Learning (DL) model delineated later.

The endeavour of predicting hourly occupancy is orchestrated via a two-tiered model strategy: initially, a Deep Learning model, specifically a Feed Forward Neural Network (FFNN), is employed to spatially and temporally correlate all available information regarding passengers' occupancy and movements across all metro stations and cellular phone locations. The FFNN is a class of neural networks well-suited for learning from multidimensional data and capturing complex relationships. The subsequent model is constructed on the probability of

occupancy levels gleaned from the DL model, with a calibrated approach accounting for the prediction uncertainty inherent in the DL model and the specific hour under prediction. This Calibration is crucial for enhancing the reliability and interpretability of the predictions. The model employed here is a proportional odds model coupled with a random effect of hours. It is estimated utilizing an advanced statistical technique called Integrated Nested Laplace Approximation (INLA). INLA is a robust method for Bayesian inference that facilitates fast and accurate approximations. Notably, in the latter model, the definition of k based on quantiles becomes immaterial, as the variable is treated as an ordinal categorical variable in the Bayesian model articulated later. This configuration affords the flexibility for alternative definitions of k diverging from the conventional utilization of quantiles, thereby providing a robust framework for analyzing the occupancy levels in a statistically rigorous and computationally efficient manner.

3.1 The Feed Forward Neural Network used in the first step

The DL model is engineered to regress the target variable Y_{st} on the predictor matrix X_l, which is a vast matrix with a sprawling count of 8,793 columns. This massive input dimensionality culminates in an FFNN (Feed Forward Neural Network) that boasts a staggering one million trainable parameters, precisely 1,093,059. The entire training sample comprises 319,956 statistical units, underscoring the sheer volume of data fed into the model. The grandiose dimensionality of this model underscores the unfeasibility of employing Bayesian Neural Networks (Bayesian NNs) in this particular applied setting, mainly due to the computational and memory demands that Bayesian approaches entail.

We explore and train four distinct FFNN models, each tailored to predict different time horizons: 1, 2, 3, and 4 hours ahead. The architectural blueprint of all these models is meticulously delineated in Figure 4, showcasing the structure of the neural network employed.

As depicted in the figure, the architecture exhibits dense and Dropout layers alternating. This design choice is imperative to avert overfitting, a common problem in machine-learning models with large parameter spaces. The dropout layers serve to distil the pertinent signal in the matrix X_l for the arduous task of predicting Y_{st}. The loss function employed to train the model is binary cross-entropy, a choice driven by the fact that the levels are deemed unordered in the DL models, and hence, a binary classification loss is apt. The model undergoes a training regimen over 20 optimization steps, and the trajectory of the loss function on the validation set is depicted in Figure 5.

The plateauing of losses, which is reached at around 15 steps when considering the confidence bounds in Figure 5, signifies a state of model learning's stabilization, where further optimization could potentially lead to overfitting —a scenario where the model learns the noise in the training data, compromising its generalization capability on unseen data.

Upon training, the model computes the estimated probability levels $\tilde{p}^{(1)}$, $\tilde{p}^{(2)}$, and subsequently $\tilde{p}^{(3)} = 1 - \tilde{p}^{(1)} - \tilde{p}^{(2)}$. As elucidated by Polson & Sokolov (2017), these probability levels represent the maximum a posteriori of the probability levels given the sample. These probabilities can be harnessed to estimate a given model's occupancy level: $\tilde{Y}_{st} = \arg\max_k \tilde{p}^{(k)}$, i.e., the level with the highest probability. Typically, $\tilde{Y}_{st}$ is juxtaposed with the observed y_{st} to derive performance metrics like accuracy, which is the Ratio of $\tilde{Y}_{st} = y_{st}$ across all y_{st} or across all $y_{st} = k$ which is the conditional accuracy to a specific occupation level k. We, however, look to the more interesting analysis done with the ROC (Receiver Operating Characteristic) and the relative Area Under the ROC Curve (AUC).

The ROC curve is a graphical representation that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

The TPR is also known as sensitivity, recall, or probability of detection, while the FPR is also known as the fall-out or probability of false alarm. The ROC curve showcases the trade-off between true positive and false positive rates, providing insight into the model's capability to discriminate between the positive and negative classes across different thresholds.

The AUC is a scalar metric that quantifies the overall ability of the model to discriminate between positive and negative classes across all thresholds. An AUC of 1.0 signifies a perfect model, while an AUC of 0.5 indicates a model with no discrimination capability akin to random guessing. The AUC is a robust measure as it is invariant to the scaling and balance of the classes. In this study, a higher AUC across different hours ahead and occupation levels underscores the model's adeptness in discriminating between different occupancy levels, thereby furnishing invaluable insights for real-time and future metro occupancy prediction and planning.

In the grand scheme of this study, the DL model serves as a stepping stone rather than the final predictive model for metro occupancy. It aids in constructing the substrate for the subsequent Bayesian model. Nevertheless, evaluating the encapsulated information within the FFNN is of merit, particularly if the DL prediction accuracy is high. To this end, we contemplate the ROC curves pertinent to the three levels of occupancy across the three models, showcased in Figure 6.

The Illustration sheds light on the superior estimation prowess of the model for low and high occupation levels. This is discerned from the corresponding areas under the ROC curves, which surpass the 80% threshold, as meticulously tabulated in Table 1. The crux of the uncertainty in prediction lies in the medium occupancy level. Despite this, the model discriminates between low and high occupancy levels, which is pivotal for devising and fine-tuning train schedules to cater to the dynamic passenger load.

Table 1: Area Under the ROC for the test set conditionally to the number of hours ahead predictions and the predicted occupation level.

	Occupation Level		
Hour Ahead	**Low**	**Medium**	**High**
1	0.925	0.802	0.840
2	0.920	0.810	0.842
3	0.920	0.803	0.837
4	0.917	0.819	0.839

Moreover, it is noteworthy that the model tasked with predicting 4 hours ahead is only marginally less accurate than the one predicting merely an hour ahead. This modest deterioration in accuracy over a longer prediction horizon is a testament to the model's robustness.

3.2 The Bayesian Proportional Odds Model (POM)

The objective of this secondary model is to furnish predictions of metro occupancy levels along with their associated uncertainties, culminating in the final and calibrated prediction model.

The Bayesian framework employs a Proportional Odds Model (POM) regression, characterized by parallel (to the K occupancy levels) fixed effects and an independent random effect about hours. Formally, the stochastic delineation of the proposed POM is encapsulated in equations (1) through (7):

$$
\begin{aligned}
Y_{st}|p_{st}^{(1)},\ldots,p_{st}^{(K)} &\sim \text{Multinomial}(p_{st}^{(1)},\ldots,p_{st}^{(K)}),\ p_{st}^{(k)} = \Pr(Y_{st}=k) && (1)\\
p_{st}^{(k)} &= F(k) - F(k-1)\\
F(k) &= \Pr(Y_{st}\le k) = \frac{\exp(\gamma_{st}^{(k)})}{1+\exp(\gamma_{st}^{(k)})}\\
\gamma_{st}^{(k)} &= \alpha_k - \eta_{st},\ \alpha_0 = -\infty < \alpha_1 < \ldots \alpha_K = \infty && (2)\\
(\alpha_1,\ldots,\alpha_{K-1}) &\sim \text{DP}(\alpha=3) && (3)\\
\eta &= \beta_1\tilde{p}_{st}^{(1)} + \beta_2\tilde{p}_{st}^{(2)} + h_t && (4)\\
\beta_1,\beta_2 &\sim \pi(\beta)\propto 1,\ \text{ for } \beta_1,\beta_2\in\mathbb{R}^2 && (5)\\
h_t|\tau &\sim \text{Normal}(0,\tau) && (6)\\
\tau &\sim \text{Gamma}(1, 0.00005) && (7)
\end{aligned}
$$

In the above equations, Multinomial$(\cdot)$ denotes the Multinomial distribution, a generalization of the binomial distribution suitable for categorical data with more than two categories. It models the probabilities of observing counts among multiple categories. The DP(α) represents the Dirichlet Process, a stochastic process used in Bayesian non-parametric to define priors and hence complex models, with a concentration parameter α set to 3. The Normal$(0,\tau)$ refers to the conventional normal distribution with mean 0 and precision τ, where precision is the reciprocal of the variance. The Gamma(a,b) symbolizes the Gamma distribution, a two-parameter family of continuous probability distributions, with mean a/b.

Due to the presence of the DP prior, this is a semiparametric model that accurately treats Y_{st} as an ordered categorical variable, liberating the analysis from a precise delineation of Y_{st} based on quantiles. Specifically, equation (3) models the cumulative distribution of Y_{st} via the prescription of the cumulative odds model. This modelling strategy is viable at this juncture since the process filtration represented in X_l has been previously accounted for in $\tilde{p}_{st}^{(k)}$, stemming from the output of the DL mentioned above model, which is incorporated in the linear predictor (4).

Moreover, this Bayesian model accommodates an independent random effect for the distinct hours to be predicted, which is pivotal as the endeavour aims to assess the reliability of model predictions across varying hours. For instance, the implication of a reliable or unreliable prediction at 8 a.m. (high traffic) vastly differs from that at 10 p.m. (low traffic). The model outlined in equations (1) through (7) virtually lacks tuning parameters, as the priors (3), (5), and (7) can be perceived as weakly informative or default priors, particularly in the case of the location parameters β_1 and β_2. This aspect substantially mitigates the necessity for robust analysis concerning the priors in the model.

The posterior distribution was estimated for the four lagged prediction hours: one hour ahead up to four hours ahead, resulting in four posterior distributions of model parameters. These posteriors were obtained by evaluating the likelihood during the calibration period spanning days 51 to 55, with the model being tested on the subsequent 5 available days as delineated in 2.3.

The ensuing section elucidates the results in terms of prediction and model interpretation.

3.3 Results

The ultimate prediction generated by the Bayesian model for each s and t is delineated by the posterior distribution of parameters $p_{st}^{(k)}$ as expressed in (1). The prediction for Y_{st} will be designated by the level k^*, under the circumstance where this level manifests the highest

posterior probability, and provided that the uncertainty, as depicted by the posterior equal tails Credible Intervals (CI) of $p_{st}^{(k^*)}$ at a nominal level $1-\delta$ (for instance, 75%, $1-\delta = 0.75$), ensures that k^* retains the position of the level with the highest posterior probability, i.e., the CI of $p_{st}^{(k^*)}$ does not intersect (in its lower bound) with the other CI of $p_{st}^{(k)}$ for levels $k \neq k^*$. This approach facilitates a prediction probable to materialize at least with a probability of $1-\delta$. A prediction is withheld if the uncertainty falls below the nominal δ.

A Credible Interval (CI) is a range of values derived from the posterior distribution, within which an unobserved parameter value falls with probability $1-\delta$. It is a Bayesian analogue to the confidence intervals in frequentist statistics but with a different interpretation, as the credible intervals are conditioned to the observed data. In contrast, confidence intervals are interpreted in the light of infinite replications (and hence for an infinite set of datasets) of the same experiments.

A significant outcome of the methodology elucidated in this discourse is illustrated in Figure 7.

For each specified δ, the probability of accurately predicting Ys in the 5-day test set, marginally about stations and hours, is approximated with the observed accuracy in the test set. Figure 7 represents this on the vertical axis. A calibrated model will exhibit an agreement between the nominal level $1-\delta$ (the conditional inference) and the accuracy level in the test set (unconditional inference), predominantly apparent at levels of $1-\delta$ spanning 70% to 80%.

It can thus be inferred that, conditional to the data, the model is calibrated to forecast hour-level occupancy with an accuracy range of 70% - 80%, provided that the same nominal level is employed in the analysis. Nominal levels exceeding (or falling below) the 70% - 80% range will yield lesser (or greater) accuracy than the nominal one. While an analyst might opt for exceedingly high nominal levels, such as $1-\delta = 0.95$, besides being unreliable, predictions may not demonstrate such a level of precision. This is also to state that the analysis is intended to not communicate as reliable any prediction that does not satisfy the nominal level or report them with a pertinent warning.

The proportion of predictions meeting the stipulated nominal levels is represented by the colour scale in Figure 7. It can be asserted that the model is calibrated at 70% - 80%, and the analyst can utilize approximately 68% of the predictions proffered by the model.

Given the available dataset, figure 7 accurately depicts the model's capability to address the hour occupancy level prediction problem. A more substantial dataset is anticipated to enhance the representation in Figure 7 according to the theory of Bayesian CI developed in Hartigan (1966). This assesses that CI will have a frequentist interpretation as the sample size increases and that the achievement between nominal coverage $1-\delta$ (accuracy here as the Y_{st} is discrete) and the frequentist one occurs at a higher speed (in term of accumulation of sample size) than frequentist procedures.

Regarding the interpretation of the latter Bayesian model, an examination of the posterior distribution of the fixed effects β_1 and β_2 showcased in Figure 8, reveals that these effects are positive, as anticipated. The positive coefficients of $p_{st}^{(1)}$ and $p_{st}^{(2)}$ engender a decline in the probability of high occupancy levels.

A higher degree of uncertainty is observed on the coefficient of $p_{st}^{(2)}$, denoted as β_2, compared to $p_{st}^{(1)}$, namely β_1. This outcome echoes the inherent challenge in predicting the medium occupancy level, as depicted in Figure 6 and Table 1.

The posterior distribution of the hour random effect is illustrated in Figure 9. Random effects model the correlation of observations within the same group and account for unobserved heterogeneity. Significant deviations from zero in the posterior distribution of the random effect for certain hours indicate these hours have a pronounced effect on the prediction of Y_{st}.

Figure 9 exhibits a pronounced clustering of hour-random effects about the opening and closing times of the Metro. Nonetheless, certain hours markedly influence the prediction of

Y_{st}, specifically hours 6, 7, 10, 13, 16, and 19. The posterior distribution of these hours significantly deviates from zero. With a negative value, it intimates that the probability of high occupancy levels is generally inclined to escalate during these hours. This insight is integral to the recalibration of the DL model for the specific hour predicted. The prominence of such an effect amplifies in the model employed for the 4 hours ahead prediction, as delineated at the bottom of Figure 10.

The relevance of the random effect in η is affirmed by the posterior distribution of the precision parameter τ in Figure 10. A diminutive precision value signifies an augmented variability among random effects emanating from a higher internal precision of the random effect.

3.4 Interpretation of the model: an analysis of the effect of the mobile antenna on metro traffic

This investigation endeavours to discern the impact of mobile phone antennae on metro traffic within the framework of the proposed model. Such exploratory endeavours and their ensuing development are commonplace in scrutinizing DL models, as delineated in Lundberg & Lee (2017); Samek et al. (2017).

Investigations like these delve into the relationships between different variables within a model to gain insights or validate the model's assumptions and predictions.

As previously discussed, there exists no direct correlation between the occupancy levels at antennae and those at metro stations: only a segment of the populace utilizes the Metro, and the objective is to ascertain the estimated size of this segment as estimated by the model.

To undertake this, we altered all antenna levels of residents, connections, and disconnections from low to high throughout October 29, 2021 (a Friday), while maintaining all other observed variables in their original state.

The resultant effect is gauged by approximating the posterior distribution of log differences (the log-risk Ratio, $\log(RR)$) in the posterior mean probability of high occupancy levels $p_{st}^{(3)}$ (the reference level in this analysis) when all three antenna levels (resident, connected and disconnected) were high as opposed to when they were designated as low. The log-risk Ratio ($\log(RR)$) is a measure used to compare the probability of a particular outcome between two groups, in this case, the probability of high occupancy levels between high and low antenna activity levels.

The findings are depicted in Figure 11. The $\log(RR)$ have been reported only if the estimations of $p_{st}^{(3)}$ under high and low conditions where both reliable at the level $1 - \delta = 80\%$. The reported values for those cases are the posterior mean of $\log(RR)$.

As inferred from Figure 11, the marginal alterations are zero at the commencement of operations. Yet, there is an anticipated decrease to 50% in the probability of high occupancy during the early hours (9-10) and a contrasting surge around (12-13). These movements of people all over the city may be interpreted as movement out of the city (at the beginning of the day) through other means that are not the Metro network and into the city (at noon) and are typically observed during the beginning of holiday periods. A general downtrend in metro traffic is expected towards the day's closure. Per the station-conditional alterations, these shifts permeate the network and do not concentrate within a particular zone.

4 A particular use of the model for planning emergencies or unexpected situations

In this section, we explore the planning for emergencies or unexpected scenarios related to metro passengers' behaviour by utilizing the fitted model and employing the same metric as

delineated in Section 3.4 to assess the impact of certain behaviours.

We illustrate, for instance, the predictions rendered by the model at a designated hour when a particular station starts to witness crowding and when a significant aggregation of mobile phones transpires in a specific urban zone.

Here, an illustration is made regarding how the model predicts metro passenger behaviour at a specific hour under two different scenarios:

1. When a particular metro station experiences crowding.

2. When a significant number of mobile phones (indicative of a large gathering of people) are detected in a specific urban zone.

These scenarios exemplify how external factors might influence metro passenger behaviour and how the model can be utilized to plan for such eventualities.

The aim is to leverage the model to provide insights into how these scenarios could impact metro operations and, in turn, aid in better planning and response to ensure the smooth operation of the metro system and the safety of its passengers.

4.1 Crowded Train Stations

In this scenario, a hypothetical situation is being considered where the occupancy levels in three specific metro stations (Beijingxi, Huilongg, and Tiantongy) in Beijing experience a surge from low to high between 8 to 10 a.m. on October 22. The objective is to understand the ripple effect this sudden surge in occupancy levels may have on the metro network.

Similar to Section 3.4, the effect is gauged by the distribution of log differences (the log-risk Ratio, $\log(RR)$) in the probability of high occupancy levels when the occupancy level was high against when it was low. All other data remains constant as collected for that particular day. The results are showcased in Figure 12.

In Figure 12 (top), the results of this hypothetical scenario are presented, showcasing how the sudden surge in occupancy levels at these stations between 8 to 10 a.m. affects the probability of high occupancy levels across the metro network.

The figures illustrate that the occupancy levels at these stations influence the network in the ensuing hours; specifically, an escalation of around 3% in the marginal probability of high occupancy levels is anticipated just at 9, and the impact will diminish as the hours progress. Initially, this change will be confined to the proximate stations at 10. Still, it will generally spread to the remainder of the line as the day advances, only to vanish in the evening since no significant alterations are foreseen towards the day's end, around 18.

This analysis suggests that sudden changes in occupancy levels at the mentioned stations during peak hours could temporarily affect the occupancy probabilities across the metro network. The model anticipates a peak effect at 9 a.m., with a ripple effect diminishing as the day progresses. By evening, around 6 p.m. (18), the effects of the morning's occupancy surge are no longer significant, returning the network to its usual occupancy probabilities. The graphical representation in Figure 12 visually elucidates these impacts, offering insights for better metro management and emergency response planning.

4.2 Tourists' concentration

This investigation seeks to grasp the impact of a shift in mobile phone antennae occupancy levels (including connection, disconnection, and resident levels) on metro traffic when such a change occurs in the city centre. For the 20 most centrally located antenna clusters, the occupancy levels were altered from low to high throughout October 29, while all other observed variables remained unchanged. The outcomes are presented in Figure 13.

Figure 13 showcases the results of this simulation, illustrating how the change in antennae occupancy levels influences the metro traffic at different hours throughout the day. From Figure 13, we can see that the marginal changes are substantial at the day's onset and diminish as the hours advance. Initially, a surge in metro traffic is anticipated at 12 and 18, whereas a notable decline in traffic is projected during hour 14. This pattern, particularly on Fridays, is typically attributed to tourist activity. Indeed, at 22, a rise in traffic is expected in the city's peripheries coupled with a significant reduction of traffic within the central city stations.

5 Discussion and concluding remarks

The analytical strategy delineated in this manuscript facilitates the forecasting of hourly occupancy levels within Beijing's metro system, along with a trustworthy measure of uncertainty surrounding such predictions. The concept of uncertainty here is pivotal as it provides a range within which the actual values are likely to lie, enhancing the predictions' reliability.

This uncertainty has been appraised unconditionally to the available data, particularly on a test set, as is customary within Machine Learning literature, while also provided conditionally in the typical Bayesian framework based on the available data. This bifurcated unconditional and conditional assessment enables a more robust evaluation of the model's performance, adhering to conventional machine learning practices and Bayesian statistical principles, which emphasize probabilistic interpretations and are particularly adept at handling uncertainty.

The dual assessment enables more dependable predictions, which could be enhanced by employing a period extending beyond the two months utilized for modelling in this study. Extending the data collection and analysis time frame could reveal more intricate patterns and dependencies, fostering improved predictive accuracy. This potential enhancement could manifest in the form of heightened prediction accuracy and the employment of a broader spectrum of occupancy levels beyond the rudimentary low, medium, and high classifications proposed herein. Broadening the categories of occupancy levels could provide a more granular insight into the occupancy dynamics, thereby enriching the model's predictive capability.

The consideration of additional levels and the original counts has been undertaken in preceding iterations of this study; however, the resultant AUC hovered around 30% at some level for $K = 4$, and the R^2 of the DL model approximated less than 10% when fitted on the original counts. This points to a certain limitation in the granularity of the model when dealing with more nuanced categorizations or the original count data.

None of the methodologies that could potentially be adapted to the dataset has demonstrated superiority over the one considered herein, at least for the data at hand. This statement underscores the robustness and suitability of the chosen model for the current dataset amidst the explored alternatives.

References

Cabras, S (2021), 'A bayesian-deep learning model for estimating covid-19 evolution in spain,' *Mathematics*, **9**(22), p. 2921.

Cabras, S & He, S (2023), 'A bayesian spatial–temporal model for predicting passengers occupancy at beijing metro,' *Spatial Statistics*, **55**, p. 100754.

Davis, RA, Fokianos, K, Holan, SH, Joe, H, Livsey, J, Lund, R, Pipiras, V & Ravishanker, N (2021), 'Count time series: A methodological review,' *Journal of the American Statistical Association*, pp. 1–15.

Eric Lin, AZ, Jinhyung D. Park (2017), 'Real-time bayesian micro-analysis for metro traffic prediction,' *the 3rd ACM SIGSPATIAL Workshop*, **4**, pp. 1–4.

Fu, Q, Liu, R & Hess, S (2012), 'A review on transit assignment modelling approaches to congested networks: a new perspective,' *Procedia-Social and Behavioral Sciences*, **54**, pp. 1145–1155.

Fu, Q, Liu, R & Hess, S (2014), 'A bayesian modelling framework for individual passenger's probabilistic route choices: a case study on the london underground,' Tech. rep., National Academies, London.

Fu, X, Zuo, Y, Wu, J, Yuan, Y & Wang, S (2022), 'Short-term prediction of metro passenger flow with multi-source data: a neural network model fusing spatial and temporal features,' *Tunnelling and Underground Space Technology*, **124**, p. 104486.

Gal, Y & Ghahramani, Z (2016), 'Dropout as a bayesian approximation: Representing model uncertainty in deep learning,' in *international conference on machine learning*, PMLR, pp. 1050–1059.

Gawlikowski, J, Tassi, CRN, Ali, M, Lee, J, Humt, M, Feng, J, Kruspe, A, Triebel, R, Jung, P, Roscher, R et al. (2023), 'A survey of uncertainty in deep neural networks,' *Artificial Intelligence Review*, pp. 1–77.

Han, Y, Wang, S, Ren, Y, Wang, C, Gao, P & Chen, G (2019), 'Predicting station-level short-term passenger flow in a citywide metro network using spatiotemporal graph convolutional neural networks,' *ISPRS International Journal of Geo-Information*, **8**(6), p. 243.

Hartigan, J (1966), 'Note on the confidence-prior of welch and peers,' *Journal of the Royal Statistical Society: Series B (Methodological)*, **28**(1), pp. 55–56.

Jiao, P, Li, R, Sun, T, Hou, Z & Ibrahim, A (2016), 'Three revised kalman filtering models for short-term rail transit passenger flow prediction,' *Mathematical Problems in Engineering*, **2016**.

Jie, W, Haitao, J & Fengjun, J (2018), 'Investigating spatiotemporal patterns of passenger flows in the beijing metro system from smart card data,' *Prog. Geogr*, **37**, pp. 397–406.

Kingma, DP, Salimans, T & Welling, M (2015), 'Variational dropout and the local reparameterization trick,' *Advances in neural information processing systems*, **28**.

Lundberg, SM & Lee, SI (2017), 'A unified approach to interpreting model predictions,' *Advances in neural information processing systems*, **30**.

Petris, G (2010), 'An r package for dynamic linear models,' *Journal of Statistical Software*, **36**(1), pp. 1–16.

Petris, G, Petrone, S & Campagnoli, P (2009), 'Dynamic linear models,' in *Dynamic Linear Models with R*, Springer, pp. 31–84.

Polson, NG & Sokolov, V (2017), 'Deep Learning: A Bayesian Perspective,' *Bayesian Analysis*, **12**(4), pp. 1275 – 1304, doi:10.1214/17-BA1082.

Rodriguez, Y, Pineda, W & Olariaga, OD (2020), 'Air traffic forecast in post-liberalization context: a dynamic linear models approach,' *Aviation*, **24**(1), pp. 10–19.

Roos, J, Bonnevay, S & Gavin, G (2016), 'Short-term urban rail passenger flow forecasting: A dynamic bayesian network approach,' in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, pp. 1034–1039.

Samek, W, Wiegand, T & Müller, KR (2017), 'Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models,' *arXiv preprint arXiv:1708.08296*.

Xue, Q, Zhang, W, Ding, M, Yang, X, Wu, J & Gao, Z (2023), 'Passenger flow forecasting approaches for urban rail transit: a survey,' *International Journal of General Systems*, **52**(8), pp. 919–947.

Yang, J, Dong, X & Jin, S (2020), 'Metro passenger flow prediction model using attention-based neural network,' *IEEE Access*, **8**, pp. 30953–30959.

Yin, D, Jiang, R, Deng, J, Li, Y, Xie, Y, Wang, Z, Zhou, Y, Song, X & Shang, JS (2023), 'Mt-mgnn: Multi-time multi-graph neural network for metro passenger flow prediction,' *GeoInformatica*, **27**(1), pp. 77–105.

Zhao, Y & Ma, Z (2022), 'Naïve bayes-based transition model for short-term metro passenger flow prediction under planned events,' *Transportation Research Record*, p. 03611981221086645.

Zhu, K, Xun, P, Li, W, Li, Z & Zhou, R (2019), 'Prediction of passenger flow in urban rail transit based on big data analysis and deep learning,' *IEEE Access*, **7**, pp. 142272–142279.

Appendix - Code

Here is a sketch of the relevant R code used in the paper.

```
library(verification)
library(readr)
library(dplyr)
library(reshape2)
library(ggplot2)
library(gganimate)
library(keras)

# Value of K: number of levels of occupancy
nq=3
lq=1:nq

batch_size <- 128 # Batch size at each iteration
epochs <- 20 # Optimization Steps
hour.before=1 # Hour ahead in the prediction
lag=4 # The value of l
test.days=20 # Number of test days (for Bayesian POM and te
nantclust=50 # Number of Antenna Clusters


# Metro occupancy data set and all covariates
metro <- read.csv("metrodata.csv")

# antenna data set
antenna <- read.csv("antenna.csv")

# Cluster of antennas
coord=antenna %>%
group_by(antenna) %>%
summarise(lat=mean(lat),lng=mean(lng)) %>%
ungroup()

ii=match(antenna$antenna,coord$antenna)
antenna$lat=coord$lat[ii]
antenna$lng=coord$lng[ii]

clantenna=antenna %>%
select(antenna,lat,lng) %>%
unique() %>%
mutate(cluster.antenna=cutree(hclust(dist(cbind(lat,LNG))),
k = nantclust)) %>%
select(antenna,cluster.antenna)
ii=match(antenna$antenna,clantenna$antenna)
antenna$clantenna=clantenna$cluster.antenna[ii]

## Definition of K levels for Metro and Antennas
```

```
antenna=antenna %>%
select(clantenna,htime,resident,arrival,departure) %>%
group_by(clantenna,htime) %>%
summarize(resident=sum(resident),
arrival=sum(arrival),
departure=sum(departure)) %>%
ungroup() %>%
group_by(htime) %>%
mutate(resident=findInterval(resident,
quantile(resident,p=seq(0,1,length=nq+1)),
all.inside = TRUE)) %>%
mutate(arrival=findInterval(arrival,
quantile(arrival,p=seq(0,1,length=nq+1)),
all.inside = TRUE)) %>%
mutate(departure=findInterval(departure,
quantile(departure,p=seq(0,1,length=nq+1)),
all.inside = TRUE)) %>%
ungroup() %>%
arrange(htime)

metro=metro %>%
arrange(htime,station) %>%
group_by(station) %>%
mutate(total=findInterval(total,
quantile(total,p=seq(0,1,length=nq+1)),
all.inside = TRUE)) %>%
mutate(enter=findInterval(enter,
quantile(enter,p=seq(0,1,length=nq+1)),
all.inside = TRUE)) %>%
mutate(depar=findInterval(depar,
quantile(depar,p=seq(0,1,length=nq+1)),
all.inside = TRUE)) %>%
ungroup()

## The FFNN
model <- keras_model_sequential() %>%
layer_dense(units = 64,  input_shape = c(ncol(X_train)),
activation = "relu",use_bias = TRUE) %>%
layer_dropout(rate = 0.2) %>%
layer_dense(units = 64,
activation = "relu",use_bias = TRUE) %>%
layer_dropout(rate = 0.2) %>%
layer_dense(units = ncol(Y_train), activation = "softmax")

model %>% compile(
loss = 'binary_crossentropy',
optimizer = optimizer_adam(),
metrics = c('accuracy'))

history <- model %>%
```

```
fit(X_train, Y_train,
epochs = epochs,
batch_size = batch_size,
validation_split = 0.2,
verbose=2)

## Predictions and AUC calculation
preds.test <- model %>% predict(X_test)
for(i in 1:nq) cat("level:",i,"␣A=",
roc.area(Y_test[,i],preds.test[,i])$A,"␣\n")
```

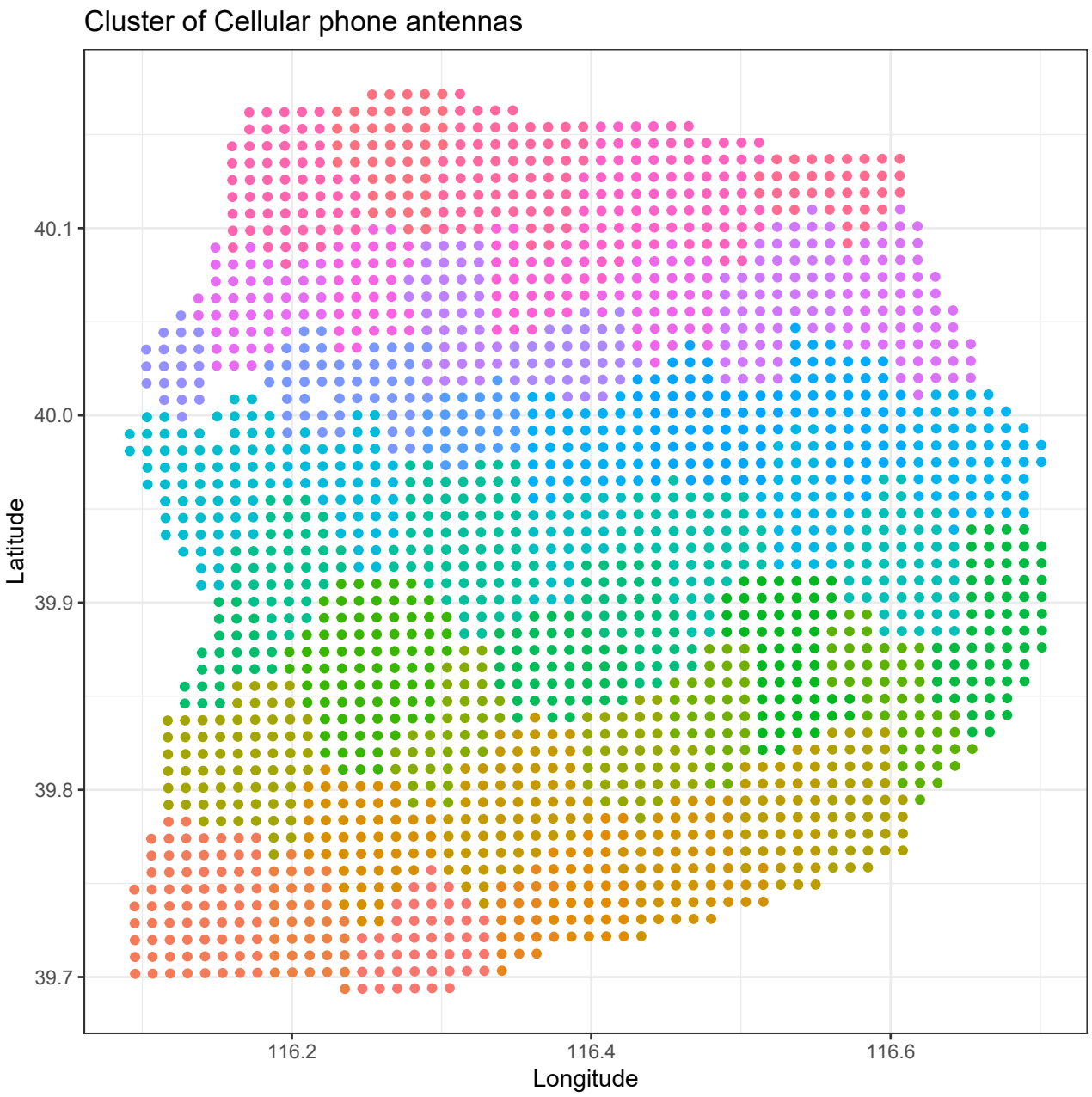

Figure 2: Depiction of the original positions of cellular phone antennas and their cluster associations. The cluster centre serves as the virtual representative antenna for that cluster.

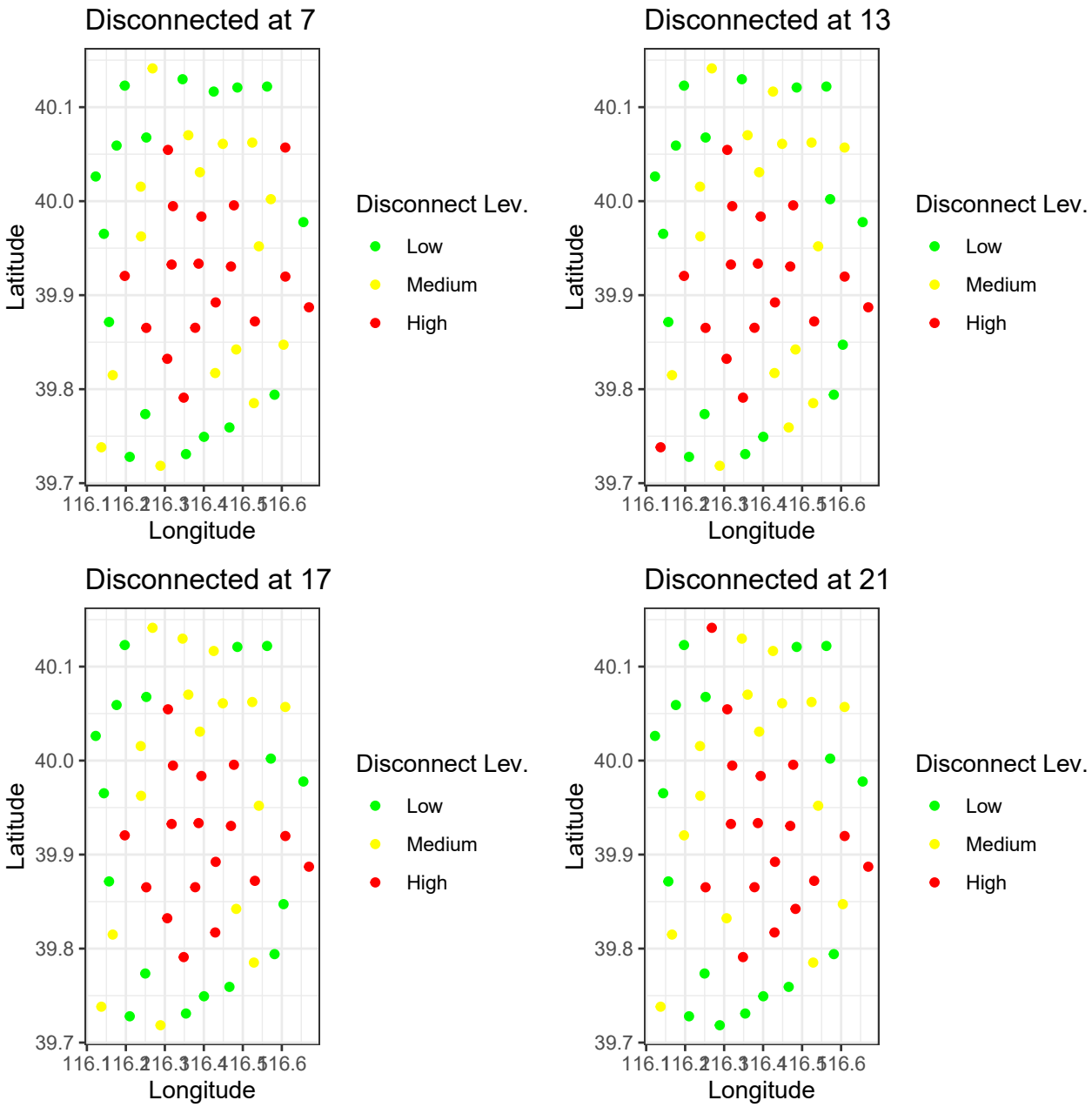

Figure 3: On the same day as shown in Figure 1, we present the levels of disconnected individuals within the antenna clusters (points represent cluster centres) one hour before the time reported in Figure 1.

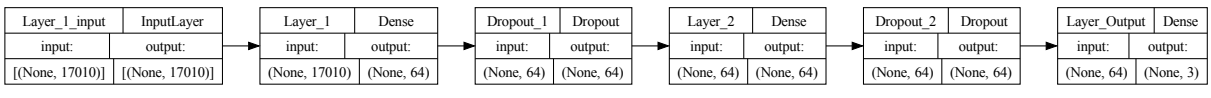

Figure 4: Structure of the deployed Feed Forward neural network.

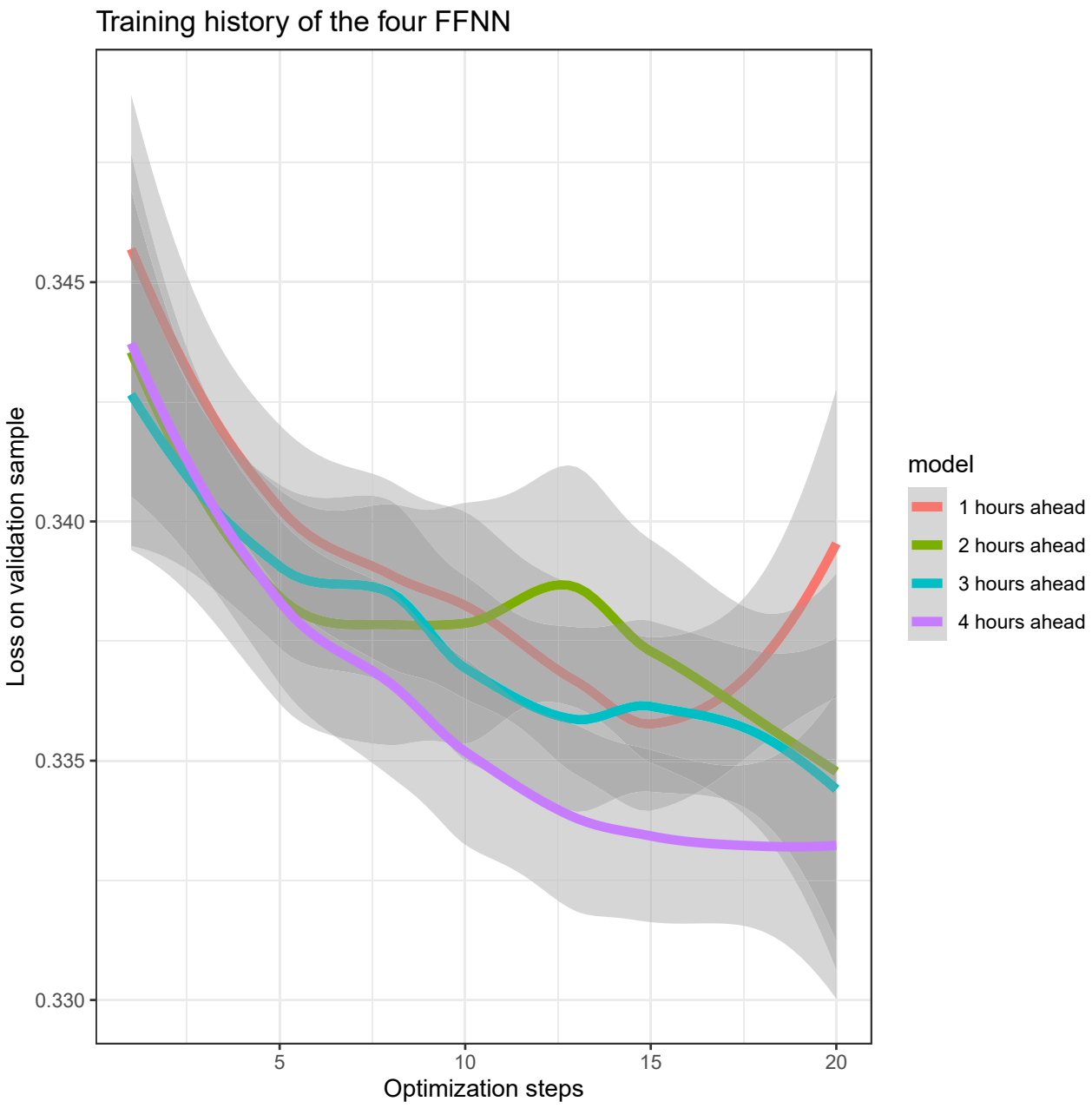

Figure 5: Illustration of loss levels across the four FFNNs during training steps. Losses on the validation sample are portrayed through smoothing splines along with their 95% confidence band.

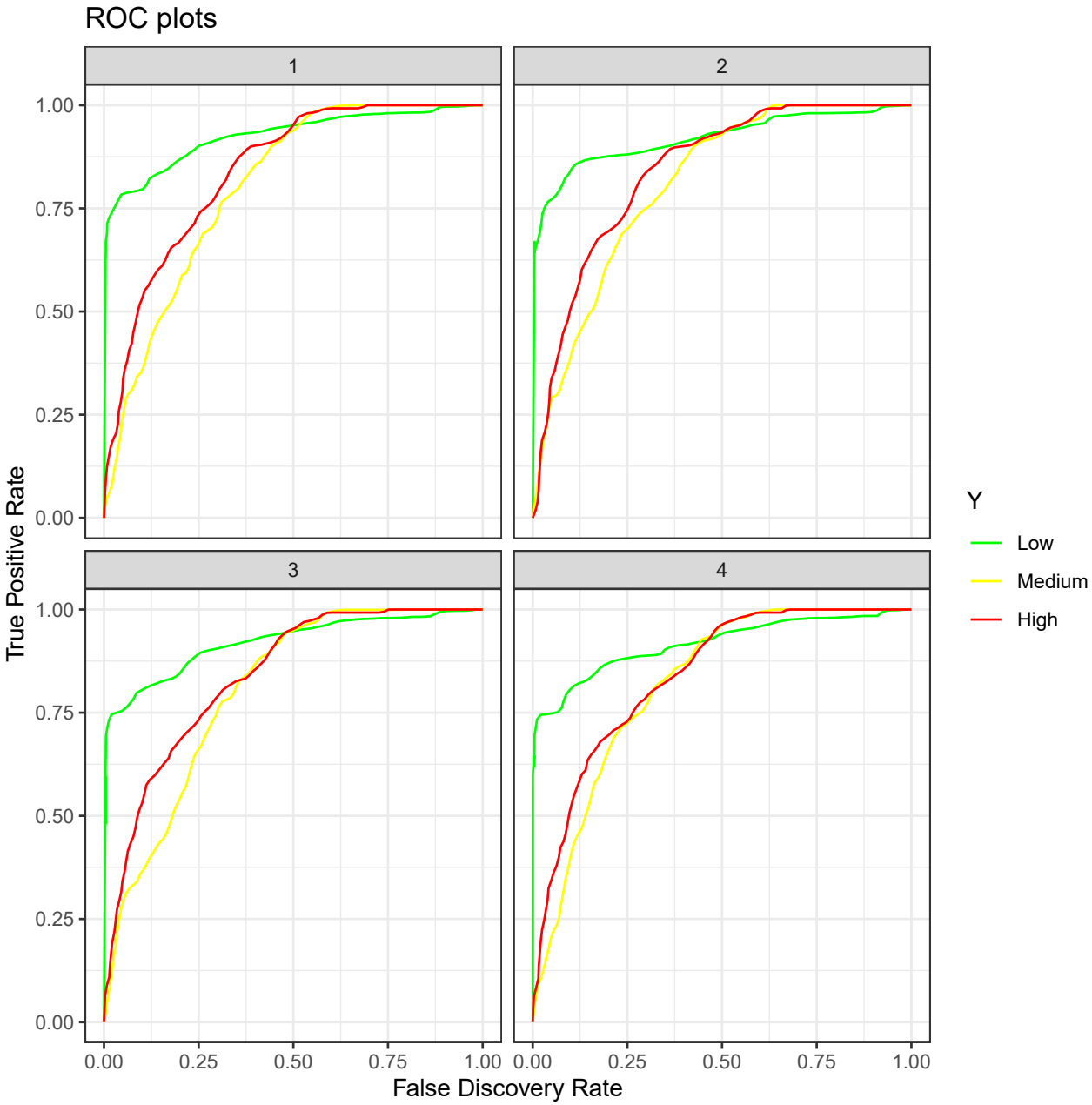

Figure 6: ROC curves conditioned on each predicted level on the test set for each model for the indicated number of hours ahead prediction.

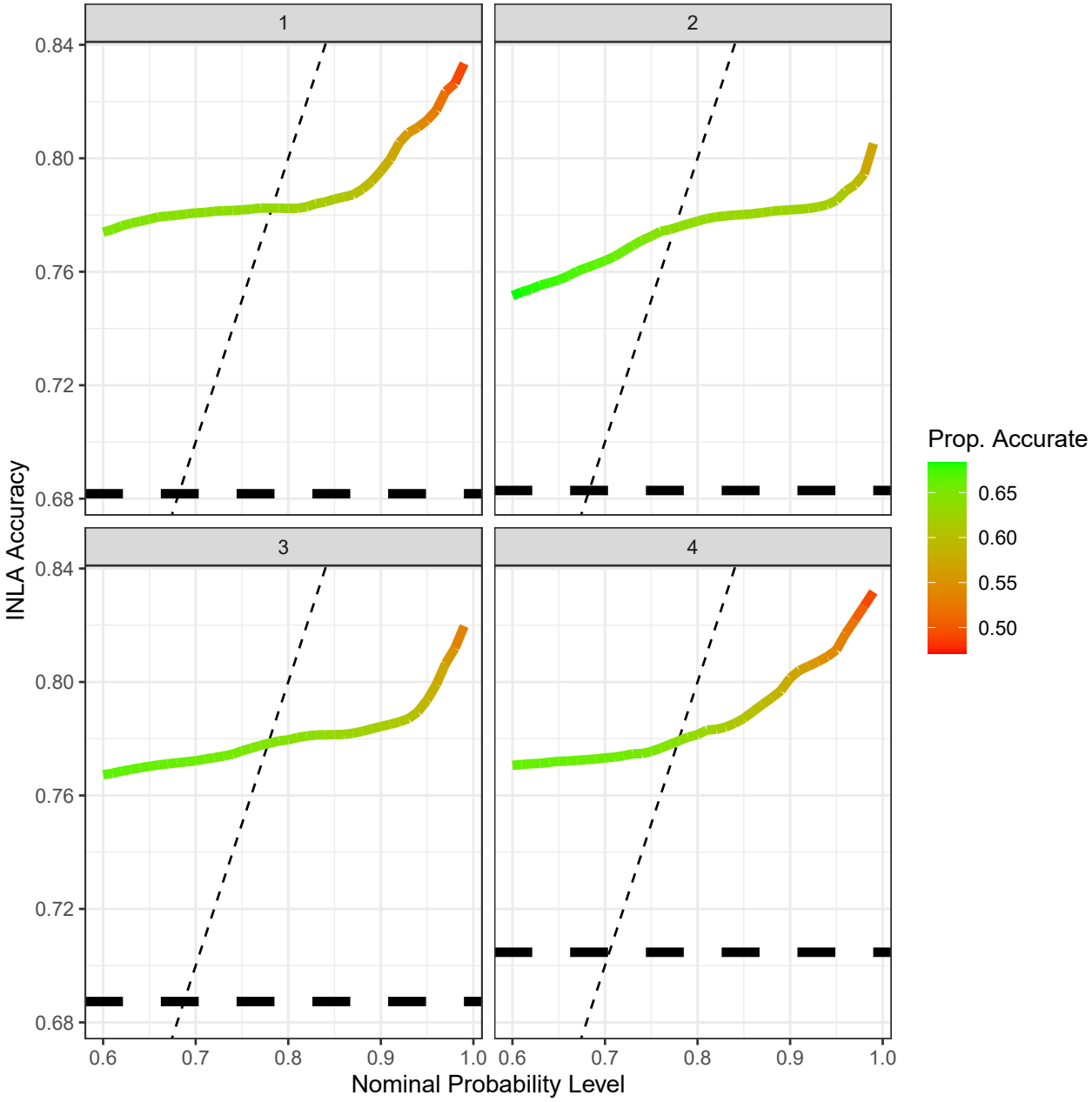

Figure 7: For each hour ahead prediction (values in the head of each panel) and varying nominal levels of probabilities $1-\delta$ (horizontal axis), the figure displays the accuracy on the test set of 5 days (vertical axis) alongside the proportion of observations deemed accurate at level $1-\delta$ (colour legend). The horizontal dashed line represents the accuracy of the DL model, which has no assessment of its conditional uncertainty.

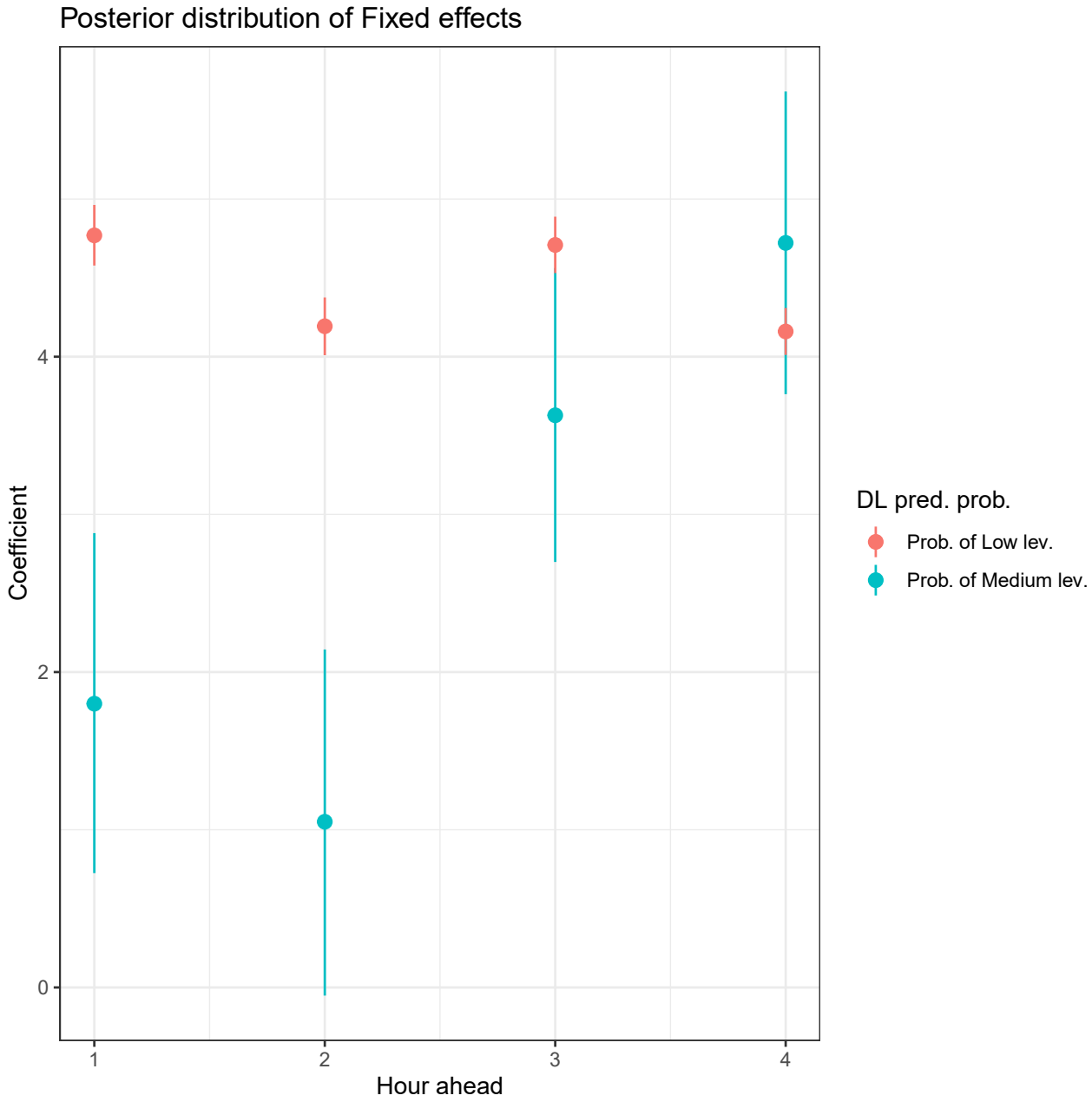

Figure 8: Summary of posterior distributions of fixed effects for each calibration model, i.e., one for each hour-ahead prediction.

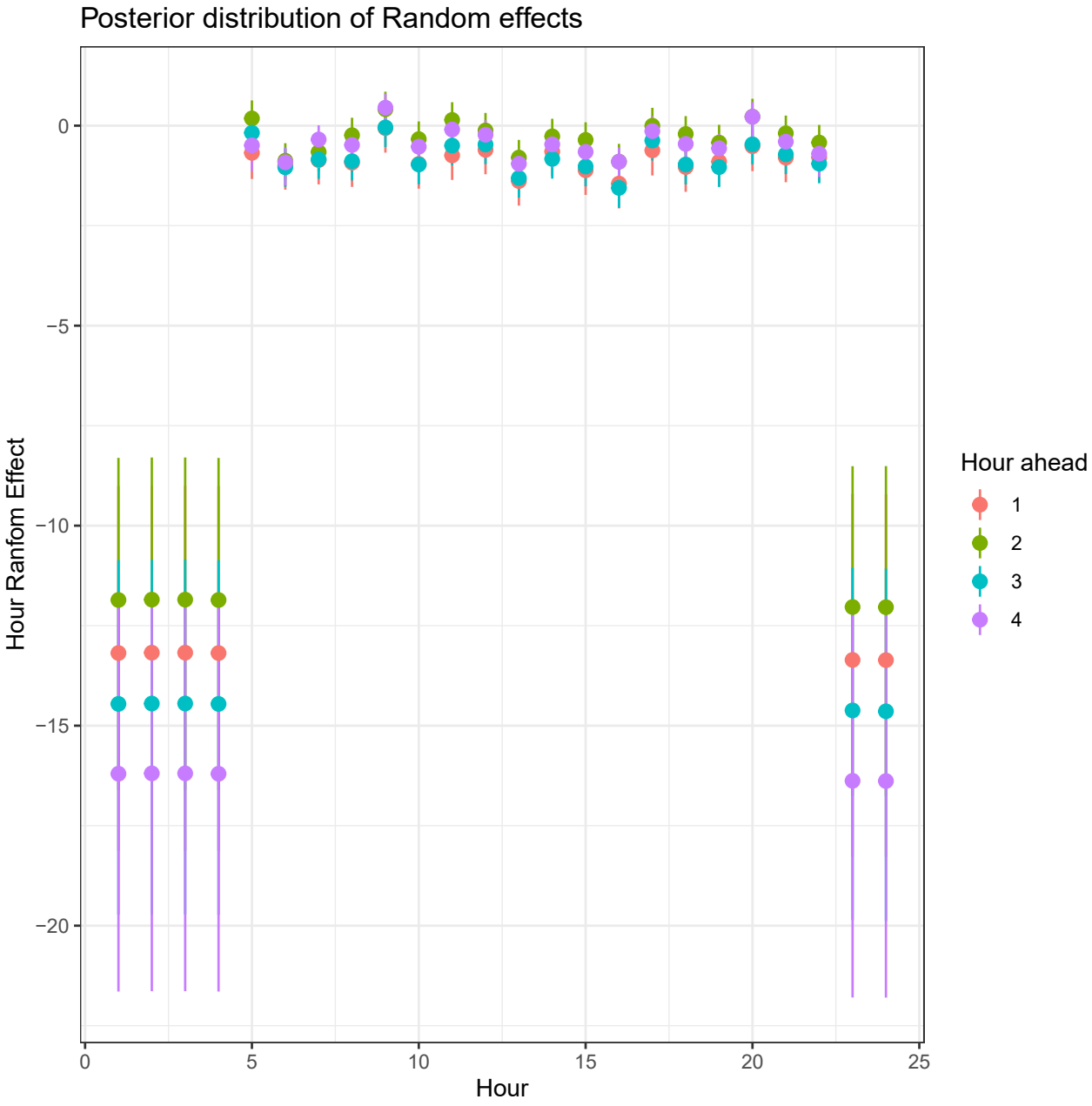

Figure 9: Summary of posterior distributions of hour random effect for each calibration model, i.e., one for each hour ahead prediction.

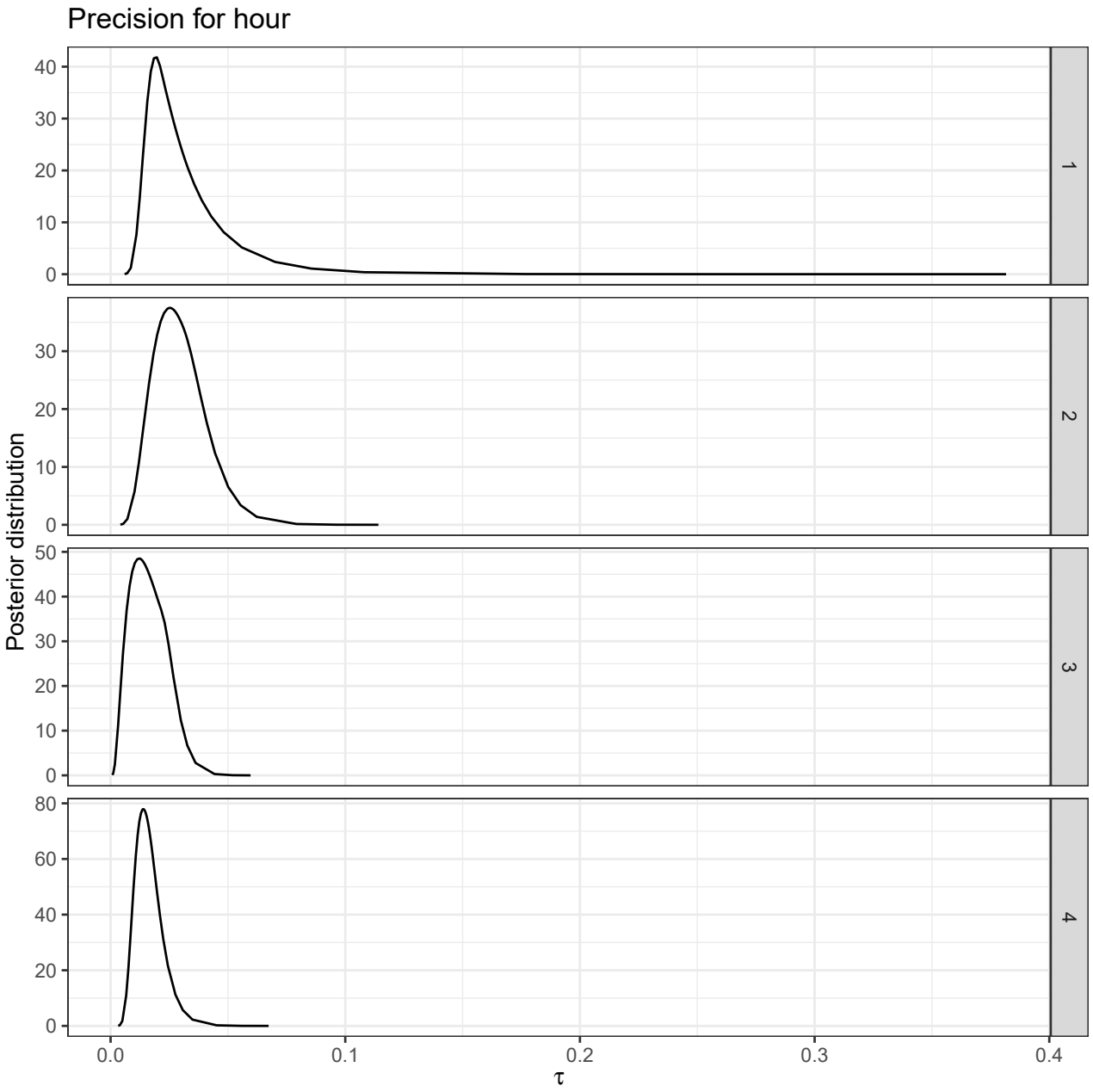

Figure 10: Posterior distributions of precision parameters of hour random effect. Lower precision of the hour random effect signifies its increased importance. Each row corresponds to each calibration model, i.e., one for each hour-ahead prediction.

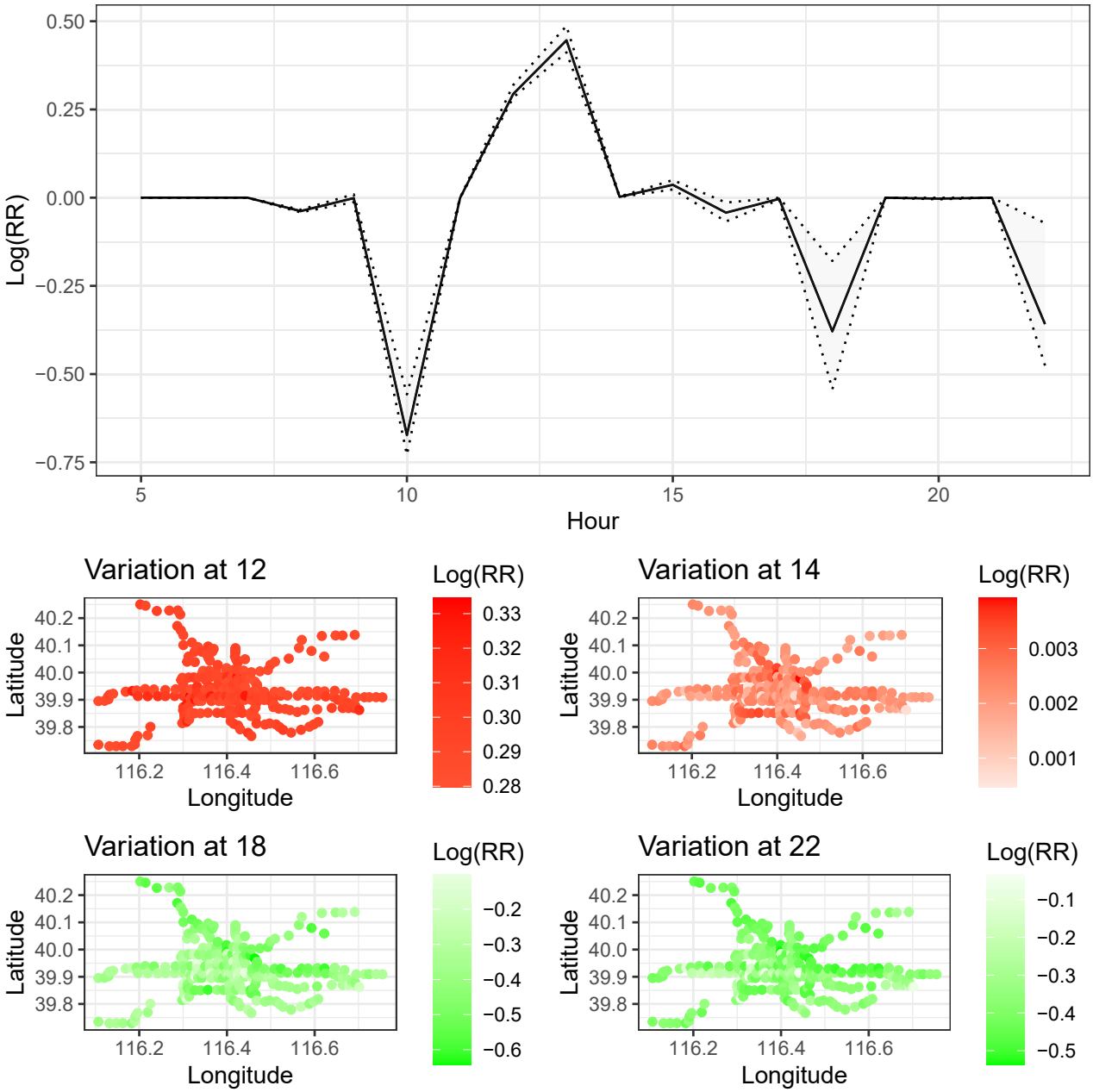

Figure 11: Posterior predictive distribution of the log differences in the probability, $\log(RR)$, of high occupancy levels $p_{st}^{(3)}$ when all three antenna levels change from low to high on October 29. Marginal (over stations) changes, along with their 95% Confidence bounds at varying hours, are presented at the top, while changes conditional to stations at specific hours are delineated at the bottom.

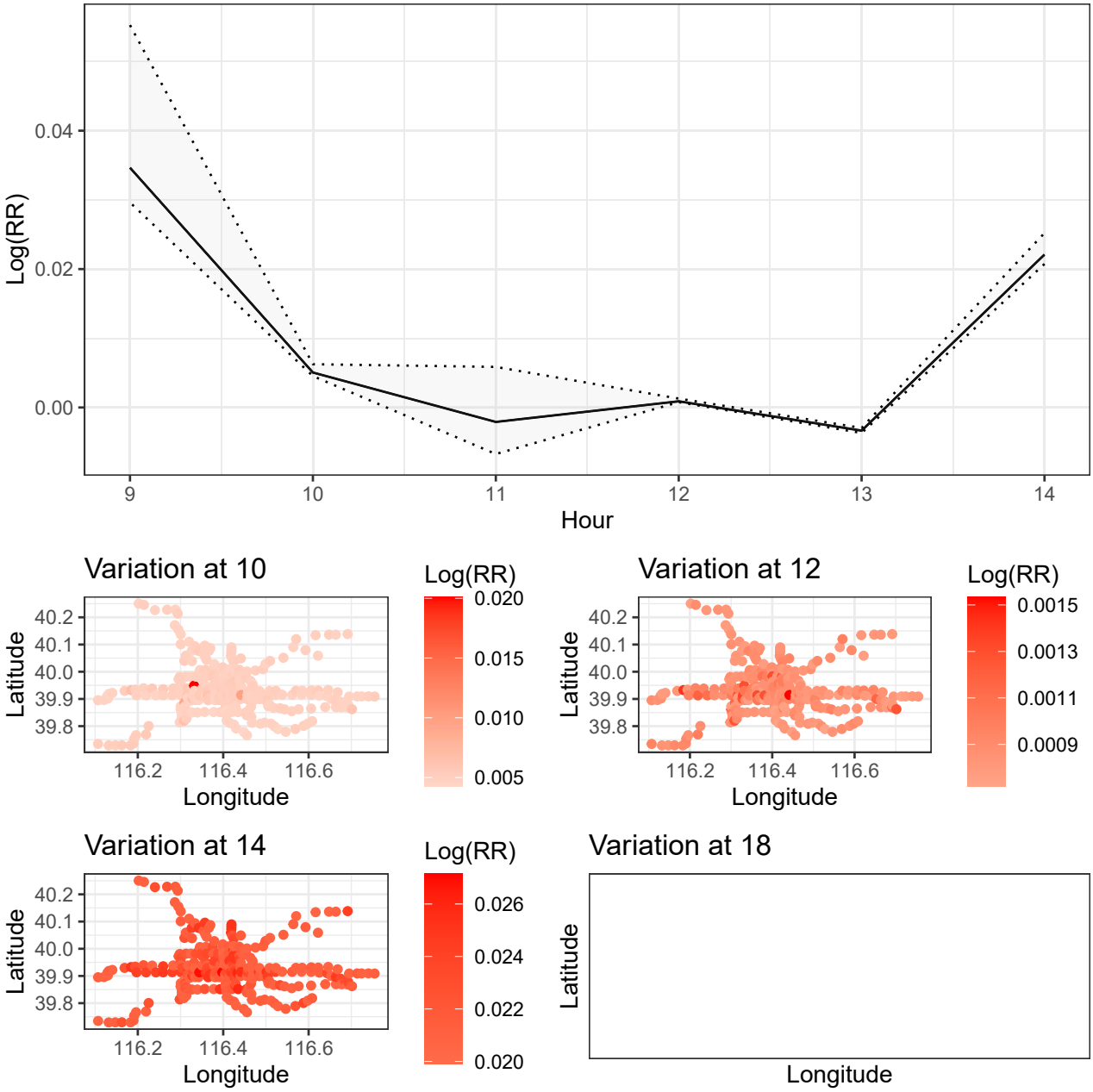

Figure 12: Posterior predictive distribution of the log differences in the probability, $\log(RR)$, of high occupancy levels $p_{st}^{(3)}$ when the occupancy level in stations Beijingxi, Huilongg, and Tiantongy shifts from 8 till 10 on October 22. Marginal (over stations) changes, along with their 95% Confidence bounds at varying hours, are presented at the top, while changes conditional to stations at specific hours are delineated at the bottom.

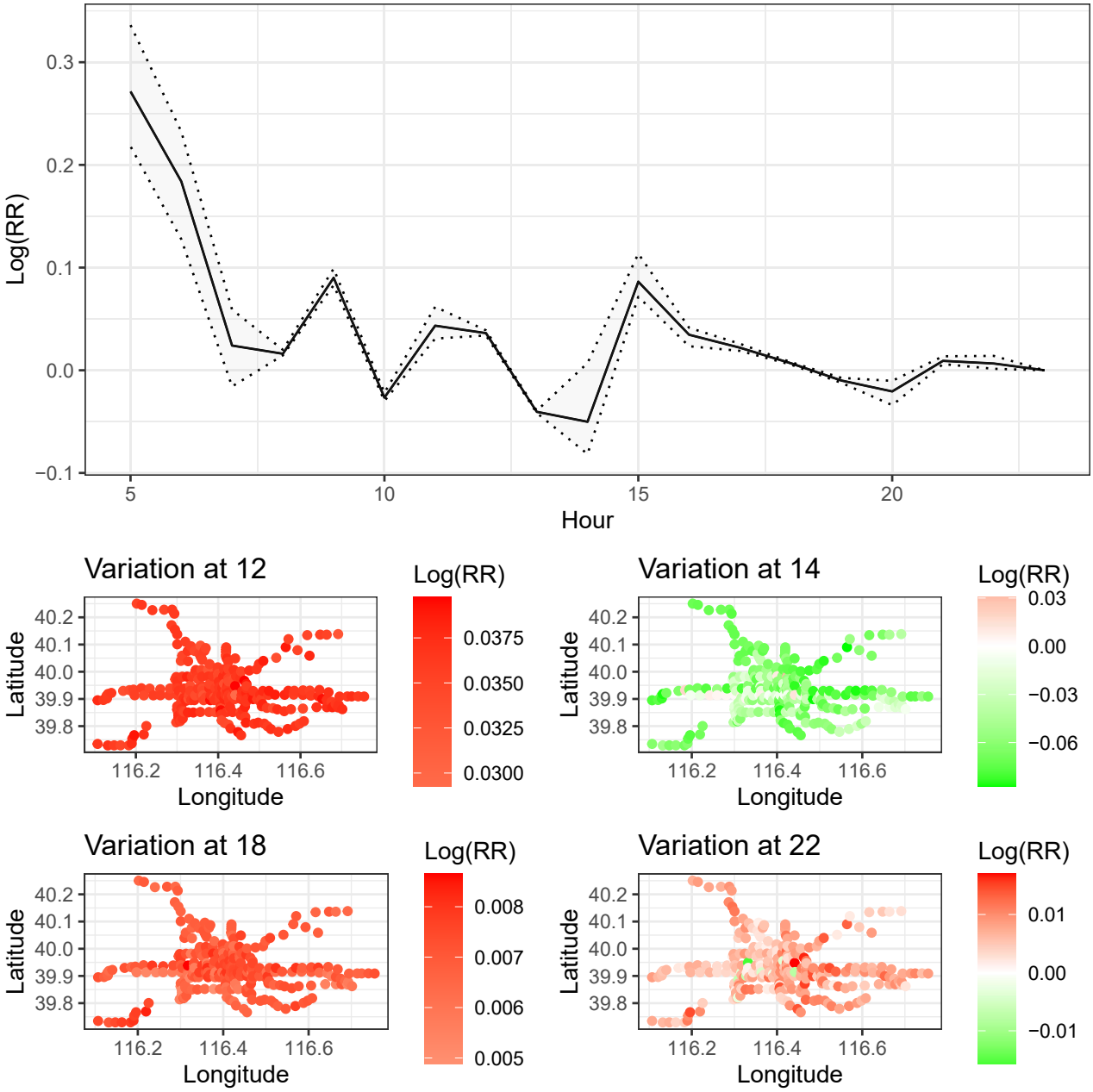

Figure 13: Posterior predictive distribution of the log differences in the probability, $\log(RR)$, of high occupancy levels $p_{st}^{(3)}$ when the resident and traffic on the 20 most centred mobile phone antenna transition from low to high on October 29. Marginal (over stations) changes, along with their 95% Confidence bounds at varying hours, are presented at the top, while changes conditional to stations at specific hours are delineated at the bottom.