Capturing network and dynamic effects in bike sharing system via fused Lasso

Yunjin Choi[1] Haeran Cho[2] Hyelim Son[3]

August 23, 2022

Abstract

Data collected from a bike-sharing system exhibit complex temporal and spatial features. We analyze shared-bike usage data collected in Seoul, South Korea, at the level of individual stations while accounting for station-specific behavior and covariate effects. We adopt a penalized regression approach with a multilayer network fused Lasso penalty. The proposed fusion penalties are imposed on networks which embed spatio-temporal linkages, and capture the homogeneity in bike usage that is attributed to intricate spatio-temporal features without arbitrarily partitioning the data. We demonstrate that the proposed approach yields competitive predictive performance and provides a new interpretation of the data.

Keywords: bike sharing system, fused Lasso, high dimensionality, network analysis

1 Introduction

bike-sharing systems (BSSs) have become increasingly popular in urban areas and have successfully complemented public transportation systems in dense metropolitan cities. In addition to its utility to bike users, the installation of BSSs has been found to reduce the usage of automobiles (Fishman et al., 2014) and thus traffic congestion and possibly green house emissions (Hamilton and Wichman, 2018). To fully realize these benefits, efficient allocation of docking stations and bike docks is essential, which in turn requires understanding the user behaviour and characteristics of the system based on the abundant data collected on the BSS, in addition to other urban and environmental factors that are known to influence bike usage. In line with the increasing popularity of BSSs, there exists a vast literature on the analysis of bike usage patterns (for and overview, see Shaheen et al. (2010); and Fishman (2016)) Below we provide a brief summary of the literature on quantitative or statistical analysis of BSS

[1]Department of Statistics, University of Seoul, South Korea. Email: ycstat@uos.ac.kr.
[2]School of Mathematics, University of Bristol, UK. Email: haeran.cho@bristol.ac.uk.
[3]School of Economics, University of Seoul, South Korea. Email: hlson@uos.ac.kr.

usage data which is categorized into two, following Etienne and Latifa (2014). The first branch addresses the problem of clustering stations in a system based on usage patterns according to some measure of similarity (Froehlich et al., 2009; Vogel et al., 2011; Etienne and Latifa, 2014). Regarding the BSS as a network, community detection algorithms have also been adopted for this purpose (Austwick et al., 2013; Borgnat et al., 2013; Zhou, 2015). Gervini and Khanal (2019) model the demand for bikes as a multivariate temporal point process and cluster stations based on functional canonical correlations of log-intensity functions.

The second line of research concerns the problem of predicting the station occupancy or the state of the system at a given time. Faghih-Imani and Eluru (2016) model incoming and outgoing traffic at multiple stations as a panel with variables accounting for spatial and temporal autoregressive structures. Liu et al. (2016) model inter-station bike transitioning for improving the effectiveness of rebalancing operations by predicting the station drop-off demand. Torti et al. (2021) adopt functional linear regression to model the directed flow between pairs of administrative divisions that aggregate multiple stations.

In all above, it is well-documented that BSS data show temporal and spatial patterns. To address such patterns, some previous works pre-process the datasets by, for example, aggregating stations into administrative regions (Torti et al., 2021), partitioning the data using subject-specific knowledge (Faghih-Imani and Eluru, 2016), or separately analyzing the data collected on weekdays and at weekends (Liu et al., 2016). In complex urban environments, however, it may be difficult to find a single clustering of the data that comprehensively accounts for the patterns in the usage of the BSS, since there exist multiple approaches to produce geographical or temporal divisions according to socioeconomic characteristics, land zones, traffic infrastructure or population composition. Then, two data points that can belong to the same division for one aspect (for example, according to administrative division) may belong to different divisions in another aspect (for example, business district versus tourist area). Besides, collecting in-depth information about the multifaceted nature of a large metropolitan city is typically costly or even impossible.

In this paper, we analyze the hourly bike rental data collected from a BSS in Seoul, South Korea, by adopting a penalized regression modelling approach. The proposed method utilizes the whole dataset without partitioning that still addresses the complex features inherent in the temporal and spatial properties of the data. We model the BSS at the granularity of individual stations by including station-specific parameters as well as trends and variables related to air quality and precipitation. Such a model enjoys considerable flexibility and captures time-dependent usage patterns at individual stations, but it potentially suffers from the risk of overfitting as the number of parameters ($> 4 \times 10^4$) increases linearly with the number of stations.

To address these issues, we propose a multilayer network fused Lasso penalty which extends the fused Lasso penalty (Tibshirani et al., 2005). In the absence of a natural ordering among

the stations, the proposed penalty imposes the penalization using a multilayer network and promotes fusion of the parameters linked by edges in the network. In doing so, we view the BSS as a multilayer network with the stations serving as its nodes. In the network, the layers correspond to different hours of a day and within each layer, a pair of stations are connected based on their geographical distance as a proxy for shared nodal features. The cross-layer edges come from that the usage patterns tend to vary over the course of a day.

The model fitted from the penalized regression method adaptively captures spatial and temporal homogeneity in bike usage, without (arbitrarily) partitioning the data which potentially leads to information loss. The degree of homogeneity is determined by the data-driven choice of the penalty parameter which does not involve the researcher's subjective decision. Our data analysis shows the superior predictive performance of the proposed multilayer fused Lasso penalty over alternative penalization approaches. Also, we propose a new network-based model complexity measure which reveals that while similarities exist, the stations exhibit fair amount of heterogeneity. This conclusion supports that partitioning the stations into a handful of clusters may be inappropriate for such large-scale urban transportation systems. The proposed idea of treating the underlying structure as a multilayer network, can be useful in problems beyond that of BSS modelling whenever the data exhibits network-like features.

The remainder of this paper is organised as follows. In Section 2, we describe the notable properties of BSSs such as station-specific temporal patterns, spatial homogeneity, and covariate effects, through an exploratory analysis of bike rental data collected in Seoul, South Korea. Section 3 introduces our proposed Poisson regression model and the accompanying penalization strategy that uses of fusion penalties, and provides a network-based interpretation of the latter. Section 4 demonstrates the effectiveness of our proposed approach on the examined BSS dataset. We conclude the paper in Section 5, and describe the algorithms for handling large-scale datasets in Appendix.

2 Exploratory analysis of BSS data

In this section, we describe data collected from "Ddareungi", a public BSS in Seoul, South Korea. We use the hourly rental records collected from individuals between April 1, 2019 and May 30, 2019. [1] By selecting the temperate months of April and May, we avoid dealing with seasonality or possibly abnormal observations due to extreme weather conditions. We also exclude three public holidays falling in this period from our analysis. Thus, our final dataset comprises observations from $T = 57$ days and $S = 1,505$ stations.

Since the launch of the BSS in 2010, the numbers of subscribed users, stations, and available bikes have steadily increased, as has the number of trips. In addition to this overall trend, the bike usage data exhibits substantial temporal variations across localities, as well as dependence

[1] The dataset is available at `https://data.seoul.go.kr/`.

on precipitation and air quality, which is in line with the observations of previous BSSs literature.

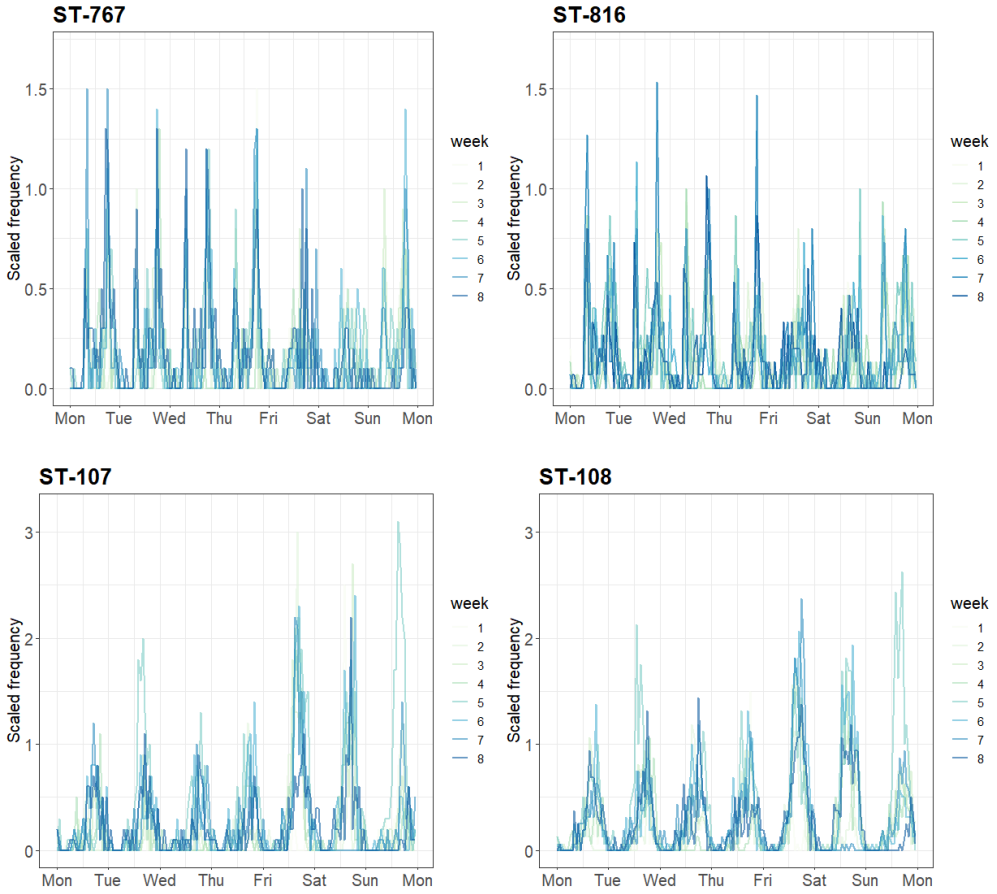

Figure 1: Hourly bike rentals from four selected stations over eight weeks, adjusted by the capacity (number of docking stations) of individual stations. Here, stations ST-767 and ST-816 are adjacent to one another and so are ST-107 and ST-108. We note that Week 5 contains three public holidays and show some anomalies.

Bike rental patterns exhibit substantial heterogeneity across stations. In Figure 1, we plot the hourly bike rental frequencies from selected pairs of stations adjacent to one another over the examined two-month period. There are clear station-specific patterns over the course of a day and a week, and nearby stations tend to display similar usage patterns. Specifically, as stations ST-767 and ST-816 are located in a commercial district with government agencies and large firms, a spike in bike rental frequency is observed during weekday commuting-time. On the other hand, stations ST-107 and ST-108 are located along a major riverside park in Seoul, and hence, are typically used for leisure activities, as evidenced by the large number of rentals concentrated on weekends. Additionally, although daily peaks and troughs in usage can be observed, these temporal patterns do not undergo abrupt changes in the sense that the number of bikes rented out between 9am and 10am is reasonably close to that between

10am and 11am.

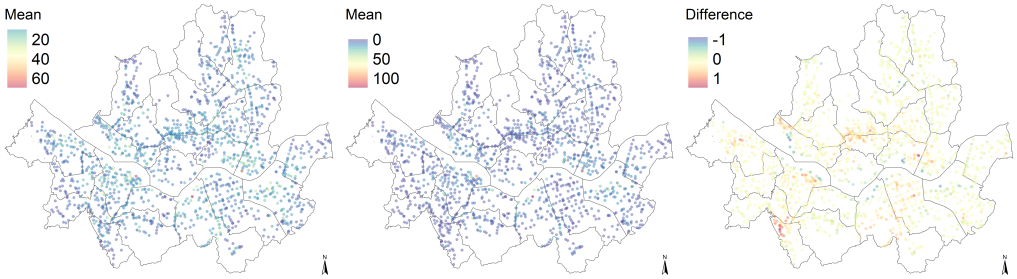

Figure 2: Average daily bike rentals at individual stations on weekdays (left), and weekends (middle) and their station-wise differences (right). The differences are taken in log-scale for better visualisation.

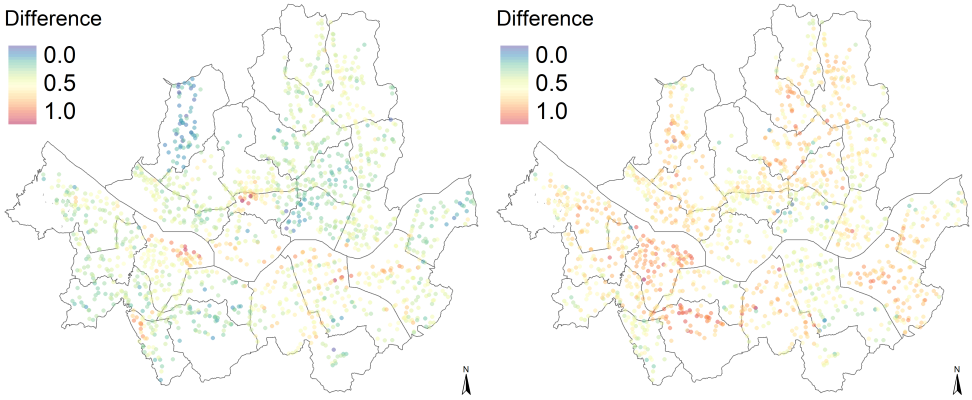

Figure 3: Difference in average daily bike rentals (in log-scale) at individual stations due to precipitation (left) and air quality (right).

Figure 2 plots the daily rental frequencies averaged over weekdays and weekends as well as their differences (in log-scale) at individual stations. The figure shows that bike usage behavior differs between weekdays and weekends, and that the degree of variation differs across stations. Additionally, the last panel of Figure 2 exhibits the presence of local clusters that share similar weekday/weekend variations.

Figure 3 shows that bike usage depends on the weather condition and air quality. We record each day as "rainy" if positive precipitation is recorded in any part of the city at any time of the day. Comparing the average rental frequencies on days with and without precipitation (in log-scale, see the left panel of Figure 3), precipitation reduces bike usage, as expected. Similar observations can be made with respect to the air quality: We plot the differences in average rental frequencies (in log-scale) on days when the air quality was "good" and "very bad", in the right panel of Figure 3. The qualitative categorization into "good", "average", "bad," and "very bad" follows classification system of the Korean Ministry of Environment based on PM10 and PM2.5 dust concentrations. The relative rental frequencies across the

BSS display a similar pattern on "good" and "very bad" days, but the volume of bike rentals is lower on "very bad" days.
In summary, the BSS dataset exhibits the following characteristics:

(C1) There exists an overall increasing trend.

(C2) Bike rentals show station-specific temporal patterns and the transition is reasonably smooth.

(C3) These patterns are shared across stations that are geographically adjacent to one another.

(C4) Bike rentals are influenced by the weather condition and the air quality.

3 Model and estimation

Motivated by the observations made in (C1)–(C4) in Section 2, we propose a Poisson regression model for hourly rental frequencies collected from the entire BSS (Section 3.1). Section 3.2 presents the accompanying estimation strategy and introduces the multilayer fused Lasso penalty, which is designed to capture the characteristics (C2)–(C3). In Section 3.3, we provide a network interpretation of the proposed penalization technique which aids in understanding and visualizing the penalty.

3.1 Poisson regression model

Let Y_i denote the ith observation representing the hourly rental frequency at station $S(i) \in \mathcal{S} = \{1, \ldots, S\}$ and hour $H(i) \in \mathcal{H} = \{0, \ldots, 23\}$, on day $D(i) \in \mathcal{D} = \{\text{Mo, Tu, } \ldots, \text{ Su}\}$ for $i = 1, \ldots, n$, with $n = 2,058,840$ denoting the number of observations in the bike usage dataset. We denote the time index of the ith observation by $t(i) \in \mathcal{T} = \{0, \ldots, T-1\}$. Since we consider the period of two months, as described in Section 2 (excluding three public holidays), we have $T = 57$. Additionally, $z_i^{\text{rain}} \in \{0, 1\}$ and $\mathbf{z}_i^{\text{air}} \in \{0, 1\}^4$ denote the variables representing the precipitation and air quality statuses associated with the ith observation, respectively. Consistent with the literature on BSSs and with the observations made in Section 2, bike usage depends on the aforementioned variables and we collect the variables relevant for the ith observation in $\mathbf{x}_i = (S(i), t(i), D(i), H(i), z_i^{\text{rain}}, (\mathbf{z}_i^{\text{air}})^\top)^\top$ (for an overview, see Table 1).
Each station in the BSS of Seoul, South Korea, has a fixed number of docks, but unlike other BSSs, this does not determine the capacity of a station, as it is possible to leave bikes even if the docks are fully occupied through chaining them to existing bikes. Therefore, we adopt a Poisson distribution for modelling Y_i, the hourly count of the bikes rented out at station

Table 1: Variables in the BSS dataset.

Variable	Range	Description
$S(i)$	$\mathcal{S} = \{1, \ldots, S\}$	Station node for the ith observation (categorical)
$H(i)$	$\mathcal{H} = \{0, \ldots, 23\}$	Hour of a day for the ith observation (categorical)
$D(i)$	$\mathcal{D} = \{\text{Mo}, \ldots, \text{Su}\}$	Day of a week for the ith observation (categorical)
z_i^{rain}	$\{0, 1\}$	Whether it rained or not for the ith observation (categorical)
$\mathbf{z}_i^{\text{air}}$	$\{0, 1\}^4$	Whether air quality was very bad, bad, average or good for the ith observation (categorical)
$t(i)$	$\{0, \ldots, T-1\}$	Time point for the ith observation (numeric)

$S(i)$, and model the relationship between Y_i and $\mathbf{x}_i$ via Poisson regression (McCullagh and Nelder, 1989, Chapter 6) as

$$Y_i|\mathbf{x}_i \sim_{\text{iid}} \text{Poisson}(\mu_i) \quad \text{with} \quad \mu_i \equiv \mu(\mathbf{x}_i) = \mathsf{E}(Y_i|\mathbf{x}_i).$$

One way to model the link between μ_i and $\mathbf{x}_i$ is to use the following linear model:

$$\log\left(\frac{\mu_i}{C_{S(i)}}\right) = \theta_{S(i)} + \alpha\, t(i) + \beta^{\text{rain}}\, z_i^{\text{rain}} + \langle \boldsymbol{\beta}^{\text{air}}, \mathbf{z}_i^{\text{air}} \rangle + \theta_{H(i)}^{\text{hod}} + \theta_{D(i)}^{\text{dow}} \tag{1}$$

where $\boldsymbol{\beta}^{\text{air}} = (\beta_j^{\text{air}},\, 0 \le j \le 3)^\top$. Here, θ_h^{hod} and θ_d^{dow}, respectively, contain "hour of a day" and "day of a week" effects common to all stations. An offset term C_s relates to the capacity of station s, that is, the number of docks, so that (1) can be interpreted as modelling the expected rental frequency per hour per station capacity. This model, referred to as the *no-interaction* model, does not permit the temporal effects to be station-specific and thus is too simple to address (C2).

Allowing for interactions between the station and temporal effects, we consider the following *full-interaction* model:

$$\begin{aligned}\log\left(\frac{\mu_i}{C_{S(i)}}\right) = \theta_{S(i)} + \alpha\, t(i) + \beta^{\text{rain}}\, z_i^{\text{rain}} + \langle \boldsymbol{\beta}^{\text{air}}, \mathbf{z}_i^{\text{air}} \rangle \\ + \theta_{H(i)}^{\text{hod}} + \theta_{D(i)}^{\text{dow}} + \theta_{S(i),H(i)}^{\text{hod}} + \theta_{S(i),D(i)}^{\text{dow}}.\end{aligned} \tag{2}$$

The station-hour and station-day interaction terms $\theta_{s,h}^{\text{hod}}$ and $\theta_{s,d}^{\text{dow}}$ allow each station to exhibit individual temporal patterns. For model identifiability, we set the baseline parameters to zero; in other words, $\beta_0^{\text{air}} = \theta_0^{\text{hod}} = \theta_{\text{Mo}}^{\text{dow}} = \theta_{s,0}^{\text{hod}} = \theta_{s,\text{Mo}}^{\text{dow}} = \theta_{1,h}^{\text{hod}} = \theta_{1,d}^{\text{dow}} = 0$.

Mmodel (2) accounts for (C1), (C4), and to a certain extent, (C2) by including the parameters α capturing the overall trend, β^{rain} and $\boldsymbol{\beta}^{\text{air}}$ capturing the effects of precipitation and air quality, and $\theta_{s,h}^{\text{hod}}$ and $\theta_{s,d}^{\text{dow}}$ addressing station-specific temporal patterns. In doing so, we take a different approach from those taken in previous studies in which, after (arbitrarily) partitioning the dataset according to temporal or spatial variables, or both, individual par-

titions are separately modelled (e.g., Austwick et al. (2013); Liu et al. (2016)). Instead, by including the interaction terms, we use the full dataset and avoid any information loss from data partitioning.

While the full-interaction model enjoys considerably more flexibility than the no-interaction model in (1), it suffers the risk of overfitting the data with the number of parameters to be estimated amounting to $p = 34 + S + (S-1) \cdot (23+6) = 45,155$. In addition, the characteristic identified in (C3), that is, stations geographically close to one another tending to exhibit similar usage patterns, is not adequately accounted for by the model fitted without any constraint. Such an approach does not benefit from the temporal ordering inherent in the parameters $\theta^{\text{hod}}_{s,h}$, and thus does not fully account for (C2). In the next section, we propose a penalized maximum likelihood estimation (MLE) methodology for model in (2) with a multilayer network fused Lasso which explicitly sets out to address these issues.

3.2 Penalized MLE via multilayer network fused Lasso

We bridge the two models (1) and (2) at extreme ends, by adopting a penalized MLE strategy with a Lasso penalty mposed on the interaction parameters, and the fusion penalties designed to capture spatial and temporal homogeneity in bike usage patterns observed in the data, that is, (C2) and (C3).

First proposed by Tibshirani et al. (2005), the fused Lasso augments the ℓ_1-penalized least squares estimation method–or Lasso (Tibshirani, 1996)– with a penalty that takes advantage of a meaningful ordering of the variables when such is available. This fusion penalty is imposed on the ℓ_1-norm of successive differences in the parameters corresponding to the ordered variables and encourages local constancy therein. Its use has been extended beyond regression problems such as trend estimation (Tibshirani, 2014), change point detection (Harchaoui and Lévy-Leduc, 2010) and graphical modelling (Danaher et al., 2014) among others.

Under full-interaction model in (2), we partition the parameters into $\Theta = \{\theta_s,\, s \in \mathcal{S}\}$, $\Theta_H = \{\theta^{\text{hod}}_h,\, \theta^{\text{hod}}_{s,h},\, h \in \mathcal{H} \setminus \{0\},\, s \in \mathcal{S} \setminus \{1\}\}$, $\Theta_D = \{\theta^{\text{dow}}_d,\, \theta^{\text{dow}}_{s,d},\, d \in \mathcal{D} \setminus \{\text{Mo}\},\, s \in \mathcal{S} \setminus \{1\}\}$ and $\Delta = \{\alpha, \beta^{\text{rain}}, \beta^{\text{air}}_1, \beta^{\text{air}}_2, \beta^{\text{air}}_3\}$. We adopt the fusion penalty to pool the information

(i) from adjacent stations for the estimation of Θ, Θ_H and Θ_D, and

(ii) over the course of a day for the estimation of Θ_H.

We also impose a standard lasso penalty to encourage sparsity. Subsequently, our objective is to minimize the following penalized negative log-likelihood

$$-\ell(\Theta, \Theta_H, \Theta_D, \Delta) + \lambda \sum_{s \in \mathcal{S}} \left(\sum_{h \in \mathcal{H}} |\theta^{\text{hod}}_{s,h}| + \sum_{d \in \mathcal{D}} |\theta^{\text{dow}}_{s,d}| \right) + \lambda_N \cdot p_N(\Theta, \Theta_H, \Theta_D) + \lambda_H \cdot p_H(\Theta_H) \tag{3}$$

with respect to Θ, Θ_H, Θ_D and Δ. Here, $\ell(\Theta, \Theta_H, \Theta_D, \Delta)$ denotes the log-likelihood

$$\ell(\Theta, \Theta_H, \Theta_D, \Delta) = -\sum_{i=1}^{n} \mu_i + \sum_{i=1}^{n} y_i \log(\mu_i) + \text{constant},$$

and $\mu_i = \mu_i(\Theta, \Theta_H, \Theta_D, \Delta)$ is a function of the parameter vectors under (2). The parameter λ controls the degree of the Lasso penalization and λ_N and λ_H control that of the fusion penalization; below we discuss the construction of the functions p_N and p_H to achieve the above goals (i) and (ii).

To capture the similarities between geographically adjacent stations, we define a set of neighboring stations for each station s as $\mathcal{N}_r(s) = \{s' \in \mathcal{S} \setminus \{s\} : d(s, s') < r\}$ and its cardinality by $|\mathcal{N}_r(s)|$, where $d(s, s')$ denotes the distance between the two stations s and s' measured as the crow flies, and r denotes a pre-specified distance. We define

$$p_N(\Theta, \Theta_H, \Theta_D) = \sum_{s \in \mathcal{S}} \sqrt{|\mathcal{N}_r(s)| \sum_{s' \in \mathcal{N}_r(s)} \left[\sum_{h \in \mathcal{H}} (\phi^{\text{hod}}_{s,h} - \phi^{\text{hod}}_{s',h})^2 + \sum_{d \in \mathcal{D}} (\phi^{\text{dow}}_{s,d} - \phi^{\text{dow}}_{s',d})^2 \right]} \tag{4}$$

with $\phi^{\text{hod}}_{s,h} = \theta_s + \theta^{\text{hod}}_h + \theta^{\text{hod}}_{s,h}$ and $\phi^{\text{dow}}_{s,d} = \theta_s + \theta^{\text{dow}}_d + \theta^{\text{dow}}_{s,d}$ which fully encode the station-specific hourly and daily patterns under (2). The penalty p_N forces the pairs of parameters $(\phi^{\text{hod}}_{s,h}, \phi^{\text{hod}}_{s',h})$ and $(\phi^{\text{dow}}_{s,d}, \phi^{\text{dow}}_{s',d})$ to fuse neighboring stations s and s', which encourages them to exhibit similar hourly and daily patterns and thus addresses the behavior noted in (C3).

Through p_N, the estimates of $\phi^{\text{hod}}_{s,h}$ and $\phi^{\text{hod}}_{s',h}$ may fuse even when $s' \notin \mathcal{N}_s(r)$ as long as they are connected via intermediate station nodes, and the same applies to the estimates of $\phi^{\text{dow}}_{s,d}$ and $\phi^{\text{dow}}_{s',d}$. This may be interpreted that by adopting p_N, we pool information across the BSS to estimate the interaction parameters (for a further discussion, see Section 3.3). The penalty in (4) is related to the spatial fusion penalties (Sun et al., 2016; Li and Sang, 2019; Sass et al., 2021) adopted in the spatial regression literature, but is clearly distinguished as we adopt the group Lasso (Yuan and Lin, 2006) in its formulation. Every parameter inside the square root is considered to belong to the same group, and the weighting applied with the size of $\mathcal{N}_r(s)$ follows the convention in the group Lasso literature. Then, the parameters associated with the stations with a large number of neighbors receive more penalization.

Remark 3.1 (Re-parameterization using $\phi^{\text{hod}}_{s,h}$ and $\phi^{\text{dow}}_{s,d}$.). We choose to decompose the station-specific hourly and daily effects as $\phi^{\text{hod}}_{s,h} = \theta_s + \theta^{\text{hod}}_h + \theta^{\text{hod}}_{s,h}$ and $\phi^{\text{dow}}_{s,d} = \theta_s + \theta^{\text{dow}}_d + \theta^{\text{dow}}_{s,d}$. Then, the proposed method applies the Lasso penalty to the ℓ_1-norm of $\theta^{\text{hod}}_{s,h}$ and $\theta^{\text{dow}}_{s,d}$ only, which gives it the interpretation of bridging between no-interaction and full-interaction models. We choose not to impose the Lasso penalization on θ_s to avoid cancelling out station-specific usage patterns, or θ^{hod}_h and θ^{dow}_d, which represent overall hourly and daily patterns shared across all stations.

Recall that, for model identifiability, we set $\theta^{\text{hod}}_{s,0} = 0$. Thus, we can write

$$\sum_{h\in\mathcal{H}}(\phi^{\text{hod}}_{s,h} - \phi^{\text{hod}}_{s',h})^2 = (\theta_s - \theta_{s'})^2 + \sum_{h=1}^{23}(\theta^{\text{hod}}_{s,h} - \theta^{\text{hod}}_{s',h})^2$$

and similarly, since $\theta^{\text{dow}}_{s,\text{Mo}} = 0$, we have

$$\sum_{d\in\mathcal{D}}(\phi^{\text{dow}}_{s,d} - \phi^{\text{dow}}_{s',d})^2 = (\theta_s - \theta_{s'})^2 + \sum_{d\in\mathcal{D}\setminus\{\text{Mo}\}}(\theta^{\text{dow}}_{s,d} - \theta^{\text{dow}}_{s',d})^2.$$

From these observations, we can re-write p_N as

$$p_N(\Theta,\Theta_H,\Theta_D) = \sum_{s\in\mathcal{S}}\sqrt{|\mathcal{N}_r(s)|\sum_{s'\in\mathcal{N}_r(s)}\left[2(\theta_s-\theta_{s'})^2 + \sum_{h=1}^{23}(\theta^{\text{hod}}_{s,h} - \theta^{\text{hod}}_{s',h})^2 + \sum_{d\in\mathcal{D}\setminus\{\text{Mo}\}}(\theta^{\text{dow}}_{s,d} - \theta^{\text{dow}}_{s',d})^2\right]},$$

which shows that p_N implicitly encourages the parameters $(\theta_s, \theta_{s'})$ to take similar values when stations s and s' are neighbors.

There is a natural temporal ordering inherent in Θ_H that gives rise to the fusion penalty in its canonical form imposed on $\phi^{\text{hod}}_{s,h}$:

$$p_H(\Theta_H) = \sum_{s\in\mathcal{S}}\sum_{h=0}^{23}\left|\phi^{\text{hod}}_{s,h} - \phi^{\text{hod}}_{s,h+1}\right| = \sum_{s\in\mathcal{S}}\sum_{h=0}^{23}\left|(\theta^{\text{hod}}_{h} + \theta^{\text{hod}}_{s,h}) - (\theta^{\text{hod}}_{h+1} + \theta^{\text{hod}}_{s,h+1})\right| \tag{5}$$

with $\theta^{\text{hod}}_{s,24} = \theta^{\text{hod}}_{s,0}$. Imposing a penalty on $p_H(\Theta_H)$ encourages the consecutive station-specific hourly effects $\phi^{\text{hod}}_{s,h}$ and $\phi^{\text{hod}}_{s,h+1}$ at a given station s, to become close to one another and suppress abrupt changes in usage. This reflects the fact that rental counts at each station rarely undergo radical shifts during the course of the day.

Jointly, p_N and p_H comprise the proposed multilayer network fusion penalty. We devote Section 3.3 to its interpretation with the description of the underlying multilayer network. The impact of the fusion penalty is determined by the sizes of λ_N and λ_H, which we select via cross-validation, as described in Section 4.1. The Alternating Direction Method of Multipliers (ADMM) algorithm (Boyd et al., 2011) is employed to solve the convex optimization problem in (3). Efficient implementation of the algorithm requires careful re-parametrization of model (2), which makes use of the specific structure of the data, and we discuss this in detail in Appendix A.1.

3.3 multilayer network interpretation of the fusion penalty

We provide a network interpretation of the proposed penalized regression method. First, we introduce the following networks that underpin the penalty functions p_N and p_H:

$$\mathfrak{N}_{\text{single}}(r) = (\mathcal{S}, \mathcal{E}_{\text{single}}(r)) \quad \text{with} \quad \mathcal{E}_{\text{single}}(r) = \cup_{s\in\mathcal{S}}\{(s, s'),\ s' \in \mathcal{N}_r(s)\}, \tag{6}$$

$$\mathfrak{N}_{\text{multi}}(r) = (\mathcal{S}\times\mathcal{H},\ \mathcal{E}_{\text{multi}}(r)) \quad \text{with}$$
$$\mathcal{E}_{\text{multi}}(r) = \left[\cup_{h\in\mathcal{H}} \cup_{s\in\mathcal{S}} \left\{\left((s,h),(s',h)\right),\ s' \in \mathcal{N}_r(s)\right\}\right] \bigcup \left[\cup_{s\in\mathcal{S}} \left\{\left((s,h),(s,h+1)\right),\ h \in \mathcal{H}\right\}\right]. \tag{7}$$

(For an illustrative example of $\mathfrak{N}_{\text{single}}$ and $\mathfrak{N}_{\text{multi}}$, see Figure 4). While both $\mathfrak{N}_{\text{single}}$ and $\mathfrak{N}_{\text{multi}}$ as well as their edge sets depend on the choice of r, we suppress this dependency for simplicity when it does not cause any confusion.

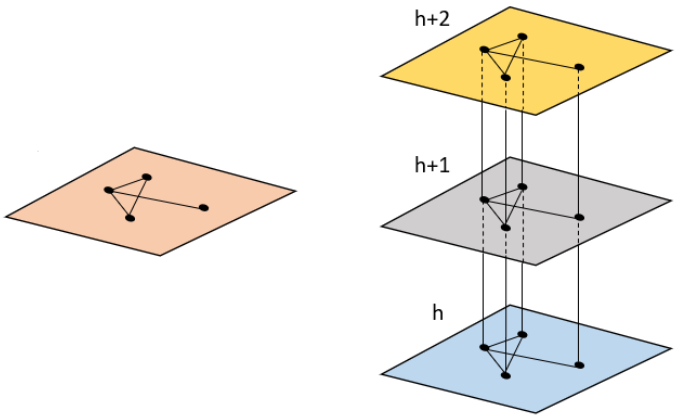

Figure 4: Illustration of a single-layer network $\mathfrak{N}_{\text{single}}$ (left) and a multilayer network $\mathfrak{N}_{\text{multi}}$ (right). Each dot represents a node (station) and a solid line represents an edge connecting the nodes within each layer and across adjacent layers.

Network $\mathfrak{N}_{\text{single}}$ is a single-layer, undirected network that is solely determined by the sets of neighbors $\mathcal{N}_r(s)$ of the stations. In this network, a pair of "day of a week" parameters $(\phi^{\text{dow}}_{s,d}, \phi^{\text{dow}}_{s',d})$ for each given day d, are encouraged to take values close to one another by the penalty function p_N, provided that the stations belong to the same connected component of $\mathfrak{N}_{\text{single}}$. On the other hand, $\mathfrak{N}_{\text{multi}}$ is a multilayer, undirected network; we follow the notational convention of Kivelä et al. (2014), in which the set $\mathcal{H}$ serves as a set of elementary layers for the hourly aspect, and each edge connects a pair of node-layer tuples (s,h) and (s',h') for some $s, s' \in \mathcal{S}$ and $h, h' \in \mathcal{H}$. Each pair of the "hour of a day" parameters $\phi^{\text{hod}}_{s,h}$ and $\phi^{\text{hod}}_{s',h'}$ is encouraged to fuse with one another by the penalty functions p_N and p_H, if the corresponding pair of nodes are connected in $\mathfrak{N}_{\text{multi}}$.

Next, we define networks whose edges are determined by the coefficient estimates from the penalized MLE. Let $\widehat{\phi}^{\text{hod}}_{s,h}$ and $\widehat{\phi}^{\text{dow}}_{s,d}$ denote the estimates of the parameters $\phi^{\text{hod}}_{s,h}$ and $\phi^{\text{dow}}_{s,d}$,

respectively; their values depend on the tuning parameters $(r, \lambda, \lambda_N, \lambda_H)$; however, we omit this dependence for simplicity. Then, the networks associated with these estimates are defined as

$$\widehat{\mathfrak{N}}_{D,d} = (\mathcal{S}, \mathcal{E}_{D,d}) \quad \text{with} \quad \mathcal{E}_{D,d} = \left\{(s, s'),\, s \neq s' : \widehat{\phi}^{\text{dow}}_{s,d} = \widehat{\phi}^{\text{dow}}_{s',d}\right\} \quad \text{for each} \quad d \in \mathcal{D},$$
$$\widehat{\mathfrak{N}}_{H} = (\mathcal{S} \times H, \mathcal{E}_{H}) \quad \text{with} \quad \mathcal{E}_{H} = \cup_{h\in\mathcal{H}} \cup_{s\in\mathcal{S}} \left\{((s,h),(s',h')),\, (s,h) \neq (s',h') : \widehat{\phi}^{\text{hod}}_{s,h} = \widehat{\phi}^{\text{hod}}_{s',h'}\right\} \tag{8}$$

As with $\mathfrak{N}_{\text{single}}$, the networks $\widehat{\mathfrak{N}}_{D,d}$ are single-layer networks and an edge joins two station nodes s and s' when their node features (i.e. parameter estimates of $\phi^{\text{dow}}_{s,d}$ and $\phi^{\text{dow}}_{s',d}$ for a given $d \in \mathcal{D}$) are identical, possibly due to the fusion penalty but not necessarily so. The network $\widehat{\mathfrak{N}}_H$, as with $\mathfrak{N}_{\text{multi}}$, is a multilayer network with the hourly layer given by $\mathcal{H}$, and an edge is formed between a pair of nodes (s, h) and (s', h') when the estimates of $\phi^{\text{hod}}_{s,h}$ and $\phi^{\text{hod}}_{s',h'}$ agree at $(s, h) \neq (s', h')$. (For an illustrative example of $\widehat{\mathfrak{N}}_H$, see Figure 5).

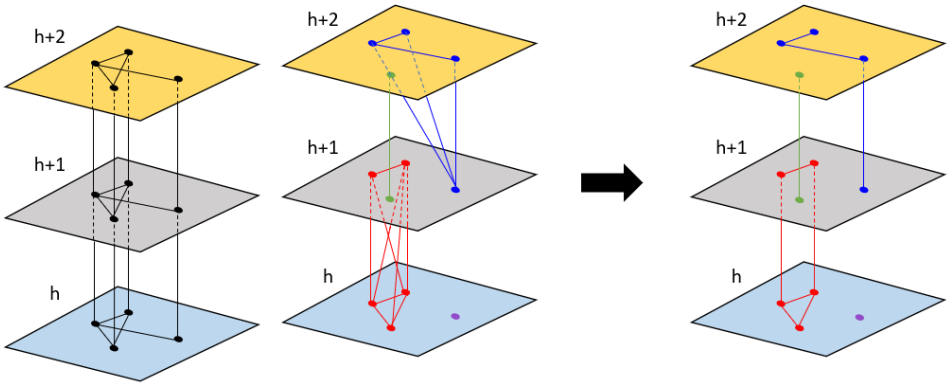

Figure 5: Illustration of multilayer networks. Left: Each layer of $\mathfrak{N}_{\text{multi}}$ (see (7)) embeds the linkages between the stations determined by their geographical distances at a given hour h, and the station at layer h is linked to itself at layers $h-1$ and $h+1$, which underpins how $\phi^{\text{hod}}_{s,h}$ are encouraged to be fused by p_N and p_H. Middle: $\widehat{\mathfrak{N}}_H$ (see (8)) is determined by the values of the estimates of $\phi^{\text{hod}}_{s,h}$ with an edge indicating that the connected estimates share the identical values. Right: A multilayer network formed with its edges obtained as an intersection of the edge sets of $\mathfrak{N}_{\text{multi}}$ and $\widehat{\mathfrak{N}}_H$, which contains four connected components.

For two networks (either single- or multilayer) $\mathfrak{N}_i = (\mathcal{V}, \mathcal{E}_i)$, $i = 1, 2$, sharing the same node set $\mathcal{V}$, denote by $\mathfrak{N}_1 \cap \mathfrak{N}_2 = (\mathcal{V}, \mathcal{E}_1 \cap \mathcal{E}_2)$ the network formed by taking the intersection of their edge sets. Our proposed penalized regression method takes as an input the observable networks $\mathfrak{N}_{\text{single}}$ and $\mathfrak{N}_{\text{multi}}$, which inform spatial and temporal proximity. Then, it outputs the networks capturing the homogeneity between the stations nodes, namely, $\mathfrak{N}_{\text{single}} \cap \widehat{\mathfrak{N}}_{D,d}$ (on a given day of a week d) and $\mathfrak{N}_{\text{multi}} \cap \widehat{\mathfrak{N}}_H$ (along the hourly layer), which can provide insights into BSS management and urban planning. Depending on the choice of penalty parameters, the output networks are not necessarily sparse; in fact, this is the case in our data

analysis reported in Section 4. This distinguishes our approach from the existing literature on clustering or partitioning the dataset using spatial or temporal variables prior to analysis.

4 Data analysis

4.1 Tuning parameter selection

For the selection of r that determines the set of neighbors $\mathcal{N}_r(s)$ for each station s, we examine the distance between each station and its 10 nearest stations, that is, $d_{s,i}$, $i = 1, \ldots, 10$, for all $s \in \mathcal{S}$. The median of the first quartile of $d_{s,i}$ over $s \in \mathcal{S}$ is 762 meters, from which we choose to consider $r \in \{750, 1500\}$. The network $\mathfrak{N}_{\text{single}}(r)$ (see (6)) with $r = 750$ has 34 connected components, whereas the network with $r = 1500$ has two connected components. Later, we show that our penalized regression approach is not sensitive to r on the BSS dataset through the adaptive selection of the penalty parameters discussed below.

The parameters λ, λ_N, and λ_H control the overall complexity of the fitted model, and the latter two, in particular control the degree of spatial and temporal homogeneity induced by the parameter estimates of Θ, Θ_H and Θ_D. We propose the selection of the tuning parameters via cross validation (CV). In the penalized regression literature, CV is typically performed by randomly partitioning the data into five or ten folds, which, in the case of the BSS dataset, ignores the temporal structure therein. Instead, we make use of the fact that our dataset covers seven weekends, and adopt seven-fold CV, in which each fold is ensured to include a balanced number of all seven days of the week.

As a CV measure, we adopt the mean squared Pearson residuals (MSPR):

$$\text{CV}(r, \lambda, \lambda_N, \lambda_H) = \frac{1}{7} \sum_{j=1}^{7} \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{(Y_i^{(j)} - \widehat{\mu}_i^{(j)}(r, \lambda, \lambda_N, \lambda_H))^2}{\widehat{\mu}_i^{(j)}(r, \lambda, \lambda_N, \lambda_H)} \tag{9}$$

where for the jth fold, n_j denotes the total number of observations, $Y_i^{(j)}$ the ith observation and $\widehat{\mu}_i^{(j)}(r, \lambda, \lambda_N, \lambda_H)$ is the corresponding estimate of the mean from the model fitted to the remaining data with the given tuning parameters. We evaluate $\text{CV}(r, \lambda, \lambda_N, \lambda_H)$ on grids of equispaced values on a log-scale for $(\lambda, \lambda_N, \lambda_H)$ and select the combination that returns the smallest CV.

4.2 Analysis of the BSS data

4.2.1 Effects of penalization

To assess the effect of penalization, we compare our multilayer network fused Lasso-based penalized regression approach (referred to as "fused Lasso") with the methods that adopt either the fusion penalty or the Lasso penalty only, which are referred to as "fusion-only" and

"Lasso-only". We also consider models (1) and (2) fitted without any penalization, referred to as "no-interaction" and "full-interaction", respectively.

Predictive performance Due to small sample size and inherent non-exchangeability of the data, we adopt CV measures as an indicator of the predictive performance. In Table 2 and Figure 6, we report the fold-wise MSPRs involved in the seven-fold CV in (9) as well as the overall CV error, all evaluated at the penalty parameters selected to minimize the overall CV error for the respective methods.

When comparing no-interaction and full-interaction methods, the flexibility afforded by allowing for station-specific temporal effects proves useful in enhancing the predictive performance as the latter model consistently attains a considerably smaller MSPR. We observe further improvement when appropriate penalization is applied to the interaction parameters. In particular, adopting the proposed fused Lasso penalty returns the minimum CV error across the seven folds compared to the other methods regardless of the choice of r. Between the fusion-only and the Lasso-only methods, the former outperforms the latter, indicating that capturing across-station homogeneity pays off by accounting for the stylized features of the BSS data that neighboring stations exhibit similar usage patterns, (see (C3)). The MSPR attained with $r = 1500$ is consistently smaller, albeit by a small margin, than that attained with $r = 750$. Recalling that the choice $r = 1500$ reduces the number of connected components in $\mathfrak{N}_{\text{single}}(r)$ compared to when $r = 750$ (from 54 to 2), this result is indicative of the benefit of pooling information across the wider spatial networks of BSSs.

Table 2: Mean squared Pearson residuals (MSPRs) from each fold used in the seven-fold CV and the overall CV error as measured in (9), with the penalty parameters selected to minimize the latter for the respective methods where relevant.

	Fold-wise CV							
Method	1	2	3	4	5	6	7	Average
Fused Lasso (750m)	1.558	1.423	1.424	1.497	1.449	1.831	1.683	1.552
Fused Lasso (1500m)	1.553	1.420	1.420	1.494	1.443	1.826	1.675	1.547
Fusion-only (750m)	1.572	1.443	1.445	1.512	1.473	1.861	1.701	1.569
Fusion-only (1500m)	1.568	1.440	1.438	1.511	1.463	1.847	1.700	1.567
Lasso-only	1.609	1.479	1.473	1.537	1.493	1.902	1.728	1.603
Full-interaction	1.729	1.601	1.556	1.629	1.596	2.016	1.841	1.710
No-interaction	1.991	1.857	1.791	1.918	1.838	2.264	2.118	1.969

Sparsity and model complexity We examine the reduction in complexity brought about by the Lasso penalty and the fusion penalties p_N and p_H in (3). The sparsity induced by the Lasso penalty is easily measured by the proportion of non-zero coefficient estimates (see the last column of Table 3). The model fitted with the fused Lasso penalty is approximately 13% (resp. 20%) less complex than the full-interaction model when $r = 750$ (resp. $r = 1500$),

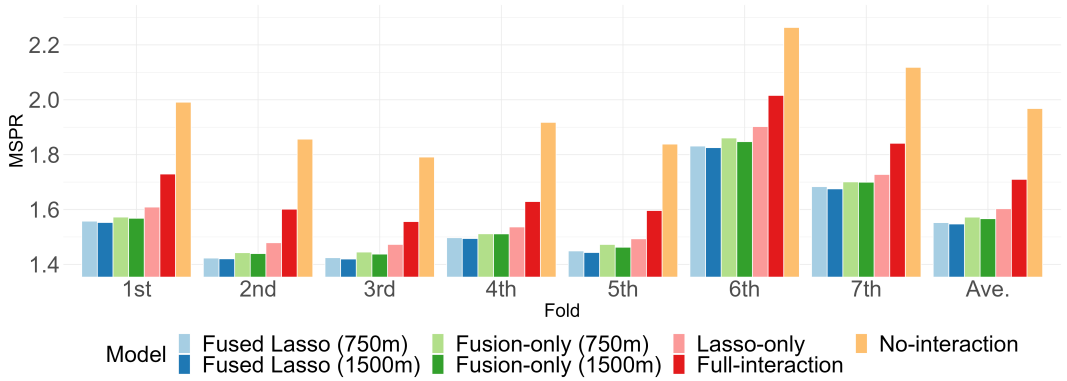

Figure 6: Mean squared Pearson residuals (MSPRs) from each fold used in the seven-fold CV and the overall CV error as measured in (9), with the penalty parameters selected to minimize the CV error for the respective methods.

yet the former has much better predictive power than the latter as Table 2 shows. At the same time, further sparsity induced by adopting the Lasso-only penalty does not improve the prediction performance. The fusion-only method does not set any parameter estimate to be exactly zero.

Table 3: Model complexity (MC) of the models fitted the penalty parameters selected to minimize the CV error for the respective methods (where relevant). We report the overall MC out of all parameters ($\Delta \cup \Theta \cup \Theta_H \cup \Theta_D$) as well as that out of the interaction parameters representing hour-of-a-day ($\phi^{\text{hod}}_{s,h}$) and day-of-a-week ($\phi^{\text{dow}}_{s,d}$) effects. Additionally, the proportion of non-zero coefficient estimates is presented.

	Parameter set			Proportion of
Method	All	$\Theta_H \setminus \{\theta^{\text{hod}}_h,\, h \in \mathcal{H}\}$	$\Theta_D \setminus \{\theta^{\text{dow}}_d,\, d \in \mathcal{D}\}$	non-zeros
Fused Lasso (750m)	0.828	0.790	0.977	0.867
Fused Lasso (1500m)	0.848	0.827	0.927	0.795
Fusion-only (750m)	0.903	0.885	0.974	1.000
Fusion-only (1500m)	0.853	0.836	0.916	1.000
Lasso-only	–	–	–	0.632

We use the networks introduced in Section 3.3 to define a model complexity (MC) measure that evaluates the effect of the fusion penalties. Simply put, the model represents the proportion of coefficient estimates that are not fused by the penalties p_N and p_H, out of the total number of parameters in model in (2), as

$$\text{MC}(r, \lambda, \lambda_N, \lambda_H) = \frac{1}{p}\left(34 + \mathcal{C}(\mathfrak{N}_{\text{multi}}(r) \cap \widehat{\mathfrak{N}}_H) + \sum_{d \in \mathcal{D}\setminus\{\text{Mo}\}} \mathcal{C}(\mathfrak{N}_{\text{single}}(r) \cap \widehat{\mathfrak{N}}_{D,d})\right) \tag{10}$$

where we suppress the dependence on $(r, \lambda, \lambda_N, \lambda_H)$ for notational simplicity. Recall that 34 represents the number of parameters that are not penalized, namely α, β^{rain}, $\{\beta^{\text{air}}_j,\, 1 \le j \le 3\}$,

$\{\theta_h^{\text{hod}}, h \in \mathcal{H}\}$ and $\{\theta_d^{\text{dow}}, d \in \mathcal{D}\}$. Denoting the number of connected components in a network $\mathfrak{N}$ by $\mathcal{C}(\mathfrak{N})$, the number of unique parameter estimates of $\phi_{s,h}^{\text{hod}}$, which are not fused by penalization, is given by $\mathcal{C}(\mathfrak{N}_{\text{multi}}(r) \cap \widehat{\mathfrak{N}}_H)$. Figure 5 shots that the network $\mathfrak{N}_{\text{multi}} \cap \widehat{\mathfrak{N}}_H$ contains four connected components; consequently, $\mathcal{C}(\mathfrak{N}_{\text{multi}} \cap \widehat{\mathfrak{N}}_H) = 4$. Similarly, We can find the number of unique parameter estimates for $\phi_{s,d}^{\text{dow}}$. In (10), we exclude the intersection network $\mathfrak{N}_{\text{single}}(r) \cap \widehat{\mathfrak{N}}_{D,\text{Mo}}$ from the numerator since, due to model identifiability constraints, we have $\phi_{s,\text{Mo}}^{\text{dow}} = \phi_{s,0}^{\text{hod}} = \theta_s$. That is, the fusion among the station-specific intercept parameters θ_s has already been accounted for by $\mathfrak{N}_{\text{multi}}(r) \cap \widehat{\mathfrak{N}}_H$ at layer $h = 0$. We refer to Appendix A.2 for the efficient calculation of MC which requires some care owing to the multilayered nature of the network in (7) underlying the penalties. We analogously evaluate the complexity of the model fitted using the fusion-only method (for the full results, see Table 3).

The MC ranges between 0 and 1, and when its value is closer to 0, it implies that most stations exhibit homogeneous behavior with their neighboring stations. The resultant intersection networks $\mathfrak{N}_{\text{multi}}(r) \cap \widehat{\mathfrak{N}}_H$ and $\mathfrak{N}_{\text{single}}(r) \cap \widehat{\mathfrak{N}}_{D,d}$ are not highly homogeneous, as evidenced by the MC being close to one, that is, many stations exhibit individual behavior. In other words, station-specific parameters account for a large portion of the variation in bike usage, which supports modelling the data at the individual station level. The high degree of heterogeneity across the BSS can be attributed to the fact that Seoul is a large city with high population density, so that each station is associated with multiple aspects of usages. This indicates that partitioning stations into a handful of clusters may ignore the complex nodal features that drive the usage of bikes at each station. Instead, our proposed multilayer fused Lasso method imposes fusion penalties through the networks $\mathcal{N}_{\text{single}}(r)$ and $\mathcal{N}_{\text{multi}}(r)$, and learns the degree of homogeneity in bike usage patterns across the BSS in a data-driven manner.

Generally, more homogeneity is observed in the hour-of-a-day effects than in the day-of-a-week effects. Also, since the choice of $r = 1500$ leads to a highly connected $\mathfrak{N}_{\text{single}}(r)$, it leads to a fitted model with a larger MC than the choice of $r = 750$. Overall, the fused Lasso method tends to further reduce the MC compared to the fusion-only method. We attribute this to the fact that the Lasso penalty enhances the effect of the fusion penalty by coercing estimated values to be zeros.

4.2.2 Parameter estimates

There exist methods for performing inference in high-dimensional generalized linear models (Belloni et al., 2016; Cai et al., 2021), but they do not easily apply to our setting because of the presence of a fusion penalty. Instead, we examine the coefficient estimates obtained from different folds used in the cross validation step in (9) (for the estimates of the parameters associated with the overall trend and the effects of the precipitation and the air quality, obtained from each fold as well as from the full data, see Table 4). We only report the results when $r = 1500$ in the main text because we obtain nearly identical estimates when $r = 750$

as shown in Table B.1 in Appendix B). Table 4 shows that, while the values of the estimates vary slightly from one fold to another, their signs and overall magnitude do not change. This confirms the observations (C1) and (C4) made in Section 2, that is, the variables have meaningful effects on overall bike usage across the system.

Table 4: Estimated coefficients for the covariate effects by the proposed fused Lasso regression method from each fold used in the seven-fold CV and from the full data when $r = 1500$. For comparison, we also report the estimates obtained with fusion Lasso-only, full-interaction and no-interaction methods.

	Fold							Full data				
	1	2	3	4	5	6	7	Fused	Fusion-only	Lasso-only	Full	No
α	0.057	0.052	0.051	0.050	0.050	0.054	0.050	0.052	0.052	0.052	0.052	0.052
β^{rain}	-2.301	-2.356	-2.380	-2.376	-2.238	-2.182	-2.382	-2.324	-2.325	-2.321	-2.323	-2.310
β_1^{air}	0.098	0.141	0.100	0.177	0.124	0.169	0.168	0.140	0.141	0.138	0.138	0.137
β_2^{air}	0.117	0.174	0.099	0.177	0.133	0.159	0.188	0.149	0.150	0.145	0.144	0.146
β_3^{air}	0.218	0.285	0.251	0.169	0.259	0.321	0.310	0.275	0.278	0.271	0.276	0.243

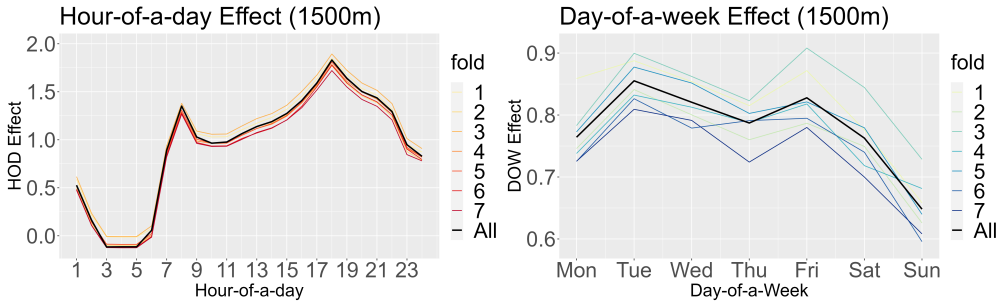

Figure 7: Parameter estimates for ϕ_h^{hod}, $h \in \mathcal{H}$ (left) and ϕ_d^{dow}, $d \in \mathcal{D}$ (right) from each fold used in the seven-fold CV and from the full data when $r = 1500$.

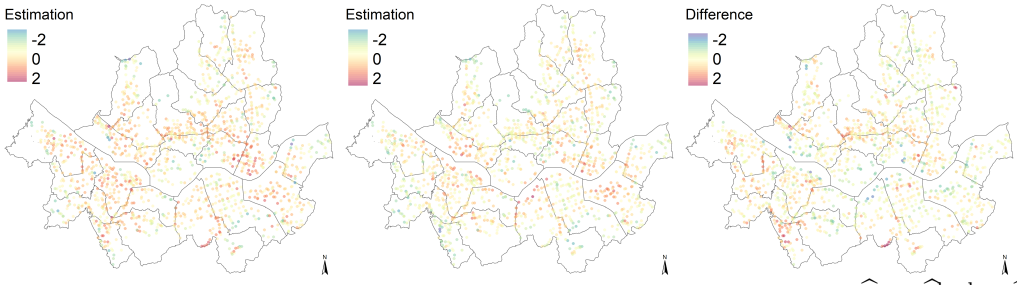

Figure 8: Estimated station-specific bike demands in log-scale (given by $\widehat{\theta}_s + \widehat{\theta}_h^{\text{hod}} + \widehat{\theta}_d^{\text{dow}} + \widehat{\theta}_{s,h}^{\text{hod}} + \widehat{\theta}_{s,d}^{\text{dow}}$) from the model fitted with $r = 1500$ at 8am on Tuesdays (left), at 8pm on Sundays (middle) and their differences (right).

In Figure 7, we plot the estimates of θ_h^{hod} and θ_d^{dow} which are shared by all the stations belonging to the BSS. We observe that the smooth transition over the course of a day and a week is well captured across the seven folds, and there are peaks corresponding to the high demand by commuters. In addition, we plot the combined effects of temporal variables on the

mean bike demand (in log-sale), namely $\widehat{\theta}_s + \widehat{\theta}_h^{\text{hod}} + \widehat{\theta}_d^{\text{dow}} + \widehat{\theta}_{s,h}^{\text{hod}} + \widehat{\theta}_{s,d}^{\text{dow}}$ under (2), for all stations $s \in \mathcal{S}$ when $(h, d) = (8, \text{Tu})$ (8am on Tuesday) and $(20, \text{Su})$ (8pm on Sunday) (see Figure 8). As expected the spatial distribution of bike usage concentration is markedly different when $(h, d) = (8, \text{Tu})$ and $(h, d) = (20, \text{Su})$, as bikes are primarily used for commuting at 8am on Tuesdays and for leisure activities at 8pm on Sundays. In Figures 7 and 8, we report the results obtained with $r = 1500$ only. (For those obtained with $r = 750$, indicating that the parameter estimates are not sensitive to the choice of r, we refer to Figures B.1 and B.2 in Appendix B.)

5 Conclusions

In this study, we examine the problem of modelling usage data collected from a BSS spanning a large metropolitan city. We model the data at the granularity of individual stations by incorporating covariate effects as well as spatial and temporal characteristics commonly observed in bike usage data. The proposed multilayer fused Lasso penalty is imposed on the networks encoding the geographical proximity of the stations with an additional hourly layer, and successfully captures the spatial and temporal homogeneity. Combined with the data-driven choice of penalty parameters, our penalized regression approach strikes a good balance between a simplistic model that does not allow for station-specific behavior, and a complex model suffering from high dimensionality, by returning a fitted model with good predictive performance.

We envision that the proposed method is applicable to different datasets with network-like features, such as those collected from large transportation, communication, or logistic systems. In particular, when information about the factors driving link homophily (such as the nodal features related to land use, slope of terrain, nearby landmarks, and other modes of transportation in the case of BSSs) is not readily available, our penalized regression approach enables learning of the linkages in an unobservable network from the fusion of parameters induced by the penalties defined on an observable network (such as those determined by the distance between pairs of stations).

Acknowledgement

Yunjin Choi was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT, No. NRF-2020R1G1A1A01014088). Haeran Cho was supported by the Leverhulme Trust Research Project Grant (RPG-2019-390).

References

Austwick, M. Z., O'Brien, O., Strano, E., and Viana, M. (2013). The structure of spatial networks and communities in bicycle sharing systems. *PloS one*, 8(9):e74685.

Belloni, A., Chernozhukov, V., and Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4):606–619.

Borgnat, P., Robardet, C., Abry, P., Flandrin, P., Rouquier, J.-B., and Tremblay, N. (2013). A dynamical network view of lyon's vélo'v shared bicycle system. In *Dynamics On and Of Complex Networks, Volume 2*, pages 267–284. Springer.

Boyd, S., Parikh, N., and Chu, E. (2011). *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Now Publishers Inc.

Cai, T. T., Guo, Z., and Ma, R. (2021). Statistical inference for high-dimensional generalized linear models with binary outcomes. *Journal of the American Statistical Association (to appear)*.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical Lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373.

Etienne, C. and Latifa, O. (2014). Model-based count series clustering for bike sharing system usage mining: a case study with the vélib'system of paris. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):1–21.

Faghih-Imani, A. and Eluru, N. (2016). Incorporating the impact of spatio-temporal interactions on bicycle sharing system demand: A case study of new york citibike system. *Journal of Transport Geography*, 54:218–227.

Fishman, E. (2016). Bikeshare: A review of recent literature. *Transport Reviews*, 36(1):92–113.

Fishman, E., Washington, S., and Haworth, N. (2014). Bike share's impact on car use: Evidence from the united states, great britain, and australia. *Transportation Research Part D: Transport and Environment*, 31:13–20.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., and Qian, J. (2020). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. R package version 4.0-2.

Froehlich, J. E., Neumann, J., and Oliver, N. (2009). Sensing and predicting the pulse of the city through shared bicycling. In *Twenty-First International Joint Conference on Artificial Intelligence*.

Gervini, D. and Khanal, M. (2019). Exploring patterns of demand in bike sharing systems via replicated point process models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3):585–602.

Hamilton, T. L. and Wichman, C. J. (2018). Bicycle infrastructure and traffic congestion: Evidence from dc's capital bikeshare. *Journal of Environmental Economics and Management*, 87:72–93.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). Multilayer networks. *Journal of Complex Networks*, 2(3):203–271.

Li, F. and Sang, H. (2019). Spatial homogeneity pursuit of regression coefficients for large datasets. *Journal of the American Statistical Association*, 114:1050–1062.

Liu, J., Sun, L., Chen, W., and Xiong, H. (2016). Rebalancing bike sharing systems: A multi-source data smart optimization. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1005–1014.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall.

Sass, D., Li, B., and Reich, B. J. (2021). Flexible and fast spatial return level estimation via a spatially fused penalty. *Journal of Computational and Graphical Statistics*, 30(4):1124–1142.

Shaheen, S. A., Guzman, S., and Zhang, H. (2010). Bikesharing in europe, the americas, and asia: past, present, and future. *Transportation research record*, 2143(1):159–167.

Sun, Y., Wang, H. J., and Fuentes, M. (2016). Fused adaptive Lasso for spatial and temporal quantile function estimation. *Technometrics*, 58(1):127–137.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:91–108.

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323.

Torti, A., Pini, A., and Vantini, S. (2021). Modelling time-varying mobility flows using function-on-function regression: Analysis of a bike sharing system in the city of Milan. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(1):226–247.

Vogel, P., Greiser, T., and Mattfeld, D. C. (2011). Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia-Social and Behavioral Sciences*, 20:514–523.

Wahlberg, B., Boyd, S., Annergren, M., and Wang, Y. (2012). An admm algorithm for a class of total variation regularized estimation problems. *IFAC Proceedings Volumes*, 45(16):83–88.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhou, X. (2015). Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in chicago. *PloS one*, 10(10):e0137922.

A Computational considerations

A.1 ADMM algorithm for optimization of penalized MLE

A.1.1 ADMM framework

We adopt the alternating direction method of multipliers (ADMM) algorithm for fitting (3) based on Wahlberg et al. (2012), where the optimization of objective functions with fusion penalties is discussed. We recall that the fused Lasso penalties lead to the objective function of the form

$$\begin{aligned}
&\min_{\Theta,\Theta_H,\Theta_D,\Delta} \sum_{i=1}^{n} \mu_i(\Theta,\Theta_H,\Theta_D,\Delta) - \sum_{i=1}^{n} y_i \log(\mu_i(\Theta,\Theta_H,\Theta_D,\Delta)) \\
&\quad + \lambda \sum_{s\in\mathcal{S}} \left(\sum_{h\in\mathcal{H}} |\theta^{\text{hod}}_{s,h}| + \sum_{d\in\mathcal{D}} |\theta^{\text{dow}}_{s,d}| \right) \\
&\quad + \lambda_N \sum_{s\in\mathcal{S}} \sqrt{|\mathcal{N}_r(s)| \sum_{s'\in\mathcal{N}_r(s)} \left[2(\theta_s - \theta_{s'})^2 + \sum_{h=1}^{23} (\phi^{\text{hod}}_{s,h} - \phi^{\text{hod}}_{s',h})^2 + \sum_{d\in\mathcal{D}\setminus\{\text{Mo}\}} (\phi^{\text{dow}}_{s,d} - \phi^{\text{dow}}_{s',d})^2 \right]} \\
&\quad + \lambda_H \sum_{s\in\mathcal{S}} \sum_{h=0}^{23} \left| \phi^{\text{hod}}_{s,h} - \phi^{\text{hod}}_{s,h+1} \right| . \qquad \text{(A.1)}
\end{aligned}$$

In this section, we treat Θ, Θ_H and Θ_D as row-vectors without confusion:

$$\begin{aligned}
\Theta &= (\theta_s,\ s\in\mathcal{S}), \\
\Theta_H &= \left((\theta^{\text{hod}}_{s,1},\ s\in\mathcal{S}), \ldots, (\theta^{\text{hod}}_{s,23},\ s\in\mathcal{S}),\ (\theta^{\text{hod}}_{1}, \ldots, \theta^{\text{hod}}_{23}) \right), \\
\Theta_D &= \left((\theta^{\text{dow}}_{s,\text{Mo}},\ s\in\mathcal{S}), \ldots, (\theta^{\text{dow}}_{s,\text{Su}},\ s\in\mathcal{S}),\ (\theta^{\text{dow}}_{\text{Mo}}, \ldots, \theta^{\text{dow}}_{\text{Su}}) \right).
\end{aligned}$$

By re-parametrizing the fused lasso penalty terms, we re-write (A.1) as,

$$\begin{aligned}
&\min_{\Theta,\Theta_H,\Theta_D,\Delta,\Gamma,\Psi} \sum_{i=1}^{n} \mu_i(\Theta,\Theta_H,\Theta_D,\Delta) - \sum_{i=1}^{n} y_i \log(\mu_i(\Theta,\Theta_H,\Theta_D,\Delta)) \\
&\qquad + \lambda \sum_{s\in\mathcal{S}} \left(\sum_{h\in\mathcal{H}} |\theta^{\text{hod}}_{s,h}| + \sum_{d\in\mathcal{D}} |\theta^{\text{dow}}_{s,d}| \right) \\
&\qquad + \lambda_N \cdot \widetilde{P}_N(\Gamma) + \widetilde{P}_H(\Psi) + \mathbf{I}_{\mathcal{C}}\left(\Theta,\Theta_H,\Theta_D,\Gamma,\Psi\right). \qquad \text{(A.2)}
\end{aligned}$$

Here, the penalty functions are defines as

$$\widetilde{P}_N(\Gamma) = \sum_{s\in\mathcal{S}} \sqrt{|\mathcal{N}_r(s)| \sum_{s'\in\mathcal{N}_r(s)} \left[\left(\gamma_{s,s'}\right)^2 + \sum_{h=1}^{23} \left(\gamma^{\text{hod}}_{s,s',h}\right)^2 + \sum_{d\in\mathcal{D}\setminus\{\text{Mo}\}} \left(\gamma^{\text{dow}}_{s,s',d}\right)^2 \right]},$$

where $\Gamma = (\boldsymbol{\gamma}_s,\ s\in\mathcal{S})$ with $\boldsymbol{\gamma}_s = \left(\gamma_{s,s'}, (\gamma^{\text{hod}}_{s,s',h},\ h\in\mathcal{H}\setminus\{0\}), (\gamma^{\text{dow}}_{s,s',d},\ d\in\mathcal{D}\setminus\{\text{Mo}\})),\ s'\in\mathcal{N}_r(s)\right)$,

and

$$\widetilde{P}_H(\Psi) = \sum_{s\in\mathcal{S}} \sum_{h\in\{0,\ldots,23\}} |\psi_{s,h}| \quad \text{with} \quad \Psi = (\boldsymbol{\psi}_s = (\psi_{s,0},\ldots,\psi_{s,23}),\ s\in\mathcal{S})\,.$$

In addition, $I_{\mathcal{C}}$ denotes an indicator function on the constraint set $\mathcal{C}$ which forces the newly defined penalty functions $\widetilde{P}_N(\Gamma)$ (resp. $\widetilde{P}_H(\Psi)$) to be the same as $P_N(\Theta,\Theta_H,\Theta_D)$ (resp. $P_H(\Theta,\Theta_H,\Theta_D)$) such that

$$I_{\mathcal{C}}(\mathbf{z}) = \begin{cases} 0 & \text{if } \mathbf{z}\in\mathcal{C} \\ \infty & \text{otherwise.} \end{cases}$$

The constraint set $\mathcal{C}$ is a subset of $|(\Theta,\Theta_H,\Theta_D,\Gamma,\Psi)|$-dimensional space whose elements fulfil

$$\begin{aligned} \gamma_{s,s'} &= \sqrt{2}(\theta_s - \theta_{s'}), \\ \gamma^{\text{hod}}_{s,s',h} &= \theta_s - \theta_{s'} + \theta^{\text{hod}}_{s,h} - \theta^{\text{hod}}_{s',h} \ \text{ for } h\in\mathcal{H}\setminus\{0\}, \\ \gamma^{\text{dow}}_{s,s',d} &= \theta_s - \theta_{s'} + \theta^{\text{dow}}_{s,d} - \theta^{\text{dow}}_{s',d} \ \text{ for } d\in\mathcal{D}\setminus\{\text{Mo}\} \end{aligned} \tag{A.3}$$

for all $s, s'\in\mathcal{S}$, and

$$\begin{aligned} \psi_{s,h} &= \theta^{\text{hod}}_{h+1} - \theta^{\text{hod}}_{h} + \theta^{\text{hod}}_{s,h+1} - \theta^{\text{hod}}_{s,h} \ \text{ for } h\in\{1,\ldots 22\}, \\ \psi_{s,0} &= \theta^{\text{hod}}_{1} + \theta^{\text{hod}}_{s,1} \quad \text{and} \quad \psi_{s,23} = -\theta^{\text{hod}}_{23} - \theta^{\text{hod}}_{s,23}, \end{aligned} \tag{A.4}$$

for all $s\in\mathcal{S}$, where (A.3) and (A.4) account for that the baseline parameters are set to be zero.

To utilize ADMM, we rewrite the objective function (A.2) as follows:

$$\begin{aligned} \min_{\substack{\Theta,\Theta_H,\Theta_D,\Delta,\Gamma,\Psi,\\ Z_\Theta,Z_{\Theta_H},Z_{\Theta_D},S_\Gamma,S_\Psi}} & \sum_{i=1}^{n} \mu_i(\Theta,\Theta_H,\Theta_D,\Delta) - \sum_{i=1}^{n} y_i \log(\mu_i((\Theta,\Theta_H,\Theta_D,\Delta))) \\ & + \lambda \sum_{s\in\mathcal{S}} \left(\sum_{h\in\mathcal{H}} |\theta^{\text{hod}}_{s,h}| + \sum_{d\in\mathcal{D}} |\theta^{\text{dow}}_{s,d}| \right) \\ & + \lambda_N \cdot \widetilde{P}_N(\Gamma) + \widetilde{P}_H(\Psi) + \mathbf{I}_{\mathcal{C}}\left(Z_\Theta, Z_{\Theta_H}, Z_{\Theta_D}, S_\Gamma, S_\Psi\right) \\ \text{subject to} \quad & (\Theta,\Theta_H,\Theta_D) = (Z_\Theta, Z_{\Theta_H}, Z_{\Theta_D}) \text{ and } (\Gamma,\Psi) = (S_\Gamma, S_\Psi). \end{aligned} \tag{A.5}$$

The ADMM optimizes (A.5) in three steps as follows:

Step 1: Update the primal variables as

$$(\Theta^{k+1},\Theta_H^{k+1},\Theta_D^{k+1},\Delta^{k+1}) = \underset{\Theta,\Theta_H,\Theta_D,\Delta}{\arg\min} \sum_{i=1}^{n} \mu_i(\Theta,\Theta_H,\Theta_D,\Delta) - \sum_{i=1}^{n} y_i \log(\mu_i((\Theta,\Theta_H,\Theta_D,\Delta)))$$

$$+\lambda\sum_{s\in\mathcal{S}}\left(\sum_{h\in\mathcal{H}}|\theta^{\text{hod}}_{s,h}|+\sum_{d\in\mathcal{D}}|\theta^{\text{dow}}_{s,d}|\right)$$
$$+\frac{\rho}{2}\left\|(\Theta,\Theta_H,\Theta_D)-(Z^k_\Theta,Z^k_{\Theta_H},Z^k_{\Theta_D})+(U^k_\Theta,U^k_{\Theta_H},U^k_{\Theta_D})\right\|^2, \tag{A.6}$$

$$\Gamma^{k+1}=\arg\min_{\Gamma}\lambda_N\widetilde{P}_N(\Gamma)+\frac{\rho}{2}\left\|\Gamma-S^k_\Gamma+T^k_\Gamma\right\|^2, \tag{A.7}$$

$$\Psi^{k+1}=\arg\min_{\Psi}\lambda_H\widetilde{P}_N(\Psi)+\frac{\rho}{2}\left\|\Psi-S^k_\Psi+T^k_\Psi\right\|^2, \tag{A.8}$$

where $(U_\Theta,U_{\Theta_H},U_{\Theta_D})$, T_Γ, and T_Ψ are dual variables associated with the constraints $(\Theta,\Theta_H,\Theta_D)=(Z_\Theta,Z_{\Theta_H},Z_{\Theta_D})$, $\Gamma=S_\Gamma$ and $\Psi=S_\Psi$, respectively.

Step 2: Update $(Z_\Theta,Z_{\Theta_H},Z_{\Theta_D},S_\Gamma,S_\Psi)$ by projecting $(\Theta^{k+1}+U^k_\Theta,\Theta^{k+1}_H+U^k_{\Theta_H},\Theta^{k+1}_D+U^k_{\Theta_D},\Gamma^{k+1}+T^k_\Gamma,\Psi^{k+1}+T^k_\Psi)$ onto the constraint set $\mathcal{C}$, as

$$\begin{aligned}&(Z^{k+1}_\Theta,Z^{k+1}_{\Theta_H},Z^{k+1}_{\Theta_D},S^{k+1}_\Gamma,S^{k+1}_\Psi)\\=&\prod_{\mathcal{C}}(\Theta^{k+1}+U^k_\Theta,\Theta^{k+1}_H+U^k_{\Theta_H},\Theta^{k+1}_D+U^k_{\Theta_D},\Gamma^{k+1}+T^k_\Gamma,\Psi^{k+1}+T^k_\Psi),\end{aligned} \tag{A.9}$$

with $\prod_{\mathcal{C}}$ denoting the projection operator.

Step 3: Update the dual variable as

$$\begin{aligned}(U^{k+1}_\Theta,\ U^{k+1}_{\Theta_H},\ U^{k+1}_{\Theta_D})&=(U^k_\Theta,U^k_{\Theta_H},U^k_{\Theta_D})+(\Theta^{k+1},\Theta^{k+1}_H,\Theta^{k+1}_D)-(Z^{k+1}_\Theta,Z^{k+1}_{\Theta_H},Z^{k+1}_{\Theta_D}),\\(T^{k+1}_\Gamma,T^{k+1}_\Psi)&=(T^k_\Gamma,T^k_\Psi)+(\Gamma^{k+1},\Psi^{k+1})-(S^{k+1}_\Gamma,S^{k+1}_\Psi).\end{aligned}$$

While Step 3 is straightforward, Steps 1 and 2 involve relatively heavy computations. The detail of Step 1 and Step 2 are discussed in the following subsections.

A.1.2 Computational details of Step 1

Let $\mathcal{P}=(\Theta,\Theta_H,\Theta_D)$.

Step 1.1: We update $(\Theta,\Theta_H,\Theta_D,\Delta)$ by minimizing the objective function in (A.6). It in turn can be optimized via Iteratively Reweighted Least Square (IRLS) method with the Lasso penalty as below, at some given fixed values for $(Z^k_\Theta,Z^k_{\Theta_H},Z^k_{\Theta_D})$ and $(U^k_\Theta,U^k_{\Theta_H},U^k_{\Theta_D})$. The $j+1$th iteration of IRLS is as follows:

$$(\Theta^{j+1},\Theta^{j+1}_H,\Theta^{j+1}_D,\Delta^{j+1})=\arg\min_{\Theta,\Theta_H,\Theta_D,\Delta}\left(\mathbf{X}\begin{pmatrix}\Theta^\top\\\Theta^\top_H\\\Theta^\top_D\\\Delta^\top\end{pmatrix}-\mathbf{z}^j\right)^\top\mathbf{W}^j\left(\mathbf{X}\begin{pmatrix}\Theta^\top\\\Theta^\top_H\\\Theta^\top_D\\\Delta^\top\end{pmatrix}-\mathbf{z}^j\right)$$
$$+\lambda\sum_{s\in\mathcal{S}}\left(\sum_{h\in\mathcal{H}}|\theta^{\text{hod}}_{s,h}|+\sum_{d\in\mathcal{D}}|\theta^{\text{dow}}_{s,d}|\right)+\frac{\rho}{2}\left\|(\Theta,\Theta_H,\Theta_D)-(Z^k_\Theta,Z^k_{\Theta_H},Z^k_{\Theta_D})+(U^k_\Theta,U^k_{\Theta_H},U^k_{\Theta_D})\right\|^2, \tag{A.10}$$

where $\mathbf{X}$ is a data matrix of dimension $n \times (|\Delta| + (|\mathcal{D}| + |\mathcal{H}| - 1) \cdot |\mathcal{S}|)$. Also, $\mathbf{W}^j$ is an $n \times n$ diagonal matrix with its ith entry being $\widehat{\mu}_i(\Theta^j, {\Theta_H}^j, {\Theta_D}^j, \Delta^j)$, the fitted value of the ith observation after the jth iteration, and $\mathbf{z}^j$ is a length-n vector defined as follows:

$$\mathbf{z}^j = \mathbf{X}(\Theta^j, \Theta_H^j, \Theta_D^j, \Delta^j)^\top + \left(\mathbf{W}^j\right)^{-1} \mathbf{y} - \mathbb{1}_n.$$

Here, $\mathbf{y}$ denotes a length-n vector, the ith entry of which is the ith response observation, and $\mathbb{1}_n$ denots an all-one vector of length n. The objective function in (A.10) can further be written as the ℓ_1-penalized least squares estimation problem as follows:

$$\begin{aligned}
&(\Theta^{j+1}, \Theta_H^{j+1}, \Theta_D^{j+1}, \Delta^{j+1}) = \\
&\underset{\Theta, \Theta_H, \Theta_D, \Delta}{\arg\min} \left\| \mathbf{X}_{\text{ext}}^j (\Theta, \Theta_H, \Theta_D, \Delta)^\top - \mathbf{z}_{\text{ext}}^j \right\|^2 + \lambda \sum_{s \in \mathcal{S}} \left(\sum_{h \in \mathcal{H}} |\theta_{s,h}^{\text{hod}}| + \sum_{d \in \mathcal{D}} |\theta_{s,d}^{\text{dow}}| \right),
\end{aligned} \tag{A.11}$$

where $\mathbf{X}_{\text{ext}}^j$ and $\mathbf{z}_{\text{ext}}^j$ are a matrix of size $(n+|\mathcal{P}|) \times (|\mathcal{P}|+|\Delta|)$ and a vector of length $(n+|\mathcal{P}|)$, respectively, such that

$$\begin{aligned}
\mathbf{X}_{\text{ext}}^j &= \begin{pmatrix} \left(\mathbf{W}^j\right)^{1/2} \mathbf{X} \\ \hdashline \sqrt{\rho} I_{|\mathcal{P}|} \,\vdots\, \mathbf{0}_{|\mathcal{P}| \times |\Delta|} \end{pmatrix}, \\
\mathbf{z}_{\text{ext}}^j &= \begin{pmatrix} \left(\mathbf{W}^j\right)^{1/2} \mathbf{z}^j \\ \sqrt{\rho}(Z_\Theta^k, Z_{\Theta_H}^k, Z_{\Theta_D}^k)^\top - \sqrt{\rho}(U_\Theta^k, U_{\Theta_H}^k, U_{\Theta_D}^k)^\top \end{pmatrix}.
\end{aligned}$$

We evaluated (A.11) using the R package `glmnet` (Friedman et al., 2020).

Step 1.2: For (A.7)–(A.8), we obtain Γ^{k+1} and Ψ^{k+1} using a soft-thershold operator $\mathbb{S}_\lambda$ that takes an input vector and outputs $\mathbb{S}_\lambda(\mathbf{v}) = (1 - \lambda/\|\mathbf{v}\|)_+ \cdot \mathbf{v}$ with $\mathbb{S}_\lambda(\mathbf{0}) = \mathbf{0}$ and $c_+ = \max\{0, c\}$.

$$\begin{aligned}
\boldsymbol{\gamma}_s^{k+1} &= \mathbb{S}_{\rho^{-1}\sqrt{\mathcal{N}_r(s)}\lambda_N} \left(S_{\boldsymbol{\gamma}_s}^k - T_{\boldsymbol{\gamma}_s}^k \right) \quad \text{for each} \quad s \in \mathcal{S}, \\
\psi_{s,h}^{k+1} &= \mathbb{S}_{\rho^{-1}\lambda_H}(S_{\Psi,s,h}^k - T_{\Psi,s,h}^k) \ \text{ for } \ s \in \mathcal{S}, \ h \in \mathcal{H}.
\end{aligned}$$

A.1.3 Computational details of Step 2

In Step 2, the update of $(Z_\Theta, Z_{\Theta_D}, Z_{\Theta_H}, S_\Gamma, S_\Psi)$ is achieved via projection in (A.9), where $\mathcal{C}$ is the constraint set specified in (A.3) and (A.4). This step is the bottleneck of the computation due to the large number of variables to be updated. In this section, the details of the procedure is illustrated. Throughout, we denote by $\mathbf{I}$ and $\mathbf{0}$ an identity matrix and a matrix of zeros, respectively, and their dimensions are determined by the context unless specified.

The projection $(A_1, A_2, A_3, B_1, B_2) = \prod_{\mathcal{C}}(E_1, E_2, E_3, F_1, F_2)$ is equivalent to the following

minimization problem

$$\min_{A_1,A_2,A_3,B_1,B_2} \|A_1 - E_1\|^2 + \|A_2 - E_2\|^2 + \|A_3 - E_3\|^2 + \|B_1 - F_1\|^2 + \|B_2 - F_2\|^2$$
$$\text{subject to} \quad B_1 = (A_1, A_2, A_3)\mathbf{D}_\Theta^\top \text{ and } B_2 = (A_1, A_2, A_3)\mathbf{D}_\mathcal{H}^\top,$$

where $\mathbf{D}_\Theta$ and $\mathbf{D}_\mathcal{H}$ are matrices encoding the constraints (A.3) and (A.4), respectively. Then, the above optimization problem can be re-written as

$$\min_{A_1,A_2,A_3} \|A_1 - E_1\|^2 + \|A_2 - E_2\|^2 + \|A_3 - E_3\|^2 + \|(A_1, A_2, A_3)\mathbf{D}_\Theta^\top - F_1\|^2 + \|(A_1, A_2, A_3)\mathbf{D}_\mathcal{H}^\top - F_2\|^2,$$

and its optimizer is the solution of its normal equation

$$(A_1, A_2, A_3)\underbrace{\left(\mathbf{I} + \mathbf{D}_\Theta^\top\mathbf{D}_\Theta + \mathbf{D}_\mathcal{H}^\top\mathbf{D}_\mathcal{H}\right)}_{\mathbf{P}} = (E_1, E_2, E_3) + F_1\mathbf{D}_\Theta + F_2\mathbf{D}_\mathcal{H}. \tag{A.12}$$

Once the inverse of $\mathbf{P}$ is available, the solution (A_1, A_2, A_3) of (A.12) can be calculated in a straightforward manner. Also, the inverse matrix remains the same throughout the iterations and thus no re-computation is required. In our problem, however, as the size of the matrix $\mathbf{P}$ is huge, its dimension reaching approximately $47,000 \times 47,000$, and inverting this matrix can be very demanding with the computational complexity of $\mathcal{O}(10^{11})$. Additionally, even if we can compute the inverse matrix, it is huge in size and occupies a large portion of memory space which hinders efficient computation. Given the situation, we avoid direct computation of $\mathbf{P}^{-1}$ and find the solution of (A.12) by utilizing the specific structure of the matrix $\mathbf{P}$.
We start by defining an $M \times |\mathcal{S}|$ matrix $\mathbf{D}_{\text{net}}$ to be a matrix that represents the network constructed from the neighborhood relations so that each row is associated with two connected stations where $M = \sum_{s\in\mathcal{S}} |\mathcal{N}_r(s)|$. Defining $\mathcal{R}_{\text{ind}} : \mathcal{S} \times \mathcal{S} \to \{1, \ldots, M\}$ to be the mapping that returns the row index of $s - s'$ connection for $s' \in \mathcal{N}_r(s)$, the $\mathcal{R}_{\text{ind}}(s, s')$-th row of $\mathbf{D}_{\text{net}}$ is given by $\mathbf{e}_s - \mathbf{e}_{s'}$, where $\mathbf{e}_i$ is a standard basis vector of length $|\mathcal{S}|$. That is, each row of $\mathbf{D}_{\text{net}}$ is composed of $\{-1, 0, 1\}$ with exactly one 1 and one -1 and the rest of the entries are all 0s. By construction, the Laplacian matrix, say $\mathbf{L}_{\text{net}}$, of the network can be represented by $\mathbf{D}_{\text{net}}$ as follows:

$$\mathbf{L}_{\text{net}} = \frac{1}{2}\mathbf{D}_{\text{net}}^\top\mathbf{D}_{\text{net}}.$$

Then, we have

$$\mathbf{D}_{\text{net}}\Theta^\top = (\theta_s - \theta_{s'},\ s' \in \mathcal{N}_r(s),\ s \in \mathcal{S})^\top,$$
$$\mathbf{D}_{\text{net}}(\Theta^\circ_{H,h})^\top = (\theta^{\text{hod}}_{s,h} - \theta^{\text{hod}}_{s',h},\ s' \in \mathcal{N}_r(s),\ s \in \mathcal{S})^\top,$$
$$\mathbf{D}_{\text{net}}(\Theta^\circ_{D,d})^\top = (\theta^{\text{dow}}_{s,d} - \theta^{\text{hod}}_{s',d},\ s' \in \mathcal{N}_r(s),\ s \in \mathcal{S})^\top,$$

where $\Theta^{\circ}_{H,h} = (\theta^{\text{hod}}_{s,h},\, s \in \mathcal{S})$ and $\Theta^{\circ}_{D,d} = (\theta^{\text{dow}}_{s,d},\, d \in \mathcal{D})$.
The matrix $\mathbf{D}_\Theta$ can be written with $\mathbf{D}_{\text{net}}$ as

$$
\begin{array}{r}
 \\
M \cdot |\mathcal{D}^\circ| \\
 \\
 \\
M \cdot |\mathcal{H}^\circ| \\
 \\
 \\
M
\end{array}
\overbrace{\phantom{\begin{array}{ccc}\mathbf{D}_{\text{net}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_{\text{net}}\end{array}}}^{|\mathcal{S}| \cdot |\mathcal{D}^\circ|}
\overbrace{\phantom{\begin{array}{c}\mathbf{0} \\ \mathbf{0} \\ \mathbf{0}\end{array}}}^{|\mathcal{D}^\circ|}
\overbrace{\phantom{\begin{array}{c}-\mathbf{D}_{\text{net}} \\ \vdots \\ -\mathbf{D}_{\text{net}}\end{array}}}^{|\mathcal{S}|}
\overbrace{\phantom{\begin{array}{ccc}\mathbf{D}_{\text{net}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_{\text{net}}\end{array}}}^{|\mathcal{S}| \cdot |\mathcal{H}^\circ|}
\overbrace{\phantom{\begin{array}{c}\mathbf{0} \\ \mathbf{0} \\ \mathbf{0}\end{array}}}^{|\mathcal{H}^\circ|}
\left(\begin{array}{ccc|c|c|ccc|c}
\mathbf{D}_{\text{net}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{D}_{\text{net}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} & \vdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathbf{D}_{\text{net}} & \mathbf{0} & -\mathbf{D}_{\text{net}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\hline
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{D}_{\text{net}} & \mathbf{D}_{\text{net}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \vdots & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{D}_{\text{net}} & \mathbf{0} & \mathbf{0} & \mathbf{D}_{\text{net}} & \mathbf{0} \\
\hline
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\sqrt{2}\mathbf{D}_{\text{net}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}
\end{array}\right)
$$

where $\mathcal{D}^\circ = \mathcal{D} \setminus \{\text{Mo}\}$ and $\mathcal{H}^\circ = \mathcal{H} \setminus \{0\}$ so that the baseline parameters are removed.
Denoting $\frac{1}{2}\mathbf{L}_{\text{net}}$ by $\widetilde{\mathbf{L}}_{\text{net}}$, the form of $\mathbf{D}_\Theta^\top \mathbf{D}_\Theta$ is as below:

$$
\begin{array}{c}
\overbrace{\qquad\qquad\qquad}^{|\mathcal{S}|\cdot|\mathcal{D}^\circ|} \quad \overbrace{\quad}^{|\mathcal{D}^\circ|} \quad \overbrace{\qquad\qquad}^{|\mathcal{S}|} \quad \overbrace{\qquad\qquad\qquad}^{|\mathcal{S}|\cdot|\mathcal{H}^\circ|} \quad \overbrace{\quad}^{|\mathcal{H}^\circ|} \\
\left(\begin{array}{ccc:c:c:ccc:c}
\widetilde{\mathbf{L}}_{\text{net}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\widetilde{\mathbf{L}}_{\text{net}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} & \vdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \widetilde{\mathbf{L}}_{\text{net}} & \mathbf{0} & -\widetilde{\mathbf{L}}_{\text{net}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\hdashline
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\hdashline
-\widetilde{\mathbf{L}}_{\text{net}} & \cdots & -\widetilde{\mathbf{L}}_{\text{net}} & \mathbf{0} & (|\mathcal{D}^\circ|+|\mathcal{H}^\circ|)\widetilde{\mathbf{L}}_{\text{net}} & -\widetilde{\mathbf{L}}_{\text{net}} & \cdots & -\widetilde{\mathbf{L}}_{\text{net}} & \mathbf{0} \\
\hdashline
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\widetilde{\mathbf{L}}_{\text{net}} & \widetilde{\mathbf{L}}_{\text{net}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \vdots & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\widetilde{\mathbf{L}}_{\text{net}} & \mathbf{0} & \mathbf{0} & \widetilde{\mathbf{L}}_{\text{net}} & \mathbf{0} \\
\hdashline
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}
\end{array}\right)
\end{array}
$$

Now, we define $\mathbf{D}_{\mathcal{H}}$, the difference matrix that addresses the association between consecutive hours of each of the stations so that

$$
\mathbf{D}_{\mathcal{H}}(\Theta, \Theta_D, \Theta_H)^\top = \left(\begin{array}{c}
(-(\theta_1^{\text{hod}} + \theta_{s,1}^{\text{hod}}),\, s \in \mathcal{S})^\top \\
\hdashline
((\theta_1^{\text{hod}} + \theta_{s,1}^{\text{hod}}) - (\theta_2^{\text{hod}} + \theta_{s,2}^{\text{hod}}),\, s \in \mathcal{S})^\top \\
\vdots \\
((\theta_{22}^{\text{hod}} + \theta_{s,22}^{\text{hod}}) - (\theta_{23}^{\text{hod}} + \theta_{s,23}^{\text{hod}}),\, s \in \mathcal{S})^\top \\
\hdashline
((\theta_{23}^{\text{hod}} + \theta_{s,23}^{\text{hod}})^\top,\, s \in \mathcal{S})^\top
\end{array}\right).
$$

The specific form of $\mathbf{D}_{\mathcal{H}}$ is as follows:

$$
\begin{array}{c}
\\
|\mathcal{H}| \\
\\
\\
\\
\\
|\mathcal{H}|\cdot|\mathcal{S}| \\
\\
\\
\\
\\
\end{array}
\overbrace{\phantom{\begin{matrix}\mathbf{0}&\mathbf{0}&\mathbf{0}\end{matrix}}}^{|\mathcal{S}|\cdot|\mathcal{D}^\circ|}
\overbrace{\phantom{\begin{matrix}\mathbf{0}\end{matrix}}}^{|\mathcal{D}^\circ|}
\overbrace{\phantom{\begin{matrix}\mathbf{0}\end{matrix}}}^{|\mathcal{S}|}
\overbrace{\phantom{\begin{matrix}\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}\end{matrix}}}^{|\mathcal{S}|\cdot|\mathcal{H}^\circ|}
\overbrace{\phantom{\begin{matrix}-1&0&0&0\end{matrix}}}^{|\mathcal{H}^\circ|}
\left(\begin{array}{ccc:c:c:cccc:cccc}
\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&-1&0&0&0\\
\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&1&-1&0&0\\
\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&0&\ddots&\ddots&0\\
\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&0&0&1&-1\\
\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&0&0&0&1\\
\hdashline
\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&-\mathbf{I}_{|\mathcal{S}|}&\mathbf{0}&\mathbf{0}&\mathbf{0}&-\mathbb{1}_{|\mathcal{S}|}&\mathbf{0}&\mathbf{0}&\mathbf{0}\\
\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{I}_{|\mathcal{S}|}&-\mathbf{I}_{|\mathcal{S}|}&\mathbf{0}&\mathbf{0}&\mathbb{1}_{|\mathcal{S}|}&-\mathbb{1}_{|\mathcal{S}|}&\mathbf{0}&\mathbf{0}\\
\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\ddots&\ddots&\mathbf{0}&\mathbf{0}&\ddots&\ddots&\mathbf{0}\\
\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{I}_{|\mathcal{S}|}&-\mathbf{I}_{|\mathcal{S}|}&\mathbf{0}&\mathbf{0}&\mathbb{1}_{|\mathcal{S}|}&-\mathbb{1}_{|\mathcal{S}|}\\
\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{I}_{|\mathcal{S}|}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbb{1}_{|\mathcal{S}|}
\end{array}\right).
$$

Thus, $\mathbf{D}_{\mathcal{H}}^\top\mathbf{D}_{\mathcal{H}}$ has the form

$$
D_{\mathcal{H}}^\top D_{\mathcal{H}} = \left(\begin{array}{c:c:c}
\mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hdashline
\mathbf{0}^\top & \mathbf{M}_1 & \mathbf{M}_2 \\
\mathbf{0}^\top & \mathbf{M}_2^\top & \mathbf{M}_3
\end{array}\right), \quad \text{where}
$$

$$
\mathbf{M}_1 = \begin{pmatrix}
2\mathbf{I}_{|\mathcal{S}|} & -\mathbf{I}_{|\mathcal{S}|} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
-\mathbf{I}_{|\mathcal{S}|} & 2\mathbf{I}_{|\mathcal{S}|} & -\mathbf{I}_{|\mathcal{S}|} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \ddots & \ddots & \ddots & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & -\mathbf{I}_{|\mathcal{S}|} & 2\mathbf{I}_{|\mathcal{S}|} & -\mathbf{I}_{|\mathcal{S}|} \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{I}_{|\mathcal{S}|} & 2\mathbf{I}_{|\mathcal{S}|}
\end{pmatrix}, \tag{A.13}
$$

$$
\mathbf{M}_2 = \begin{pmatrix}
2\mathbb{1}_{|\mathcal{S}|} & -\mathbb{1}_{|\mathcal{S}|} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
-\mathbb{1}_{|\mathcal{S}|} & 2\mathbb{1}_{|\mathcal{S}|} & -\mathbb{1}_{|\mathcal{S}|} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \ddots & \ddots & \ddots & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & -\mathbb{1}_{|\mathcal{S}|} & 2\mathbb{1}_{|\mathcal{S}|} & -\mathbb{1}_{|\mathcal{S}|} \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbb{1}_{|\mathcal{S}|} & 2\mathbb{1}_{|\mathcal{S}|}
\end{pmatrix},
$$

$$\mathbf{M}_3 = \begin{pmatrix} 2(|\mathcal{S}|+1) & -(|\mathcal{S}|+1) & 0 & 0 & 0 \\ -(|\mathcal{S}|+1) & 2(|\mathcal{S}|+1) & -(|\mathcal{S}|+1) & 0 & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & -(|\mathcal{S}|+1) & 2(|\mathcal{S}|+1) & -(|\mathcal{S}|+1) \\ 0 & 0 & 0 & -(|\mathcal{S}|+1) & 2(|\mathcal{S}|+1) \end{pmatrix}.$$

From these, we can illustrate the structure of $\mathbf{P}$. One notable thing is that both $\mathbf{D}_\Theta$ and $\mathbf{D}_\mathcal{H}$ have their columns corresponding to $(\theta_{\text{Tu}}^{\text{dow}}, \ldots, \theta_{\text{Su}}^{\text{dow}})$ set exactly to be zero. Thus, in computing $\mathbf{P}^{-1}$, we remove their corresponding columns so that the inverse matrix is applied only to $(\Theta, \Theta_{D^\circ}, \Theta_H)$. The structure of $\bar{\mathbf{P}}$, a sub-matrix of $\mathbf{P} = \mathbf{I} + \mathbf{D}_\Theta^\top \mathbf{D}_\Theta + \mathbf{D}_\mathcal{H}^\top \mathbf{D}_\mathcal{H}$ without the columns corresponding to the daily parameters, is as follows:

$\bar{\mathbf{P}} =$

$$\left(\begin{array}{ccc:c:ccccc:c}
\widetilde{\mathbf{L}}_{\text{net}}+\mathbf{I} & \mathbf{0} & \mathbf{0} & -\widetilde{\mathbf{L}}_{\text{net}} & & & \mathbf{0} & & & \mathbf{0} \\
\mathbf{0} & \ddots & \mathbf{0} & \vdots & & & \vdots & & & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \widetilde{\mathbf{L}}_{\text{net}}+\mathbf{I} & -\widetilde{\mathbf{L}}_{\text{net}} & & & \mathbf{0} & & & \mathbf{0} \\ \hdashline
-\widetilde{\mathbf{L}}_{\text{net}} & \cdots & -\widetilde{\mathbf{L}}_{\text{net}} & (|\mathcal{D}|+|\mathcal{H}|)\widetilde{\mathbf{L}}_{\text{net}}+\mathbf{I} & -\widetilde{\mathbf{L}}_{\text{net}} & & \cdots & & -\widetilde{\mathbf{L}}_{\text{net}} & \mathbf{0} \\ \hdashline
\mathbf{0} & \cdots & \mathbf{0} & -\widetilde{\mathbf{L}}_{\text{net}} & \widetilde{\mathbf{L}}_{\text{net}}+3\mathbf{I} & -\mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \\
 & & & & -\mathbf{I} & \widetilde{\mathbf{L}}_{\text{net}}+3\mathbf{I} & -\mathbf{I} & \mathbf{0} & & \\
\vdots & \ddots & \vdots & \vdots & \mathbf{0} & \ddots & \ddots & \ddots & \mathbf{0} & \mathbf{M}_2 \\
 & & & & \mathbf{0} & \mathbf{0} & -\mathbf{I} & \widetilde{\mathbf{L}}_{\text{net}}+3\mathbf{I} & -\mathbf{I} & \\
\mathbf{0} & \cdots & \mathbf{0} & -\widetilde{\mathbf{L}}_{\text{net}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{I} & \widetilde{\mathbf{L}}_{\text{net}}+3\mathbf{I} & \\ \hdashline
\mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & & & \mathbf{M}_2^\top & & & \mathbf{M}_3+\mathbf{I}_{|\mathcal{H}^\circ|}
\end{array}\right)$$

where $\mathbf{M}_2$ and $\mathbf{M}_3$ are as in (A.13). Here, the size of each block is $(|\mathcal{S}| \cdot |\mathcal{D}^\circ|)$, $|\mathcal{S}|$, $(|\mathcal{S}| \cdot |\mathcal{H}^\circ|)$, and $|\mathcal{H}^\circ|$ from left to right and top to bottom, which match that of $\mathbf{D}_\Theta^\top \mathbf{D}_\Theta$ and $\mathbf{D}_\mathcal{H}^\top \mathbf{D}_\mathcal{H}$ specified above after removing corresponding columns of $(\theta_{\text{Tu}}^{\text{dow}}, \ldots, \theta_{\text{Su}}^{\text{dow}})$.
Let $\widetilde{\mathbf{L}}_{\text{net}} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^\top$ denote the eigenvalue decomposition of $\widetilde{\mathbf{L}}_{\text{net}}$ with $\mathbf{\Lambda}$ denoting a diagonal matrix having the eigenvalues as its diagonal entries. Then,

$\mathbf{W}^\top \bar{\mathbf{P}} \mathbf{W} =$

$$\left(\begin{array}{ccc:c:ccccc:c}
\mathbf{\Lambda}+\mathbf{I} & \mathbf{0} & \mathbf{0} & -\mathbf{\Lambda} & & & \mathbf{0} & & & \mathbf{0} \\
\mathbf{0} & \ddots & \mathbf{0} & \vdots & & & \vdots & & & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathbf{\Lambda}+\mathbf{I} & -\mathbf{\Lambda} & & & \mathbf{0} & & & \mathbf{0} \\ \hdashline
-\mathbf{\Lambda} & \cdots & -\mathbf{\Lambda} & (|\mathcal{D}|+|\mathcal{H}|)\mathbf{\Lambda}+\mathbf{I} & -\mathbf{\Lambda} & & \cdots & & -\mathbf{\Lambda} & \mathbf{0} \\ \hdashline
\mathbf{0} & \cdots & \mathbf{0} & -\mathbf{\Lambda} & \mathbf{\Lambda}+3\mathbf{I} & -\mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \\
 & & & & -\mathbf{I} & \mathbf{\Lambda}+3\mathbf{I} & -\mathbf{I} & \mathbf{0} & \mathbf{0} & \\
\vdots & \ddots & \vdots & \vdots & \mathbf{0} & \ddots & \ddots & \ddots & \mathbf{0} & \widetilde{\mathbf{M}}_2 \\
 & & & & \mathbf{0} & \mathbf{0} & -\mathbf{I} & \mathbf{\Lambda}+3\mathbf{I} & -\mathbf{I} & \\
\mathbf{0} & \cdots & \mathbf{0} & -\mathbf{\Lambda} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{I} & \mathbf{\Lambda}+3\mathbf{I} & \\ \hdashline
\mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & & & \widetilde{\mathbf{M}}_2^\top & & & \mathbf{M}_3+\mathbf{I}_{|\mathcal{H}^\circ|}
\end{array}\right) \tag{A.14}$$

where $\widetilde{\mathbf{M}}_2$ is

$$\widetilde{\mathbf{M}}_2 = \begin{pmatrix} 2\mathbf{E}^{(\mathrm{r})} & -\mathbf{E}^{(\mathrm{r})} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathbf{E}^{(\mathrm{r})} & 2\mathbf{E}^{(\mathrm{r})} & -\mathbf{E}^{(\mathrm{r})} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{E}^{(\mathrm{r})} & 2\mathbf{E}^{(\mathrm{r})} & -\mathbf{E}^{(\mathrm{r})} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{E}^{(\mathrm{r})} & 2\mathbf{E}^{(\mathrm{r})} \end{pmatrix} \text{ with } \mathbf{E}^{(\mathrm{r})} = \mathbf{E}\mathbb{1}_{|\mathcal{S}|}, \text{ and } \mathbf{W} = \left(\begin{array}{ccc:c} \mathbf{E} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{E} & \mathbf{0} \\ \hdashline & \mathbf{0} & & \mathbf{I}_{|\mathcal{H}^\circ|} \end{array}\right).$$

Thus, the solution of $\bar{\mathbf{P}}\mathbf{z} = \mathbf{b}$ can be achieved by solving

$$\left(\mathbf{W}^\top \bar{\mathbf{P}} \mathbf{W}\right) \widetilde{\mathbf{z}} = \widetilde{\mathbf{b}} \tag{A.15}$$

with $\widetilde{\mathbf{z}} = \mathbf{W}^\top \mathbf{z}$ and $\widetilde{\mathbf{b}} = \mathbf{W}^\top \mathbf{b}$, and then finally setting $\mathbf{z} = \mathbf{W}\widetilde{\mathbf{z}}$. Solving (A.15) can be achieved efficiently by taking into account the structure of $\mathbf{W}$ and $\mathbf{W}^\top \bar{\mathbf{P}} \mathbf{W}$. Specifically, $\mathbf{W}$ is block diagonal with repeated blocks which reduces the matrix multiplication complexity. The matrix $\mathbf{W}^\top \bar{\mathbf{P}} \mathbf{W}$ is very sparse as described in (A.14) with all blocks involving $\mathbf{\Lambda}$ being diagonal, which facilitates efficient computation. Therefore, (A.15) can be solved in an iterative manner using LU decomposition. The details of this procedure is illustrated later in this section. Having the full illustration of the structures, our suggested method for the projection step is described in Algorithm 1.

Algorithm 1 Projection in Step 2 of ADMM

1: **Inputs:**
$\left(\Theta^{k+1},\ \Theta_D^{k+1},\ \Theta_H^{k+1},\ \Gamma^{k+1},\ \Psi^{k+1}\right),\ \left(U_\Theta^k,\ , U_{\Theta_D}^k,\ U_{\Theta_H}^k,\ T_\Gamma^k,\ , T_\Psi^k\right),$
$\mathbf{D}_\Theta,\ \mathbf{D}_\mathcal{H},\ \mathbf{W}$
2: Compute $\mathbf{b}_1 \leftarrow \left(\Theta^{k+1},\ \Theta_D^{k+1},\ \Theta_H^{k+1}\right)^\top + \left(U_\Theta^k,\ , U_{\Theta_D}^k,\ U_{\Theta_H}^k\right)^\top$
3: Compute $\mathbf{b}_2 \leftarrow \mathbf{D}_\Theta \left(\Gamma^{k+1} + T_\Gamma^k\right)^\top + \mathbf{D}_\mathcal{H} \left(\Psi^{k+1} + T_\Psi^k\right)^\top$
4: Compute $\widetilde{\mathbf{b}} \leftarrow \mathbf{W}^\top (\mathbf{b}_1 + \mathbf{b}_2)$
5: Solve $\left(\mathbf{W}^\top \left(\mathbf{I} + \mathbf{D}_\Theta^\top \mathbf{D}_\Theta + \mathbf{D}_\mathcal{H}^\top \mathbf{D}_\mathcal{H}\right) \mathbf{W}\right)^{-1} \widetilde{\mathbf{z}} = \widetilde{\mathbf{b}}$ for $\widetilde{\mathbf{z}}$
6: Compute $\mathbf{z} \leftarrow \mathbf{W}\widetilde{\mathbf{z}}$
7: Set $(Z_\Theta^{k+1}, Z_{\Theta_D}^{k+1}, Z_{\Theta_H}^{k+1}) \leftarrow (\mathbf{z}_\Theta, \mathbf{z}_{\Theta_D}, \mathbf{z}_{\Theta_H})$ where $\mathbf{z} = (\mathbf{z}_\Theta, \mathbf{z}_{\Theta_D}, \mathbf{z}_{\Theta_H})$
8: Set $(S_\Gamma^{k+1}, S_\Psi^{k+1}) \leftarrow (\mathbf{D}_\Theta \mathbf{z},\ \mathbf{D}_\mathcal{H} \mathbf{z}_{\Theta_H})$
9: **Ouputs:**
$(Z_\Theta^{k+1}, Z_{\Theta_D}^{k+1}, Z_{\Theta_H}^{k+1}, S_\Gamma^{k+1}, S_\Psi^{k+1})$

In order to solve (A.15) in line 5 of Algorithm 1, we can utilize the LU decomposition of the matrix in (A.14). Writing the LU decomposition of the matrix (A.14) by $\mathbf{L}\mathbf{L}^\top$, the matrix $\mathbf{L}$

has the banded block diagonal structure as follows:

$$\mathbf{L} = \left(\begin{array}{c|c|c|c} \mathbf{A}_{11} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{A}_{32} & \mathbf{A}_{33} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{A}_{43} & \mathbf{A}_{44} \end{array}\right). \tag{A.16}$$

The blocks match the size of their counterparts in (A.14). Here, the relatively large matrices $\mathbf{A}_{11}$, $\mathbf{A}_{21}$, $\mathbf{A}_{22}$, $\mathbf{A}_{32}$, and $\mathbf{A}_{33}$ have specific repetitive sparse structures which facilitate memory saving and efficient computation. Specifically, these contain repeated sub-matrices of dimension $|\mathcal{S}| \times |\mathcal{S}|$ which are of the following forms:

$$\mathbf{A}_{11} = \begin{pmatrix} \mathbb{D}_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbb{D}_{11} \end{pmatrix}, \quad \mathbf{A}_{21} = \begin{pmatrix} \mathbb{D}_{21} \dots \mathbb{D}_{21} \end{pmatrix}, \quad \mathbf{A}_{22} = \mathbb{D}_{22}, \quad \mathbf{A}_{32} = \begin{pmatrix} \mathbb{D}_{32} \cdots \mathbb{D}_{32} \end{pmatrix}^\top,$$

$$\mathbf{A}_{33} = \begin{pmatrix} d_{1,1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ g_{2,1} & d_{2,2} & \mathbf{0} & \vdots & \vdots & \vdots \\ f_1 & g_{3,2} & d_{3,3} & \mathbf{0} & & \\ \vdots & f_2 & \ddots & \ddots & \mathbf{0} & \\ \vdots & \vdots & & \ddots & \ddots & \mathbf{0} \\ f_1 & f_2 & \cdots & f_{|\mathcal{H}^\circ|-1} & g_{|\mathcal{H}^\circ|,|\mathcal{H}^\circ|-1} & d_{|\mathcal{H}^\circ|,|\mathcal{H}^\circ|} \end{pmatrix}.$$

Using this banded diagonal structure of the LU-decomposition, solving (A.15) can be done in two steps, (i) solving $\mathbf{L}\mathbf{y} = \widetilde{\mathbf{b}}$, and then (ii) solving $\mathbf{L}^\top \widetilde{\mathbf{z}} = \mathbf{y}$. The details are presented below.

(i) Solve $\mathbf{L}\mathbf{y} = \widetilde{\mathbf{b}}$ with $\mathbf{y} = (\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(4)})^\top$:

(a) Solve $\mathbf{A}_{11}\mathbf{y}^{(1)} = \widetilde{\mathbf{b}}^{(1)}$:
$\mathbf{y}_i^{(1)} = \mathbb{D}_{11}^{-1}\widetilde{\mathbf{b}}_i^{(1)}$ for $i = 1, \dots, |\mathcal{D}^\circ|$.

(b) Solve $\mathbf{A}_{22}\mathbf{y}^{(2)} = \widetilde{\mathbf{b}}^{(2)} - \mathbf{A}_{21}\mathbf{y}^{(1)}$:
$\mathbf{y}^{(2)} = \mathbb{D}_{22}^{-1}\left(\widetilde{\mathbf{b}}^{(2)} - \mathbb{D}_{21}\sum_{i=1}^{|\mathcal{D}^\circ|}\mathbf{y}_i^{(1)}\right)$.

(c) Solve $\mathbf{A}_{33}\mathbf{y}^{(3)} = \widetilde{\mathbf{b}}^{(3)} - \mathbf{A}_{32}\mathbf{y}^{(2)}$:

$$\begin{aligned} \mathbf{y}_1^{(3)} &= d_{1,1}^{-1}\left(\widetilde{\mathbf{b}}_1^{(3)} - \mathbb{D}_{32}\mathbf{y}_1^{(3)}\right), \\ \mathbf{y}_2^{(3)} &= d_{2,2}^{-1}\left(\left(\widetilde{\mathbf{b}}_2^{(3)} - \mathbb{D}_{32}\mathbf{y}_2^{(3)}\right) - g_{2,1}\mathbf{y}_1^{(3)}\right), \\ \mathbf{y}_i^{(3)} &= d_{i,i}^{-1}\left(\left(\widetilde{\mathbf{b}}_i^{(3)} - \mathbb{D}_{32}\mathbf{y}_i^{(3)}\right) - g_{i,i-1}\mathbf{y}_{i-1}^{(3)} - \sum_{j=1}^{i-2} f_j \mathbf{y}_j^{(3)}\right) \text{ for } i = 3, \dots, |\mathcal{H}^\circ|. \end{aligned}$$

(d) $\mathbf{y}^{(4)} = \mathbf{A}_{44}^{-1}\left(\widetilde{\mathbf{b}}^{(4)} - \mathbf{A}_{43}\mathbf{y}^{(3)}\right)$.

(ii) Solve $\mathbf{L}^\top \widetilde{\mathbf{z}} = \mathbf{y}$ with $\widetilde{\mathbf{z}} = (\widetilde{\mathbf{z}}^{(1)}, \ldots, \widetilde{\mathbf{z}}^{(4)})^\top$:

(a) $\widetilde{\mathbf{z}}^{(4)} = \left(\mathbf{A}_{44}^\top\right)^{-1}\mathbf{y}^{(4)}$

(b) Solve $\mathbf{A}_{33}^\top \widetilde{\mathbf{z}}^{(3)} = \mathbf{y}^{(3)} - \mathbf{A}_{43}^\top \widetilde{\mathbf{z}}^{(4)}$:

$$\widetilde{\mathbf{z}}^{(3)}_{|\mathcal{H}^\circ|} = d^{-1}_{|\mathcal{H}^\circ|,|\mathcal{H}^\circ|}\left(\mathbf{y}^{(3)}_{|\mathcal{H}^\circ|} - \left(\mathbf{A}_{43}^\top\right)_{|\mathcal{H}^\circ|}\widetilde{\mathbf{z}}^{(4)}\right),$$

$$\widetilde{\mathbf{z}}^{(3)}_{|\mathcal{H}^\circ|-1} = d^{-1}_{|\mathcal{H}^\circ|-1,|\mathcal{H}^\circ|-1}\left(\mathbf{y}^{(3)}_{|\mathcal{H}^\circ|-1} - \left(\mathbf{A}_{43}^\top\right)_{|\mathcal{H}^\circ|-1}\widetilde{\mathbf{z}}^{(4)} - g_{|\mathcal{H}^\circ|,|\mathcal{H}^\circ|-1}\widetilde{\mathbf{z}}^{(3)}_{|\mathcal{H}^\circ|}\right),$$

$$\widetilde{\mathbf{z}}^{(3)}_{i} = d^{-1}_{i,i}\left(\mathbf{y}^{(3)}_{i} - \left(\mathbf{A}_{43}^\top\right)_{i}\widetilde{\mathbf{z}}^{(4)} - g_{i+1,i}\widetilde{\mathbf{z}}^{(3)}_{i+1} - (|\mathcal{H}^\circ| - i - 1)f_i\sum_{j=i+2}^{|\mathcal{H}^\circ|}\widetilde{\mathbf{z}}^{(4)}_{j}\right)$$

for $i = |\mathcal{H}^\circ| - 2, \ldots, 1$.

(c) Solve $\mathbf{A}_{22}^\top \widetilde{\mathbf{z}}^{(2)} = \mathbf{y}^{(2)} - \mathbf{A}_{32}^\top \widetilde{\mathbf{z}}^{(3)}$:
$\widetilde{\mathbf{z}}^{(2)} = \mathbb{D}_{22}^{-1}\left(\mathbf{y}^{(2)} - \mathbb{D}_{32}\left(\sum_{i=1}^{|\mathcal{H}^\circ|}\widetilde{\mathbf{z}}^{(3)}_i\right)\right)$

(d) Solve $\mathbf{A}_{11}^\top \widetilde{\mathbf{z}}^{(1)} = \mathbf{y}^{(1)} - \mathbf{A}_{21}^\top \widetilde{\mathbf{z}}^{(2)}$:
$\widetilde{\mathbf{z}}^{(1)}_i = \mathbb{D}_{11}^{-1}\left(\mathbf{y}^{(1)}_i - \mathbb{D}_{21}\widetilde{\mathbf{z}}^{(2)}\right)$ for $i = 1, \ldots, |\mathcal{D}^\circ|$.

Here, $(\mathbf{A}_{43})_i$ denotes its ith row. The vector arguments are partitioned as

$$\mathbf{x}^\top = \left((\mathbf{x}^{(1)}_1)^\top, \ldots, (\mathbf{x}^{(1)}_{|\mathcal{D}^\circ|})^\top,\ (\mathbf{x}^{(2)})^\top,\ (\mathbf{x}^{(3)}_1)^\top, \ldots, (\mathbf{x}^{(3)}_{|\mathcal{H}^\circ|})^\top,\ (\mathbf{x}^{(4)})^\top\right)^\top,$$

where $\mathbf{x}_i^{(j)}$ for $j = 1, 3$ and $\mathbf{x}^{(2)}$ are of length $|\mathcal{S}|$ and $\mathbf{x}^{(4)}$ is of length $|\mathcal{H}^\circ|$.

A.2 Computation of model complexity

The numerator of MC in (10) counts the number of connected components of the given graphs. Typically, the number of connected components of a graph can be found by counting the number of zero-eigenvalue of the graph's Laplacian matrix. The Laplacian matrix of $\mathfrak{N}_{\text{single}}(r) \cap \widehat{\mathfrak{N}}_{D,d}$ has the dimension of $|\mathcal{S}| \times |\mathcal{S}|$ for each d, and it is feasible to compute its eigenvalues and calculate $\mathcal{C}(\mathfrak{N}_{\text{single}}(r) \cap \widehat{\mathfrak{N}}_{D,d})$.
However, computing $\mathcal{C}(\mathfrak{N}_{\text{multi}}(r) \cap \widehat{\mathfrak{N}}_H)$ is not so straightforward since the Laplacian matrix of the graph $\mathfrak{N}_{\text{multi}}(r) \cap \widehat{\mathfrak{N}}_H$ has the dimension of $|\mathcal{S}||\mathcal{H}| \times |\mathcal{S}||\mathcal{H}|$, it is not practical to compute its eigendecomposition. Instead, we propose to obtain $\mathcal{C}(\mathfrak{N}_{\text{multi}}(r) \cap \widehat{\mathfrak{N}}_H)$ by first finding the layer-specific connected components, and then coalescing the components of two consecutive layers if they are connected transversely over the layers.
Before describing the proposed method, we introduce some notations relevant to $\mathfrak{N}_{\text{single}}(r) \cap \widehat{\mathfrak{N}}_{H,h}$, a single-layer network for each $h \in \mathcal{H}$:

$$
\begin{aligned}
\widehat{\mathfrak{N}}_{H,h} &= (\mathcal{S},\, \mathcal{E}_{H,h}) \quad \text{with} \quad \mathcal{E}_{H,h} = \left\{ (s, s'),\, s \neq s' : \widehat{\phi}^{\text{hod}}_{s,h} = \widehat{\phi}^{\text{hod}}_{s',h} \right\}, \\
\mathfrak{G}_h &= \mathfrak{N}_{\text{single}}(r) \cap \widehat{\mathfrak{N}}_{H,h}, \\
N_h &= \mathcal{C}(\mathfrak{G}_h), \\
\mathbf{M}^{(0)} &= |\mathcal{H}| \times |\mathcal{S}| \text{ matrix that encodes the layer-specific connected components of } \mathfrak{G}_h, \\
\mathbf{L}_h &= \text{the Laplacian matrix of } \mathfrak{G}_h \\
\mathbf{c}_{h,k}(\mathbf{M}) &= \text{the index vector of entries that corresponds to the elements in the } k\text{th cluster} \\
&\qquad \text{of the } h\text{th row of a cluster label matrix } \mathbf{M}.
\end{aligned}
$$

Specifically, for each connected component, the corresponding (h, s) elements of $\mathbf{M}^{(0)}$ take a unique value, and thus the number of unique values in $\mathbf{M}^{(0)}$ agrees with the number of total connected components $\sum_{h=0}^{23} N_h$.
We begin by describing how to construct $\mathbf{M}^{(0)}$. Recall that the number of zero eigenvalues of $\mathbf{L}_h$ corresponds to N_h. The eigenvector associated with the zero eigenvalue provides some information of the connected components. Namely, it is a linear combination of the connected component indicator vectors $\mathbf{c}_{h,k}(\mathbf{M}^{(0)})$. For an arbitrary index vector $\mathbf{g} = (g_i : g_i \in \{1, \ldots, |\mathcal{S}|\})$, we denote the indicator vector of $\mathbf{g}$ by $\mathbf{e}_{\mathbf{g}}$: It has its g_ith entry to be one for $i = 1, \ldots, |\mathbf{g}|$, and all the rest are zeros. Denote by $\mathbf{v}^{(0)}_{h,i}$ the ith eigenvector associated with the zero eigenvalues of $\mathbf{L}_h$ for $i = 1, \ldots, N_h$. Then, we have $\mathbf{v}^{(0)}_{h,i} = \sum_{k=1}^{N_h} a_k^{(h,i)} \mathbf{e}_{\mathbf{c}_{h,k}(\mathbf{M}^{(0)})}$ for some constants $a_k^{(h,i)}$s. Thus, we utilize the eigenvectors to identify the cluster label of each stations and construct $\mathbf{M}^{(0)}$. Precisely, we find the partition $\mathcal{P}$ of the index set $\{1, \ldots, |\mathcal{S}|\}$ with the smallest cardinality such that for each P belonging to the partition, all the elements of $\mathbf{v}^{(0)}_{h,i}$ located at P take the same value, for all $i = 1, \ldots, N_h$. This procedure is described in Algorithm 2.

Algorithm 2 Construction of $\mathbf{M}^{(0)}$.

Inputs:
  The Laplacian matrix $\mathbf{L}_h$ for $h \in \mathcal{H}$
Initialize:
  $\mathbf{M}^{(0)} \leftarrow$ a $|\mathcal{H}| \times |\mathcal{S}|$-matrix of zeros
  $\texttt{cluster\_label} \leftarrow 1$
for $h \in \mathcal{H}$ **do**
  Perform eigenvalue decomposition of $\mathbf{L}_h$
  $N_h \leftarrow$ the number of zero eigenvalues
  $\mathbf{V} \leftarrow$ a $|\mathcal{S}| \times N_h$ matrix having the eigenvector $\mathbf{v}_{h,i}^{(0)}$ as its ith column
  $\texttt{unlabeled} \leftarrow \{1, 2, \ldots, |\mathcal{S}|\}$
  for $s \in \{1, \ldots, |\mathcal{S}|\}$ **do**
    if $s \notin \texttt{unlabeled}$ **then**
      $\texttt{cluster} \leftarrow \{1, \ldots, |\mathcal{S}|\}$
      for $i \in \{1, \ldots, N_h\}$ **do**
        $\texttt{value} \leftarrow V_{s,i}$, the sth entry of the ith eigenvector $\mathbf{v}_{h,i}^{(0)}$
        $\texttt{value\_set} \leftarrow \{s' : V_{s',i} = \texttt{value}\}$
        $\texttt{cluster} \leftarrow \texttt{cluster} \cap \texttt{value\_set}$
      end for
      $M_{h,s'}^{(0)} \leftarrow \texttt{cluster\_label}$ for $s' \in \texttt{cluster}$
      $\texttt{unlabeled} \leftarrow \texttt{unlabeled} \setminus \texttt{cluster}$
      $\texttt{cluster\_label} \leftarrow \texttt{cluster\_label} + 1$
    end if
  end for
end for
Ouputs:
  $\mathbf{M}^{(0)}$

Once the layer-specific cluster label matrix $\mathbf{M}^{(0)}$ is provided, we link the connected components throughout the hourly layers as described in Algorithm 3. Within the procedure, the sub-routine given in Algorithm 4 is utilized, which sequentially links connected components lying in two consecutive layers. In each run, Algorithm 4 links a component in evaluation to exactly one component in another layer while there can be more than one component that are supposed to be linked. Thus, Algorithm 3 keeps running the sub-algorithm until there exists no more component left to be linked. One notable feature is that the sub-algorithm is executed twice with different order vector $\mathbf{r}$ each time. This is to account for the circular feature of hour-of-a-day. The function $\texttt{is}(\cdot)$ in the sub-algorithm 4 is defined to return the Boolean of the input statement.

Algorithm 3 Counting the connected components of $\mathfrak{N}_{\text{multi}}(r) \times \widehat{\mathfrak{N}}_H$.

Inputs:
  The layer-specific cluster label matrix $\mathbf{M}^{(0)}$
  The estimates $\widehat{\phi}_{s,h}^{\text{hod}}$ for $s \in \mathcal{S}$ and $h \in \mathcal{H}$
$\mathbf{M} \leftarrow \mathbf{M}^{(0)}$
$\texttt{changed} \leftarrow \texttt{true}$
while `changed` **do**
  $\texttt{changed} \leftarrow \texttt{false}$
  Make links across the layers in a forward manner by running Algorithm 4
  with $\mathbf{r} = (0, 1, \ldots, 23)$
  Connect the layers of $h = 0$ and $h = 23$ in a forward manner by running Algorithm 4
  with $\mathbf{r} = (23, 0, 1, \ldots, 22)$
end while
Ouputs:
  the number of unique values in $\mathbf{M}$

Algorithm 4 Link across ordered layers.

Inputs:
The estimates $\widehat{\phi}^{\text{hod}}_{s,h}$ for $s \in \mathcal{S}$ and $h \in \mathcal{H}$
An order vector $\mathbf{r}$ of length $|\mathcal{H}|$
An $|\mathbf{r}| \times |\mathcal{S}|$ cluster label matrix $\mathbf{M}$
A Boolean variable `changed`
$\mathbf{M}^{\text{old}} \leftarrow \mathbf{M}$
for $i \in \{1, \ldots, |\mathbf{r}|\}$ **do**
  // link the r_ith and the r_{i+1}th layers:
  $\mathbf{M}^{\text{new}} \leftarrow \mathbf{M}^{\text{old}}$
  for $k \in \{1, \ldots, N_{r_i}\}$ **do**
    // investigate whether each component is subject to further connection
    ind $\leftarrow \mathbf{c}_{r_i,k}(\mathbf{M}^{\text{old}})$
    $j \leftarrow 0$
    `matched` $\leftarrow$ `false`
    `exhausted` $\leftarrow$ `is`$(k \geq |\mathbf{c}_{r_i,k}(\mathbf{M}^{\text{old}})|)$
    while `matched` $=$ `false` and `exhausted` $=$ `false` **do**
      $j \leftarrow j + 1$
      $c_1 \leftarrow \widehat{\phi}^{\text{hod}}_{\text{ind}_j, r_i}$
      $c_2 \leftarrow \widehat{\phi}^{\text{hod}}_{\text{ind}_j, r_{i+1}}$
      ind$^{(j,r_{i+1})}$ $\leftarrow$ the index vector of the cluster to which the jth entry of `ind` belongs, in the r_{i+1}th row of $\mathbf{M}^{\text{new}}$ i.e. $\mathbf{c}_{r_{i+1},k'}(\mathbf{M}^{\text{new}})$ for some k' such that $\texttt{ind}_j \in c_{r_{i+1},k'}(\mathbf{M}^{\text{new}})$
      $m_1 \leftarrow \min\ \mathbf{M}^{\text{old}}{}_{r_i,\text{ind}}$
      $m_2 \leftarrow \min\ \mathbf{M}^{\text{new}}_{r_{i+1},\text{ind}^{(j,r_{i+1})}}$
      `same_estimate` $\leftarrow$ `is`$(c_1 = c_2)$
      `same_label` $\leftarrow$ `is`$(m_1 = m_2)$
      `should_connect` $\leftarrow$ `is`(`same_estimate` and not `same_label`)
      if `should_connect` **then**
        // connect the to components and update their labels
        $m = \min\{m_1, m_2\}$
        $\mathbf{M}^{\text{new}}_{r_i,\text{ind}} \leftarrow m$
        $\mathbf{M}^{\text{new}}_{r_{i+1},\text{ind}^{(j,r_{i+1})}} \leftarrow m$
        `changed` $\leftarrow$ `true`
        `matched` $\leftarrow$ `true`
        `exhausted` $\leftarrow$ `is`$(k \geq |\mathbf{c}_{r_i,k}(\mathbf{M}^{\text{old}})|)$
      end if
    end while
  end for
  $\mathbf{M}^{\text{old}} \leftarrow \mathbf{M}^{\text{new}}$
end for
$\mathbf{M} \leftarrow \mathbf{M}^{\text{new}}$
Ouputs:
The updated cluster label matrix $\mathbf{M}$
The updated Boolean variable `changed`

B Additional data analysis results

Table B.1 presents the results obtained with $r = 750$, see Table 4 in the main text for the results obtained with $r = 1500$.

Table B.1: Estimated coefficients for the covariate effects by the proposed fused Lasso regression method from each fold used in the 7-fold CV and from the full data when $r = 750$. For comparison, we also report the estimates obtained with fusion Lasso-only, full-interaction and no-interaction methods.

	Fold							Full data				
	1	2	3	4	5	6	7	Fused	Fusion-only	Lasso-only	Full	No
α	0.057	0.052	0.051	0.050	0.050	0.054	0.050	0.052	0.052	0.052	0.052	0.052
β^{rain}	-2.301	-2.356	-2.380	-2.377	-2.238	-2.182	-2.383	-2.324	-2.325	-2.321	-2.323	-2.310
β_1^{air}	0.098	0.141	0.100	0.177	0.124	0.169	0.168	0.140	0.141	0.138	0.138	0.137
β_2^{air}	0.116	0.173	0.098	0.177	0.133	0.159	0.187	0.149	0.149	0.145	0.144	0.146
β_3^{air}	0.218	0.285	0.250	0.175	0.259	0.321	0.310	0.275	0.277	0.271	0.276	0.243

Figures B.1 and B.2 plot the results obtained with $r = 750$, see Figures 7 and 8 in the main text for the results obtained with $r = 1500$.

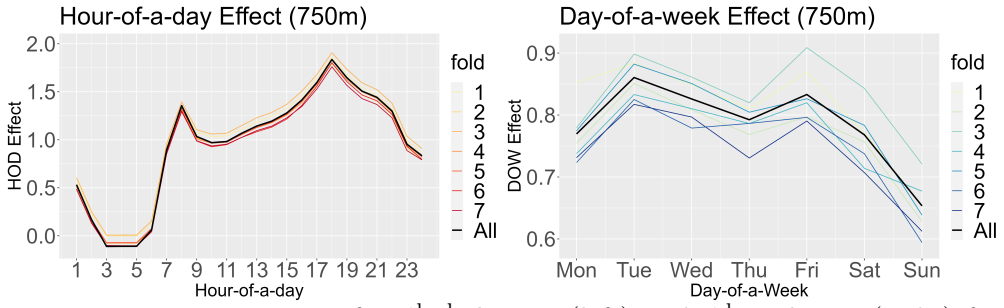

Figure B.1: Parameter estimates for ϕ_h^{hod}, $h \in \mathcal{H}$ (left) and ϕ_d^{dow}, $d \in \mathcal{D}$ (right) from each fold used in the 7-fold CV and from the full data when $r = 750$.

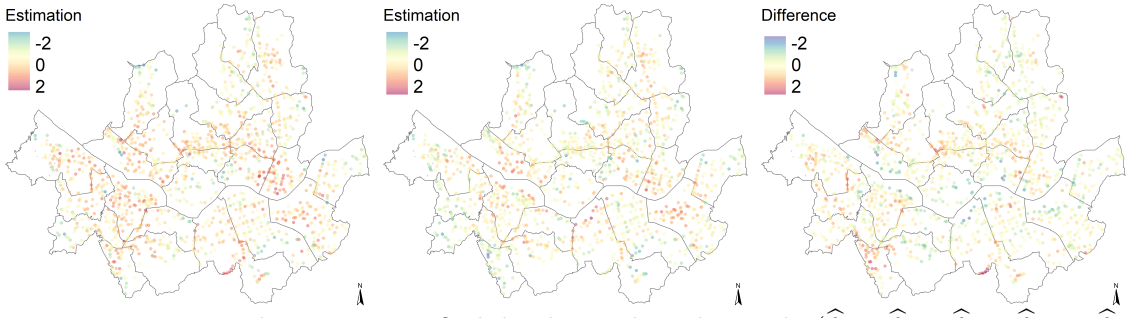

Figure B.2: Estimated station-specific bike demands in log-scale ($\widehat{\theta}_s + \widehat{\theta}_d + \widehat{\theta}_h + \widehat{\theta}_{s,d} + \widehat{\theta}_{s,h}$) from the model fitted with $r = 750$ at 8am on Tuesdays (left), at 8pm on Sundays (middle) and their differences (right).