Received: 18 December 2020 | Accepted: 12 February 2022

DOI: 10.1111/rssa.12834

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Series A
Statistics
in Society

ORIGINAL ARTICLE

The Helsinki bike-sharing system—Insights gained from a spatiotemporal functional model

Andreas Piter[1] | Philipp Otto[1,2] | Hamza Alkhatib[1]

[1]Leibniz University Hannover, Hannover, Germany

[2]University of Göttingen, Göttingen, Germany

Correspondence
Philipp Otto, University of Göttingen, Göttingen, Niedersachsen, Germany.
Email: philipp.otto@uni-goettingen.de

Abstract

Understanding the usage patterns for bike-sharing systems is essential in terms of supporting and enhancing operational planning for such schemes. Studies have demonstrated how factors such as weather conditions influence the number of bikes that should be available at bike-sharing stations at certain times during the day. However, the influences of these factors usually vary over the course of a day, and if there is good temporal resolution, there could also be significant effects only for some hours/minutes (rush hours, the hours when shops are open and so forth). Thus, in this paper, an analysis of Helsinki's bike-sharing data from 2017 is conducted that considers full temporal and spatial resolutions. The station hire data are analysed in a spatiotemporal functional setting, where the number of bikes at a station is defined as a continuous function of the time of day. For this completely novel approach, we apply a functional spatiotemporal hierarchical model to investigate the effect of environmental factors and the magnitude of the spatial and temporal dependence. Challenges in computational complexity are faced using a Monte Carlo subsampling approach. The results show the necessity of splitting the bike-sharing stations into two clusters based on the similarity of their spatiotemporal functional observations in order to model the station hire data of Helsinki's bike-sharing system effectively.

KEYWORDS
bike-sharing system, functional data analysis, spatiotemporal statistics, subsampling

1 | INTRODUCTION

Bicycle-sharing systems have become popular in all the cities in which they have been implemented. Among these cities is the capital of Finland, Helsinki, where a station-based system has provided a flexible transport option since 2016 (City of Helsinki, a; Helsinki Region Transport, a). That is, a bike can be taken from a bike-sharing station and returned to any other station. The extension of the existing system to neighbouring cities and the rebalancing of the bike-sharing stations are challenging tasks for the operator and city planners (cf. Schuijbroek et al., 2017). Support can be provided via empirical research examining the demand at all stations over time, and determining the factors that influence the usage of a bike-sharing system is one of the main research interests. Furthermore, the environmental impact of implementing bike-sharing schemes is an important question in current research (e.g. Maranzano et al., 2020; Zhang & Mi, 2018).

In studies concerning bike-sharing systems in other cities, a variety of influential factors have been discussed and analysed using different statistical models (El-Assi et al., 2017; Wang et al., 2020; Yang et al., 2016). The significance and magnitudes of such factors have been analysed by Eren and Uz (2020) in a review. Moreover, Yang et al. (2020) focused on the analysis and prediction of the bike-sharing usage at different points in time. In these studies, the amount of data is often reduced through the aggregation of the observations in time spans consisting of a set number of hours (e.g. El-Assi et al., 2017) or a day (e.g. Buck & Buehler, 2012). So far, no study has analysed how factors' influences vary by time of day. Additionally, no study has utilised an entire dataset in its analyses, that is, all the studies to date have aggregated the data or reduced the dimensionality.

This gap in knowledge is addressed in this paper by means of a comprehensive analysis of the freely available station hire data for the bike-sharing system in Helsinki from 2017. The station hire data are meant to represent spatiotemporal functional observations. Thus, we apply a complex spatiotemporal functional hierarchical model implemented in the software package D-STEM (cf. Fassò et al., 2018; Finazzi & Fassò, 2014; Wang et al., 2021). This model can be used to predict and map the spatiotemporal process and its uncertainty over a geographical region across time. Applications of such dynamic coregionalisation models are used in Fassò et al. (2016), Fassò and Finazzi (2013), Finazzi et al. (2013) and Taghavi-Shahri et al. (2019) to assess the air quality in Europe and model the concentrations of several airborne pollutants in a multivariate setting or for land use regression in Teheran, Iran. In contrast to previous approaches, which handle the purely temporal dynamics separately from the purely spatial correlation component, the approach presented in Calculli et al. (2015) combines the spatial and temporal dependencies in an autoregressive spatial component. It is known as hidden dynamic geostatistical model (HDGM). The parameters are estimated using the maximum likelihood approach and an EM algorithm (cf. Finazzi & Fassò, 2014). Alternatively, Bayesian approaches can also be used (cf. Rue et al., 2009). These approaches are mostly based on computationally efficient integrated nested

Laplace approximations (INLA). In this paper, we focus on the EM estimation for functional HDGM implemented in D-STEM, because there is no prior information about the model parameters (in this case, non-informative priors would be appropriate). Moreover, functional data can efficiently be handled. Although this technique was originally developed to handle spatiotemporal functional data from environmental sciences, such as atmospheric radiosonde profiles, its potential for modelling the number of allocated bikes at the bike-sharing system in Helsinki is demonstrated in this paper.

Because of the large amount of data from 140 stations (measured in 5-min intervals), we propose to combine this estimation with a Monte Carlo subsampling approach. That is, we repeatedly draw a smaller subset from all available spatial locations, bike-sharing stations. Using all functional observations of these stations over time, the spatiotemporal model was estimated. Thus, we will be able to estimate the standard errors of the estimated functional model parameters in a very efficient way, allowing a rich interaction model with spatiotemporal interactions to be estimated from the full data. Moreover, all results are validated in a cross-validation study.

The remainder of the paper is structured as follows. First, we provide an overview on different bike-sharing systems and previous empirical findings. Moreover, we briefly introduce the bike-sharing service in Helsinki and present some descriptive statistics for a first exploratory analysis. In the ensuing section, we introduce the spatiotemporal functional model from a theoretical perspective and explain the applied subsampling principle. The concept of functional data and the construction of a continuous function from discrete observations are described. These theoretical sections are followed by the empirical analysis of the data from Helsinki. Initially, we discuss several descriptive statistics and figures in detail to provide a comprehensive understanding of the data, which is highly complex (i.e. spatial, temporal domain; daily, weekly periodicity; high frequency; and so forth). Eventually, the estimated functional parameters are shown and the results are discussed in Section 5. In this section, we also explain how the specific model (hyper-)parameters are chosen. Section 6 concludes the paper.

2 | BIKE-SHARING SYSTEMS

Many of the larger cities across the world have expanded their public transport systems by introducing bike-share schemes, which provide an alternative and sustainable transport mode. Starting in 2000, the number of bike-share systems worldwide has increased rapidly, and many studies have been conducted to improve these services and understand their usage patterns (Fishman, 2016; Gervini & Khanal, 2019).

The systems differ in the way they handle bike usage (Eren & Uz, 2020). On the one hand, there are station-based systems, where users retrieve a bike from a particular station, take a ride and return it to any other station. Here, a positive effect is that the station locations are fixed and thus users know where to search for bikes. However, a bike station may already be full when a user wants to return a bike, as there are only a limited number of docks at each station. Then the bike can only be returned to another station. On the other hand, dockless sharing systems offer more flexibility, as users can return bikes anywhere and they are not bound to stations. However, the disadvantage is that users who want to pick up a bike need to be lucky to find a bike close by. Bike-sharing systems have become popular for various reasons (O'brien et al., 2014). City administrations aim at increasing the number of cyclists and reducing the car traffic in the cities (Fishman, 2016). Shared bikes can be used to overcome distances between public transport options, such as the metro or train, to reach specific destinations, such as the workplace or recreational areas. Hence, bike sharing improves the public transport network and helps users cover gaps in that network or the last miles (Willberg et al., 2019).

Research in this area proceeds in various directions but mainly aims to understand users' behaviour and the different facets of bike-sharing demand. The knowledge gained from the investigations helps improve bike-sharing systems and support operators in operational planning. Understanding usage patterns of, and dependencies between, stations may help when introducing similar systems to other cities (Tran et al., 2015). Martinez et al. (2012) and García-Palomares et al. (2012) address finding appropriate station locations and determining bike fleet size. Data from user registrations or from user surveys provide the users' perspective and give insights into the socio-demographic factors influencing the usage of bike-sharing systems (Willberg et al., 2019).

A different focus is set by data-driven demand analysis. On the one hand, there is station hire data, which give either the number of bikes or the number of check-outs and check-ins at each station at a certain point in time. On the other hand, trip-based data give information about the origin, destination and duration of each bicycle trip. Thus, there are several studies that investigate spatial and temporal factors influencing the demand on a station level or trip basis that try to predict future usage (Li et al., 2015; Rixey, 2013; Yang et al., 2016). However, station-based bike-sharing systems suffer from an unbalanced spatial distribution of bikes at the stations due to different levels of demand across space and time. Hence, this optimization problem must find the most effective rebalancing strategy for the bikes in the network (Shi et al., 2019).

A recently published literature review by Eren and Uz (2020) on the factors influencing bike-sharing demand focuses on six categories. The categories used are weather conditions, built environment, public transport and temporal factors that are used in many studies on station hire data. One of the main findings is that precipitation affects bike-sharing demand the most among the meteorological covariates. Its negative correlation with demand was found in almost all examined studies. Furthermore, increases in the humidity and wind speed decrease the demand, whereas air temperatures between 0 and 30°C lead to more bicycle trips. The strongest positive correlation was found for temperatures between 20 and 30°C, but the demand is less for temperatures under 0°C and over 30°C. Infrastructure and land use are widely investigated factors in the built environment category. Bicycle lanes and the proximity of bike-sharing stations to them are found to have high positive impacts on a station's demand. Furthermore, changes in the elevation across the area of a bike-sharing service are correlated with its demand. From trip-based data, it can be seen that users tend to use shared bikes to go downhill more than uphill. Moreover, considerable differences in demand are found for bike-sharing stations in commercial and residential areas, and a station's proximity to infrastructure, such as museums, shopping centres, schools, universities and restaurants, is investigated by many studies. Also, public transport options seem to be related to the bike-sharing demand. The more train, tram, bus and metro stations near a station, the higher its demand. Moreover, many studies have shown that the bike-sharing demand varies along the temporal dimension, with the most apparent differences occurring between weekdays and weekends due to different user travel motivations. The usage of bike sharing for commuting to work becomes visible via the peak usage during morning and afternoon rush hours on weekdays. On the other hand, trips during the weekend are more often for recreational purposes.

Most of the studies use linear models (e.g. generalised linear models (Chastenet de Castaing, 2017), hierarchical linear mixed effect models (El-Assi et al., 2017) or negative binomial models (Gebhart & Noland, 2014; Nair et al., 2013)) without explicitly addressing spatial dependence. Yang et al. (2020), Ji et al. (2018) and Wang et al. (2020) model the demand using ordinary least squares regression, which is inconsistent in the presence of spatial dependence, but they subsequently check residuals for spatial autocorrelation using Moran's I (Lee & Li, 2017). In other studies, spatial dependence is mostly addressed via cluster analyses (e.g. Froehlich et al., 2009; Lathia et al., 2012; Li et al., 2015; Raninen, 2018; Vogel et al., 2011; Zhou, 2015). In contrast, temporal dependence has been studied more accurately. For instance, Shi et al. (2018) studied metro riderships explicitly addressing its temporal dimension, while El-Assi et al. (2017) consider a first-order temporal autoregressive model. In some studies, temporal dependence has been ruled out for the dependent/independent variables through aggregation over time, for example, average number of trips per month (Rixey, 2013), per day (Buck & Buehler, 2012) or during the peak hours in the morning or afternoon (Nair et al., 2013; Tran et al., 2015; Wang et al., 2020).

The city of Helsinki introduced the station-based public bike sharing scheme in 2016 with 50 stations and further expanded it in 2017 with 100 additional stations (see City of Helsinki, a and Jäppinen et al., 2013). As a consequence of its high usage and popularity, the system was extended to the neighbouring cities Espoo (2018) (Helsinki Region Transport, a) and Vantaa (2019) (Helsinki Region Transport, b). Hence, the bike-sharing system covers wide areas of the larger Helsinki region and has become a dense network of bike-sharing stations, making this system an alternative transport mode. Helsinki's bike-sharing scheme has been addressed before in a few studies (see Chastenet de Castaing, 2017; Raninen, 2018; Tarnanen, 2017).

3 | HELSINKI'S BIKE-SHARING SYSTEM AND DATA DESCRIPTION

The City of Helsinki is located at the coast, the Gulf of Finland, and is characterised by a primarily flat area with a few hills. In our empirical study, we exploited the data from the bike sharing season 2017 for which the bike sharing service was already operational in Helsinki city centre shown in Figure 1. In general, the area shown is widely covered with bike sharing stations, although they are denser and more evenly distributed in the city centre. In the north of Helsinki city centre, where there are more residential areas, coverage becomes sparse and rather irregular. Moreover, a public transport network connects the city centre with the greater Helsinki area (outside the map extent) consisting of its suburbs and neighbouring cities Vantaa and Espoo. In Figure 1, the metro and train lines are indicated by the orange and purple lines, respectively, with the markers indicating the stations. In addition, there is a dense tram and bus network in Helsinki, but this is not shown on the map. The background colours of the map correspond to the four city land use categories assigned by the 2016 Helsinki master plan for city development (City of Helsinki, b). These four categories consist of the city centre (orange), shops (red), recreational areas (green) and predominantly residential areas (yellow).

The company Helsinki Region Transport has provided an API to enable individuals and organizations to develop their own applications and investigate the data related to transport in Helsinki and neighbouring municipalities (cf. Kainu, 2017 Helsinki Region Transport, c). We selected 176 days in 2017, starting on May 9, which was the first day with full records, and ending on October 31, which appears to have been the end of the biking season that year. However, the data are incomplete, as shown in Figure 2, where black entries depict missing values. It is worth noting that the functional HDGM can be estimated even when the responses are not available for all stations and/or time points.

The station hire data contain information on the observed number of bikes recorded every 5 min at 140 bike-share stations in Helsinki. Thus, there are over 7 million bicycle counts in

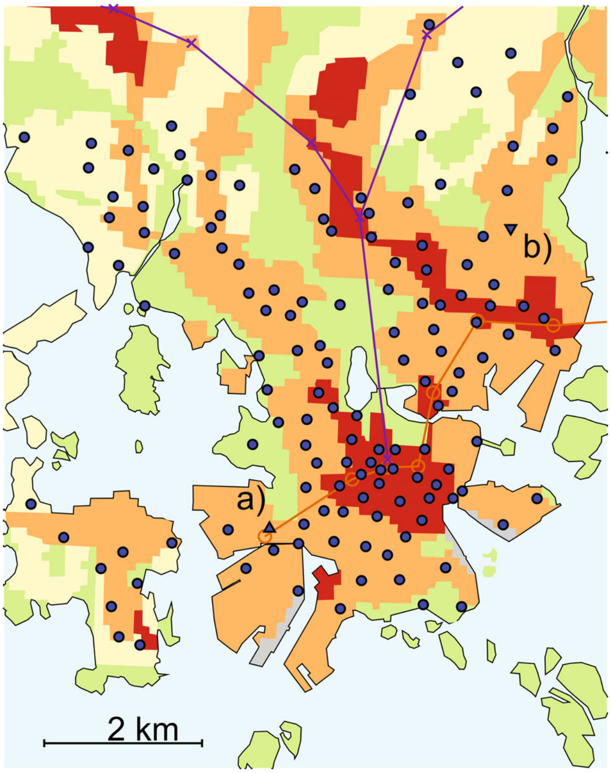


FIGURE 1 Map of Helsinki's city centre showing bike sharing stations (blue dots), metro (orange line) and train (purple line). The coloured areas indicate different land use as provided by Helsinki's master plan for city development 2016 (City of Helsinki, b). Please note that the metro is shown in the status of 2017 before the western extension of the metro line. Two triangles indicate the location of the bike-sharing stations (a) *Itämerentori* and (b) *Haukilahdenkatu* which are further analysed in Figure 4 [Colour figure can be viewed at wileyonlinelibrary.com]

total. In a preliminary analysis, the time series for the stations were analysed in the frequency domain (cf. Brockwell & Davis, 2016; Cooley & Tukey, 1965). The Fourier transform was applied to decompose the time series of each individual bike sharing station into the weighted sum of its underlying periodic signals. The resulting periodograms depict the magnitude for each temporal frequency present in each time series. Analysing the time series of the bike sharing stations in this frequency domain shows us that the stations differ with respect to their magnitudes of the dominating temporal frequencies. Periodograms were computed for all the stations separately and are depicted in Figure 3c via a glyph-map (Eden et al., 2010; Wickham et al., 2012) with the station *Arabiankatu* shown in Figure 3a to illustrate the scale and axes definition of the glyphs. The periodograms are shown as small glyphs at the locations of the respective bike-sharing stations.

The glyph-map reveals differences in the periodograms that appear to be spatially correlated. In the city centre close to the main station, the magnitudes are higher than in most parts in the north of Helsinki. Moreover, the bike-sharing stations in the city centre have periodograms with several dominant peaks. However, most of the periodograms for the stations have one outstanding peak in common. This dominant peak corresponds to the daily frequency. This finding is further highlighted in the histogram for the relative magnitudes in Figure 3b. The histogram shows an

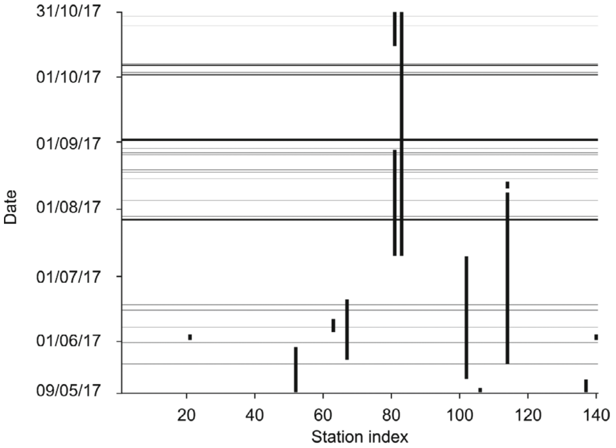


FIGURE 2 Missing data in the station hire dataset from 2017 in Helsinki. Black lines show the data gaps over time (horizontal) and for certain bike-sharing stations (vertical). The width of the horizontal lines corresponds to the length of the period that the data are missing

aggregation of all the periodograms from the glyph-map. Here, the magnitude $M(f_r)$ [%] of the r-th frequency f_r relative to the total signal of all stations is computed with

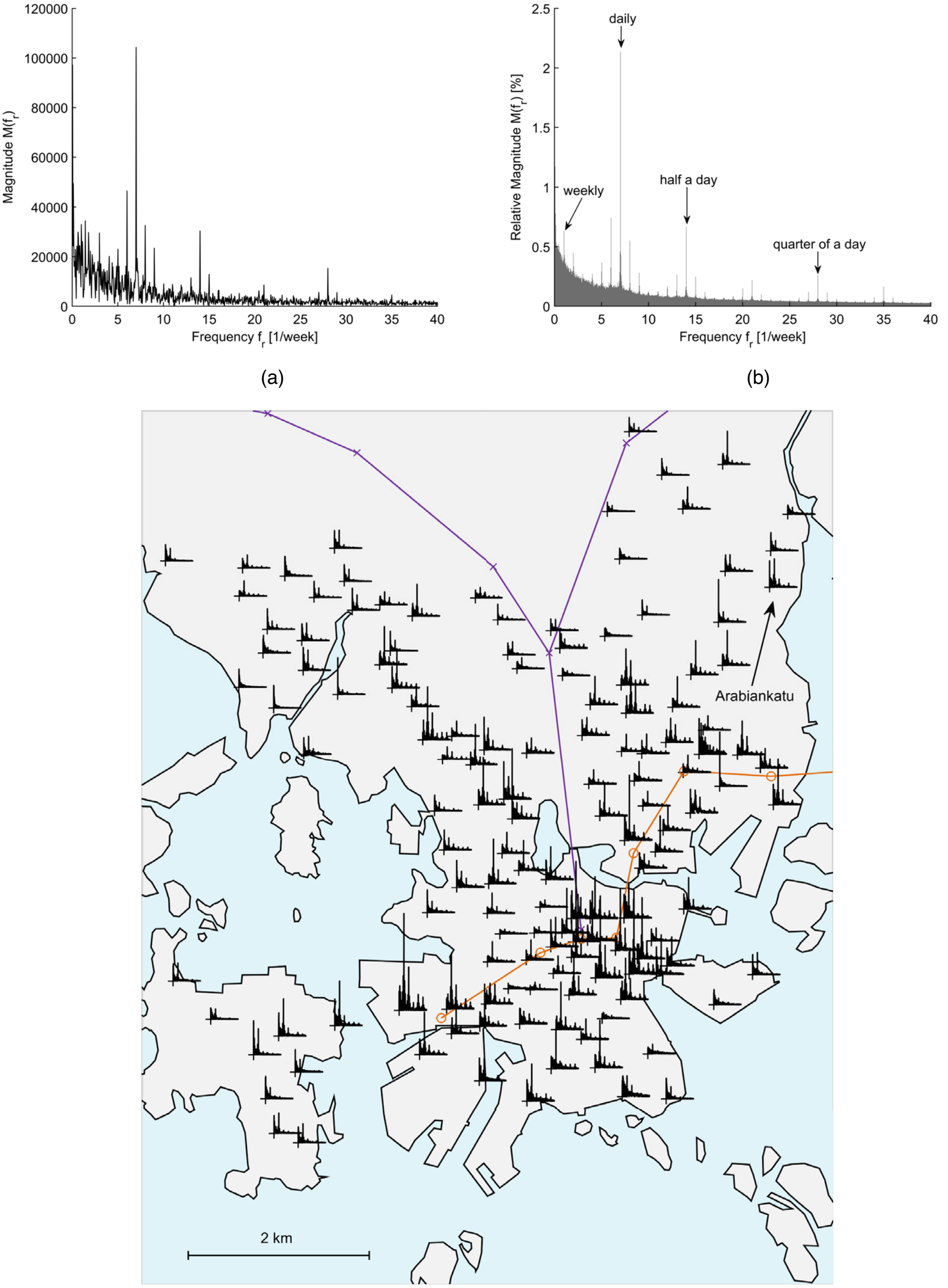
$$M(f_r)\ [\%] = \frac{\sum_{i=1}^{n} M_i(f_r)}{\sum_{i=1}^{n}\sum_{j=1}^{N_f} M_i(f_j)} \cdot 100, \tag{1}$$

where n is the number of stations, and N_f is the number of frequencies in the periodogram. According to the histogram in Figure 3b, the daily cycle is most prominent, representing over 2% of the total signal from all stations.

In general, time series can be subdivided into linear time granularities, for example, a sequence of subsequent days, and cyclic time granularities, for example, daily, weekly or yearly periodicities (Andrienko et al., 2010; Gupta et al., 2021). Moreover, cyclic data are often classified into regular and irregular cycles. For regular cycles, the time series are constructed by subdividing the linear time into pieces, and stacking them to match the cycles, for example, days, weeks, years, etc. The station hire data contain both temporal types of data, that is, linear time as the sequence of days and regular cyclic patterns on each weekday/weekend.

There are two predominant periodicities that can be understood as cyclic time. The most prominent cycle is the length of one day. Its prominence could be due to daily activity and sleep periods. Moreover, there is a cycle with a length of one week that is connected to the transition between workdays (Monday through Friday) and the weekend (Saturday and Sunday). Nevertheless, these repetitive structures appear for a sequence of days, which represents linear time. Both types of time have to be considered in the analysis of the bike-sharing data in order to cover all spatial and temporal dependencies within the data.

Therefore, this study makes the novel proposal that the cyclic time of one day should be treated as a functional observation (cf. Ferraty & Vieu, 2006; Ramsay & Silverman, 2007). To be precise, the number of bikes is one continuous function of time across a day, that is, a function that maps $h \in [0, 24]$ to the non-negative integer $y \in \mathbb{N}_0$. Thus, the daily cycle is incorporated into the functional observations for every day at every station. This daily cycle is denoted by h (time during the



(a)

(b)

(c)

FIGURE 3 Glyph-map of the periodograms of all the bike-sharing stations in Helsinki, with an additional periodogram of the bike-sharing station *Arabiankatu* and the histogram showing relative magnitudes. (a) Periodogram of bike-sharing station *Arabiankatu* illustrating scale and axes definition of the glyphs in the glyph-map. (b) Histogram of relative magnitudes of each frequency computed from the data for all bike-sharing stations. (c) Glyph-map [Colour figure can be viewed at wileyonlinelibrary.com]

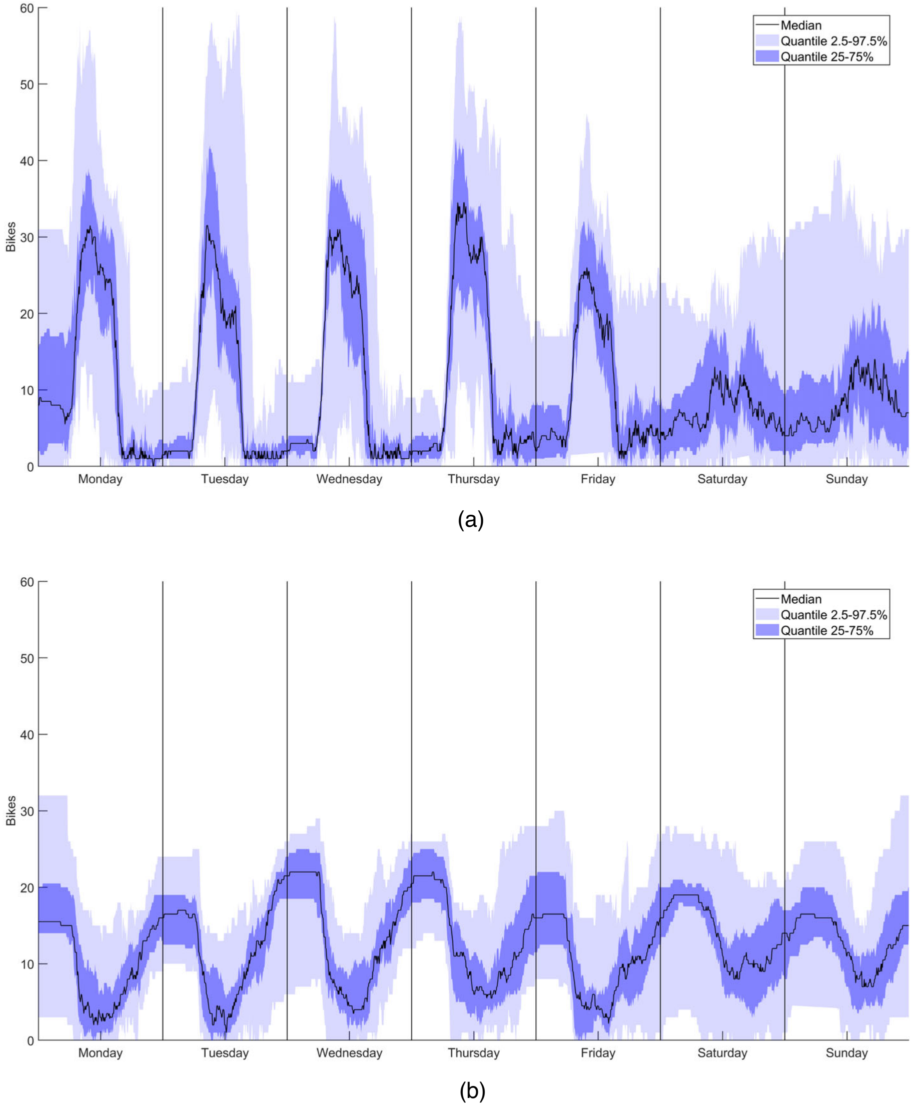


FIGURE 4 Functional boxplots of two bike-sharing stations summarizing their spatiotemporal functional observations. The spatial location of the two stations are marked with triangles in Figure 1. (a) Station *Itämerentori*. (b) Station *Haukilahdenkatu* [Colour figure can be viewed at wileyonlinelibrary.com]

day), whereas the day will be denoted by t below. It is important to note that h is not restricted to full hours, but can be any time point within the day.

To analyse the temporal dependence, functional boxplots are used (cf. Sun & Genton, 2011). Figure 4 show these plots for the stations *Itämerentori* (Figure 4a) and *Haukilahdenkatu* (Figure 4b). Both functional boxplots highlight the periodic behaviour belonging to the cycle of one day. The stations are characterised by a change in the number of bikes allocated during the morning hours from 7 to 10 and another major change in the afternoon hours from about 14 to 18 o'clock. However, the directions of change, as well as the ranges of observed bikes, are different. At *Itämerentori* station, the number of bikes increases in the morning and decreases in the afternoon. In addition, up to 58 bikes were observed at maximum. The opposite happens at *Haukilahdenkatu* station. There is a decrease in the number of bikes in the morning and an increase in the afternoon. Here, the maximum number of bikes observed was 32 bikes.

Because we observed these two different type of stations, we performed a preliminary cluster analysis. We found no evidence for including more than two clusters. The analysis of the explained variance shows that only very weak improvements are possible by including more clusters. To be precise, k-means clustering was applied to the median function of all observations, ensuring that these clusters are robust against outliers. Moreover, we chose Pearson's correlation as a distance measure between the median curves and the cluster centres.

The clustering was conducted separately for each day of the week. Hence, the clustering yielded $k = 2$ cluster centres $\tilde{\mu}_k(h)$ for each day of the week. These centres are shown in Figure 5a. The solid lines represent the cluster centres $\tilde{\mu}_k(h)$ for Monday through Friday, and the cluster centres for the weekend are denoted by dashed lines. The weekend's dashed lines are similar to each other but differ from the solid lines for Monday through Friday by a positive shift on the time axis of approximately 4 h. Additionally, the magnitude of change in the function is less for the cluster centres for the weekend than for the cluster centres $\tilde{\mu}_k(h)$ for Monday through Friday. Due to doing separate clusters for all days of the week, a station could be assigned to different clusters over the course of one week. However, for all days, the numbers of stations assigned to the clusters are similar, with roughly 40% of the stations belonging to the first cluster and consequently about 60% to the second cluster.

The assignments are shown in Figures 5b and c, where the sizes of the symbols denoting the locations of the stations are scaled by the number of assignments out of seven to that respective cluster. Looking more closely at the locations of both clusters, we can see that type 1 stations are mostly located in the centre, in areas where people work, while type 2 stations are located in regions where people live. We therefore interpret these main characteristic temporal patterns of the two clusters as the users commuting from home to work places and back. Hence, the clusters are hereafter named 'Work' and 'Home' respectively.

4 | MODELLING SPATIOTEMPORAL DEPENDENCE IN FUNCTIONAL DATA

Functional data analysis deals with a functional random variable Y that is continuously defined, as in, for example, Ferraty and Vieu (2006). Observations y of a functional random variable are either measured on a regular grid or at random discrete points, thus leading to a set of q discrete measurements $y_{i,1}, \ldots, y_{i,q}$ of the functional data $i \in 1, \ldots, m$. It is worth noting that q can be different for different functional data y_i. However, for functional data analysis, a continuous function is needed to evaluate the function $y(h)$ for any argument h. Hence, the functional form is reconstructed, for instance, using the basis function expansion (cf. Ndongo, 2017; Wang et al., 2016). That is, the reconstruction is accomplished with a set of K known basis functions ϕ_k with respective coefficients c_k, $y(h)$ that can be expressed as

$$y(h) = \sum_{k=1}^{K} \phi_k(h) c_k = \boldsymbol{\phi}^T(h)\mathbf{c}. \qquad (2)$$

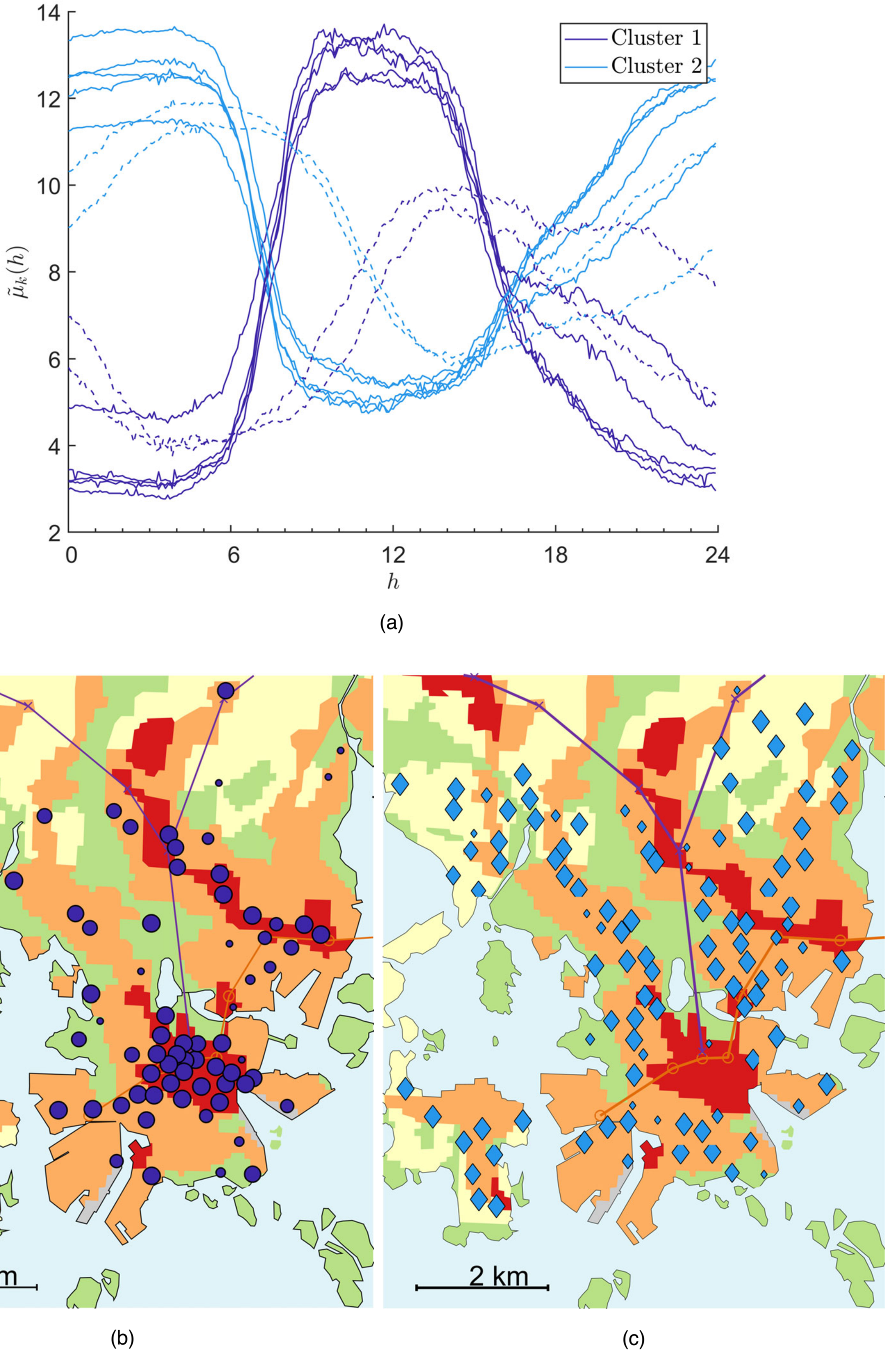


FIGURE 5 Assignment of the bike-sharing stations to the two clusters. The sizes of the markers are proportional to the number of assignments out of 7 weekdays to the corresponding cluster. (a) Centres of the two clusters for each day of the week. Monday through Friday centres are shown with solid lines, while the Saturday and Sunday centres are depicted by dashed lines. (b) Spatial allocation of cluster 'Work'. (c) Spatial allocation of cluster 'Home' [Colour figure can be viewed at wileyonlinelibrary.com]

Typical choices for the basis functions are the Fourier series for periodic data or the B-splines for non-periodic data (Ramsay & Silverman, 1997). For this analysis, we focus on the B-spline approach, where the number of free parameters is given by the order of the piecewise polynomials and the number of interior knots. The compact support of the B-spline basis functions has the advantage that the computational complexity increases only linearly with K. Furthermore, B-spline basis functions are flexible in the sense that the location of the break points can be chosen in order to approximate the function better in segments where it changes more frequently.

Let the number of available bikes in a bike-sharing station be described by a functional space–time random variable $Y(\boldsymbol{s}, t, h)$, and let $y(\boldsymbol{s}, t, h)$ be the observed functions from $h \in [0, 24] \subseteq \mathbb{R}$ to $\mathbb{R}$ at day t and station $\boldsymbol{s} \in D$, with $D \subset S^2$ and S^2 being a sphere in $\mathbb{R}^3$ (surface of the Earth). For this application, we consider that D is a discrete set of N bike-sharing stations. Even though the curvature of the Earth might be neglectable in this case because of the small extent of the city of Helsinki, we have used spherical coordinates. Thus, all distance measures reported below correspond to great-circle distances (in metres). Time is assumed to be discrete, with $t \in \{1, \dots, T\}$. Furthermore, the actual observations of $Y(\boldsymbol{s}, t, h)$ are made at q discrete points along the dimension of the function $y(\boldsymbol{s}, t, h)$, meaning, in the course of the day. More precisely, the number of available bikes is available in a 5-min frequency. Hence, the observation at $(\boldsymbol{s}, t)$ is the q-dimensional vector $\boldsymbol{y}(\boldsymbol{s}, t) = (y_1(\boldsymbol{s}, t), \dots, y_q(\boldsymbol{s}, t))^T$, where q equals 288 5-min intervals within 24 h. In this paper, we do not explicitly account for integer-valued observations of $Y(\boldsymbol{s}, t, h)$, because this would result in non-smooth function over the day, but we consider the number of available bikes in a station as continuous variable. The sample size m of the dataset is then given by the number of functional observations $m = NT = 24{,}640$ with $N = 140$ being the number of stations and $T = 176$ the number of days. The spatiotemporal functional variable $Y(\boldsymbol{s}, t, h)$ is assumed to be first-order stationary (i.e. the mean of the spatiotemporal process does not depend on the location).

A hierarchical model is used to model the mean spatiotemporal functional process and its variation by splitting up the total uncertainty into separate components. For multivariate data, the model is commonly known as hidden dynamic geostatistical model (HDGM, cf. Calculli et al., 2015; Wang et al., 2021). The first level is given by

$$y(\boldsymbol{s}, t, h) = \mu(\boldsymbol{s}, t, h) + \omega(\boldsymbol{s}, t, h) + \epsilon(\boldsymbol{s}, t, h), \tag{3}$$

where the $\mu(\boldsymbol{s}, t, h)$ are fixed effects, and the $\omega(\boldsymbol{s}, t, h)$ are spatially and temporally correlated random effects. The model errors $\epsilon(\boldsymbol{s}, t, h)$ are assumed to be from independent Gaussian white noise processes with a constant variance that is allowed to vary over the functional domain. More precisely,

$$\epsilon(\boldsymbol{s}, t, h) \sim N(0, \tilde{\sigma}^2(h)), \tag{4}$$

with

$$\tilde{\sigma}^2(h) = \boldsymbol{\phi}_\epsilon^T(h)\boldsymbol{\sigma}_\epsilon^2. \tag{5}$$

It is important to note that the spline basis functions could be chosen differently for each term. Hence, the basis functions ϕ_a^T and their dimension p_a are denoted by the subscript corresponding to the terms $a \in \{\mu, \omega, \varepsilon\}$ of the hierarchical model given in Equation (3).

The fixed effect model

$$\mu(\boldsymbol{s},t,h) = \sum_{i=1}^{d} x_{\mu,i}(\boldsymbol{s},t,h)\boldsymbol{\phi}_{\mu}^{T}(h)\boldsymbol{\beta}_i \tag{6}$$

consists of d space–time varying functional covariates $x_{\mu,i}(\boldsymbol{s},t,h)$, where the unknown coefficients $\boldsymbol{\beta}_i$ must be estimated. It is worth noting that these covariates could also be constant across space or time and/or in the functional dimension. Furthermore, the random effects model is given by

$$\omega(\boldsymbol{s},t,h) = \boldsymbol{\phi}_{\omega}^{T}(h)\boldsymbol{z}(\boldsymbol{s},t). \tag{7}$$

For each time step and each location a function is sampled from the estimated random effect model. It covers both spatial and temporal dependencies by modelling the respective variation using a basis function expansion. Specifically, the spatiotemporal latent component $\boldsymbol{z}(\boldsymbol{s},t)$ has the Markovian dynamics

$$\boldsymbol{z}(\boldsymbol{s},t) = \mathbf{G}\boldsymbol{z}(\boldsymbol{s},t-1) + \boldsymbol{\eta}(\boldsymbol{s},t). \tag{8}$$

Thus, the random effect merely depends on the previous time step and the innovation $\boldsymbol{\eta}(\boldsymbol{s},t)$. The degree of dependence from the previous time step is specified by the transition matrix $\mathbf{G}$ that is assumed to be stable. Here, $\mathbf{G}$ is a diagonal matrix $\mathbf{G} = diag(g_1, \ldots, g_{p_\omega})$ and accordingly the latent components $\boldsymbol{z}(\boldsymbol{s},t)$ are not cross-correlated across the functional dimension. For each day t an innovation $\boldsymbol{\eta}(t) = vec(\boldsymbol{\eta}(t,\boldsymbol{s}_1), \ldots, \boldsymbol{\eta}(t,\boldsymbol{s}_N))$ is sampled from the following spatially dependent Gaussian process

$$\boldsymbol{\eta}(t) \sim N\big(0, \mathbf{V} \otimes \rho\big(\|\boldsymbol{s}-\boldsymbol{s}'\|, \boldsymbol{\theta}, \nu\big)\big). \tag{9}$$

Here, $\otimes$ stands for the Kronecker product and *vec* is the vectorisation operator. While the cross-covariance matrix $\mathbf{V}$ describes the correlation between all components, the spatial covariance function $\rho(\|\boldsymbol{s}-\boldsymbol{s}'\|, \boldsymbol{\theta}, \nu)$ captures the correlation across space. For instance, this function could be a Matérn covariance function, which is an isotropic covariance function depending on the distance $\|\boldsymbol{s}-\boldsymbol{s}'\|$ between spatial locations only. The range of the spatial dependence is described by the coefficients $\boldsymbol{\theta}$ and ν cover potential further parameters. Moreover, the variance of the random effects is given by the matrix $\mathbf{V} = diag(\sigma^2_{\eta_1}, \ldots, \sigma^2_{\eta_{p_\omega}})$. It is worth noting that the cross-covariance matrix $\mathbf{V}$ is restricted to a diagonal matrix in order to reduce computational effort and, hence, it acts as a scaling matrix of the random effects. To estimate the parameters, we follow the maximum likelihood approach using an EM algorithm implemented in D-STEM (see Wang et al., 2021), which uses a functional hierarchical model called the f-HDGM model. For more details on the closed form and the numerical computations of the parameters in the EM algorithm, we refer the reader to Wang et al. (2021), Calculli et al. (2015), Fassò and Finazzi (2011) and Fassò and Cameletti (2009).

However, for this approach, the computational costs (in particular, for the computation of variance–covariance matrix of the estimated parameters) increase drastically with the number of spatial locations N and the number of splines p_a chosen for the basis function expansion. In our study of Helsinki's bike-sharing system, the number of bikes at 140 stations was observed every 5 min for 176 days. Thus, we propose to use the following subsampling procedure to estimate the standard errors of the model parameters more efficiently.

First, B independent samples of bike-sharing locations, $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_B \in \{\boldsymbol{s} \in D : y(\boldsymbol{s}, t, h)\}$, are generated, each consisting of M locations that are randomly drawn without replacement from all locations where the process is being observed. To reduce the computational costs, we have chosen $M \ll N$. Moreover, we followed a stratified sampling technique (see, e.g. Cochran, 2007), such that the stations of each cluster appear in their correct proportions in the subsamples $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_B$ (i.e. 40% of cluster 'Work', 60% of cluster 'Home'). Then the parameter of interest is estimated separately for each of the B samples.

The estimator of the i-th sample $\{y(\boldsymbol{s}, t, h) : \boldsymbol{s} \in \boldsymbol{b}_i\}$ is denoted by $\hat{\gamma}_i$. Lastly, $\hat{\gamma}$ is given by

$$\hat{\gamma} = \frac{1}{B}\sum_{i=1}^{B}\hat{\gamma}_i. \tag{10}$$

Furthermore, the standard error

$$\hat{\sigma} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\gamma}_i - \hat{\gamma})^2} \tag{11}$$

is the estimate of the true standard error of the reduced sample. Finally, the confidence interval of the estimate $\hat{\gamma}$ can be computed using the percentile method (see, e.g. Efron & Hastie, 2016). Here, the $(1-\alpha)$ confidence interval is approximated by the respective empirical quantiles of the distribution of $\{\hat{\gamma}_i\}$.

Since we have reduced the total number of stations in each subsample (i.e. M locations in each sample vs. N locations of the full sample) to make the calculation of the standard errors of the parameters computationally feasible, the estimated standard error must be interpreted as an upper bound of the true standard error of the whole sample. That is, all statistically significant results which we will report below would stay significant for full sample as well, while insignificant effects might be significant (but probably weak) when considering the full sample.

5 | EMPIRICAL RESULTS AND INTERPRETATION

In the following section, the focus is on the results of the empirical analysis. That is, the bike-sharing usage in Helsinki is analysed in the spatiotemporal functional framework to gain insights into the influence of functional covariates and interaction effects on the number of allocated bikes. This knowledge can be used to understand and predict the number of bikes at existing stations over the course of the day due to the functional setting. Moreover, rebalancing strategies could be improved based on our empirical results. To a limited degree, the model could also be used for predicting the bikes at new locations, that is, kriging. However, one has to keep in mind that the overall demand is not arbitrarily scalable by introducing new stations.

All included covariates with the notation of their corresponding coefficients are listed in Table 1. More precisely, we have fitted a model with two intercepts corresponding to the two cluster types 'Work' and 'Home'. Note that the cluster assignment was based on the preliminary analysis of the response variable. Moreover, a set of nine covariates varying either in space, in both time and the functional domain, or in time were included as interactions with each intercept. That is, the fixed effects models are considered like independent models for each cluster.

TABLE 1 Summary of the selected covariates and notation of the corresponding coefficients. The covariates are either varying in space or both in time or functional domain

	Coefficients		**Covariates are varying across**		
	Cluster 'Work'	**Cluster 'Home'**	**space (station s)**	**time (calendar day t)**	**functional domain (hour of the day h)**
Cluster intercepts					
Intercept	$\beta_{\text{Work}}(h)$	$\beta_{\text{Home}}(h)$			
Weather covariates					
Precipitation	$\beta_{\text{Work}*\text{Precipitation}}(h)$	$\beta_{\text{Home}*\text{Precipitation}}(h)$		✓	✓
Temperature	$\beta_{\text{Work}*\text{Temperature}}(h)$	$\beta_{\text{Home}*\text{Temperature}}(h)$		✓	✓
Wind speed	$\beta_{\text{Work}*\text{WindSpeed}}(h)$	$\beta_{\text{Home}*\text{WindSpeed}}(h)$		✓	✓
Cloud coverage	$\beta_{\text{Work}*\text{CloudCoverage}}(h)$	$\beta_{\text{Home}*\text{CloudCoverage}}(h)$		✓	✓
Geographical covariates					
Elevation	$\beta_{\text{Work}*\text{Elevation}}(h)$	$\beta_{\text{Home}*\text{Elevation}}(h)$	✓		
Infrastructure covariates					
Distance to metro	$\beta_{\text{Work}*\text{Metro}}(h)$	$\beta_{\text{Home}*\text{Metro}}(h)$	✓		
Distance to train	$\beta_{\text{Work}*\text{Train}}(h)$	$\beta_{\text{Home}*\text{Train}}(h)$	✓		
Weekend indicators					
Saturday	$\beta_{\text{Work}*\text{Saturday}}(h)$	$\beta_{\text{Home}*\text{Saturday}}(h)$		✓	
Sunday	$\beta_{\text{Work}*\text{Sunday}}(h)$	$\beta_{\text{Home}*\text{Sunday}}(h)$		✓	
Spatiotemporal dependence					
Temporal dependence	$\mathbf{G}$				
Range of spatial dependence	$\boldsymbol{\theta}$				
Scale matrix	$\mathbf{V}$				

However, these two models are linked by the common random effects model, which accounts for the spatiotemporal dependence of the data.

In the final model, we selected several meteorological covariates. Three observatories are located within the spatial extent of the bike-sharing stations in Helsinki, but we only have used data from the Kaisaniemi observatory located in the central city. The spatial differences in these weather covariates are neglectable; hence, the observations are assumed to be constant over space but not time. Figure 6 shows the four weather covariates for the period from May 9 to the October 31, 2017. Furthermore, the respective histogram shows the distribution of the meteorological observations with the relative frequencies of occurrence. The histograms on the right-hand side of Figure 6 are aligned with the observations over time on the left-hand side.

In summary, we included dummy variables for Saturday and Sunday showing the weekend effects, meteorological variables (i.e. temperature in [°C], cloud coverage in [%], wind speed in [m/s], precipitation in [mm]), a geographical variable, namely the elevation of the station in [m], and two infrastructure variables (i.e. first, the distance of a bike-sharing station to its closest metro and second, to its closest train station in [km]). Doing so leads to two intercept functions and nine interactions for each cluster; in total, 20 parameter functions must be estimated (each consisting of several spline coefficients).

The model set-up (i.e. the choice of the spatial covariance function, splines basis, knots and so on) was determined via a cross-validation study using $B = 1000$ subsamples and a sample size of 30 stations for both the in-sample and out-of-sample cases. Note that the choice of size is a trade-off between reliability and computational complexity. The stations were drawn from the first and the second cluster according to the proportions described above (i.e. 41.4% for cluster 'Work', and 58.6% for cluster 'Home') to obtain distinct in-sample and out-of-sample sets in each subsample.

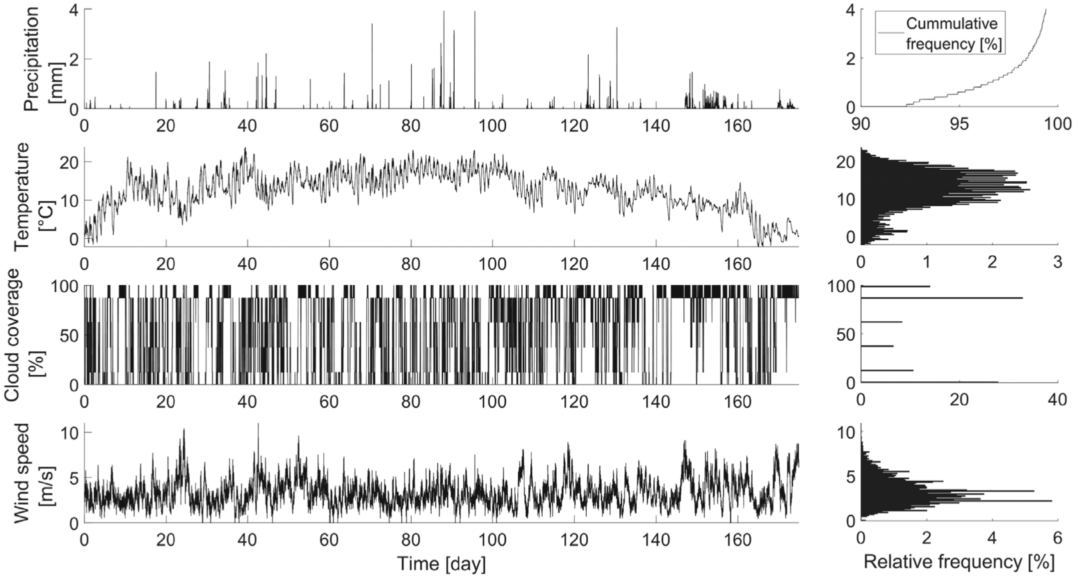


FIGURE 6 Time series and histograms for four meteorological variables from May 9 to October 31, 2017. The histograms on the right-hand side refer to the time series on the left-hand side. A cumulative histogram is shown for precipitation

In our case, an exponential covariance function fits the data the best. Hence, there are no anisotropic dependencies, which would indicate a prevalent direction of bike usage. Furthermore, for the basis function expansion, the B-splines approach was chosen, although the spectral time series analysis revealed periodic structures in the time series for all the stations. However, the B-splines allow the knot positions to be adapted according to the variation in the data along the functional dimension. For this study, we have determined the knot positions in such a manner that the standard deviation $\tilde{\sigma}(h)$ of the modelling error $\epsilon(\boldsymbol{s}, t, h)$ remained less than three bikes. To be precise, the position of the break points was set to

$$\text{break points} = \{0, 5, 7.14, 9.29, 11.43, 13.57, 15.71, 17.86, 20, 24\} \text{ o' clock}$$

with only a few B-splines supporting the morning and evening hours, as there is little variation in the spatiotemporal functional observations during these periods. In contrast, there is high variation in the functional observations during the middle of the day; thus, the splines basis functions are denser during this period.

5.1 | Fixed effects

The two intercepts referring to the stations' cluster memberships are shown in Figure 7. The mean curve of $\hat{\boldsymbol{\beta}}_{\text{Work}}(h)$ shows about seven bikes during the night and in the evening, while the number of bikes increases during the day and has a first peak at approximately 11 o'clock, with about 17 bikes, and a second peak at 15 in the afternoon, with 19 bikes. The confidence interval has the widest range at the peaks. As expected, $\hat{\boldsymbol{\beta}}_{\text{Home}}(h)$ shows the opposite shape. In the beginning and the end of the day, approximately 17 bikes are located at stations from cluster 'Home'. The number grows smaller during the day, with the minimum of nine bikes occurring at approximately 9 in the morning. Between 10 and 14 o'clock, there are 10 bikes; afterwards, the number of bikes increases again. For both intercepts, the 95%-confidence interval along the entire function indicates a range of possible values of up to ±2 bikes around the mean. However, these intercept curves represent the expected number of bikes in the case that all other covariates would be zero. Thus, they have to be rather interpreted along with the regressive effects.

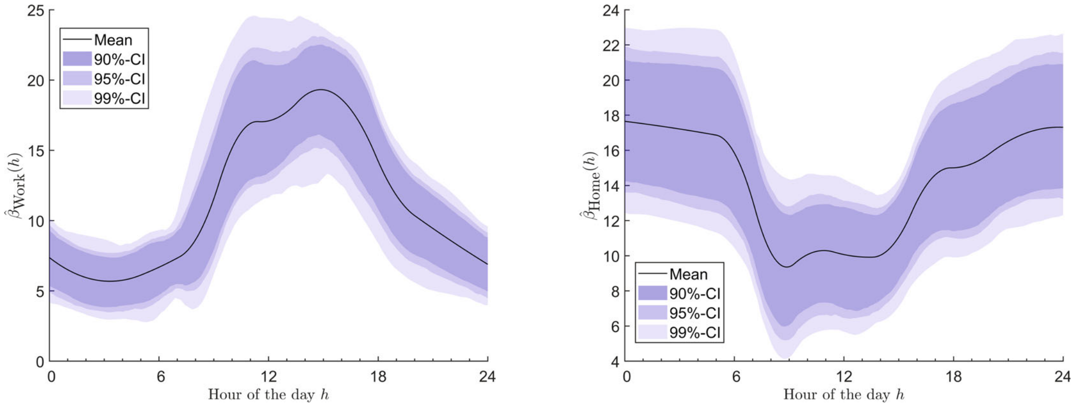


FIGURE 7 Intercept: estimated functional intercepts for $\hat{\boldsymbol{\beta}}_{\text{Work}}(h)$ (left) and $\hat{\boldsymbol{\beta}}_{\text{Home}}(h)$ (right) [Colour figure can be viewed at wileyonlinelibrary.com]

For instance, in Figure 8, the temperature effects in both clusters are depicted (i.e. $\hat{\boldsymbol{\beta}}_{\text{Work}*\text{Temperature}}(h)$ and $\hat{\boldsymbol{\beta}}_{\text{Home}*\text{Temperature}}(h)$). In the areas for working, we observe that the number of allocated bikes changes by −0.2 bikes/°C in the night and evening. Between 8 and 15 o'clock, the influence of the temperature is not significantly different from zero. As a consequence, the higher the temperature in Helsinki, the fewer bikes are located at the stations from cluster 'Work' in the evening and night, while a temperature change between 8 and 15 o'clock has no significant effect on the number of bikes. Regarding residential areas, the interaction starts in the night at around 0.1 bikes/°C, decreases rapidly to −0.15 bikes/°C from 11 to 16 o'clock and increases afterwards up to 0.14 bikes/°C. The shape of the function is similar to the valley-like shape of the intercept $\hat{\boldsymbol{\beta}}_{\text{Home}}(h)$. Consequently, the higher the temperature in Helsinki, the more bikes are located at the stations from cluster 'Home' in the night and evening, while the number of allocated bikes decreases during the daytime.

Both interactions of the temperature with the cluster effects yield functions with shapes similar to that of the function of the intercept itself, meaning that an increase in the temperature amplifies the already existing mountain- and valley-like shape of the intercepts. The effect of the temperature shows that usage of the bike-sharing scheme is higher when it is warmer, as the change in the number of allocated bikes at stations from both clusters increases. This effect becomes more clear when combining each estimated functional interaction covariate $\hat{\boldsymbol{\beta}}_{\text{Work}*\text{Temperature}}(h)$ and $\hat{\boldsymbol{\beta}}_{\text{Home}*\text{Temperature}}(h)$ with their corresponding functional intercepts $\hat{\boldsymbol{\beta}}_{\text{Work}}(h)$ and $\hat{\boldsymbol{\beta}}_{\text{Home}}(h)$ respectively. Figure 9 illustrates the predicted number of bikes of a station at cluster 'Work' or 'Home' depending on the temperature. Importantly, all other covariates and the random effects are considered to be zero. The solid black lines show the respective intercept curves. For example, the number of allocated bikes at a station from cluster 'Home' at noon (indicated by the vertical black line) ranges from roughly 7 to 9 bikes, merely due to temperature variation when at noon temperature is realistically assumed to range between 10 and 25°C.

The influence of a bike-sharing station's elevation on the number of allocated bikes is given by $\hat{\boldsymbol{\beta}}_{\text{Work}*\text{Elevation}}(h)$ and $\hat{\boldsymbol{\beta}}_{\text{Home}*\text{Elevation}}(h)$, as shown in Figure 10. Both functions are significant and have negative signs along their entire domains. The influence of the interaction of elevation with the cluster 'Home' is around −0.4 bikes/m. There is little variation around the mean, which does not change significantly. In comparison, the interaction $\hat{\boldsymbol{\beta}}_{\text{Work}*\text{Elevation}}(h)$ varies more. Due

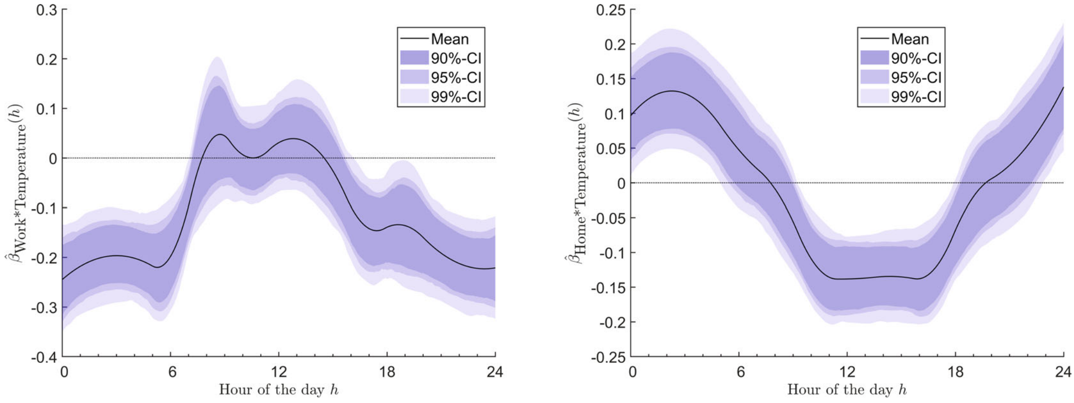


FIGURE 8 Temperature: estimated functional influences of $\hat{\boldsymbol{\beta}}_{\text{Work}*\text{Temperature}}(h)$ (left) and $\hat{\boldsymbol{\beta}}_{\text{Home}*\text{Temperature}}(h)$ (right) [Colour figure can be viewed at wileyonlinelibrary.com]

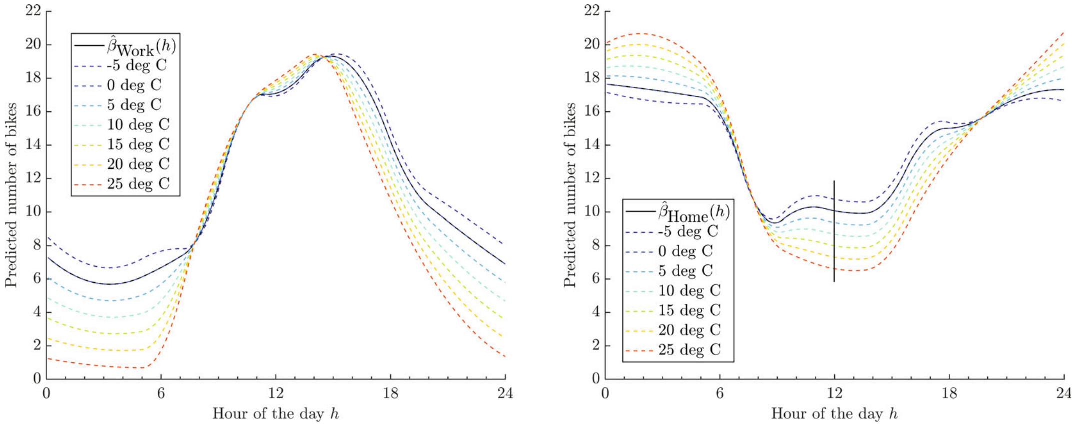


FIGURE 9 Temperature: combination of the estimated functional influences $\hat{\boldsymbol{\beta}}_{\text{Work}*\text{Temperature}}(h)$ (left) and $\hat{\boldsymbol{\beta}}_{\text{Home}*\text{Temperature}}(h)$ (right) with their functional intercepts $\hat{\boldsymbol{\beta}}_{\text{Work}}(h)$ and $\hat{\boldsymbol{\beta}}_{\text{Home}}(h)$, respectively, for temperature in the range of −5 to 25°C. All other interaction covariates and the random effects are considered to be zero [Colour figure can be viewed at wileyonlinelibrary.com]

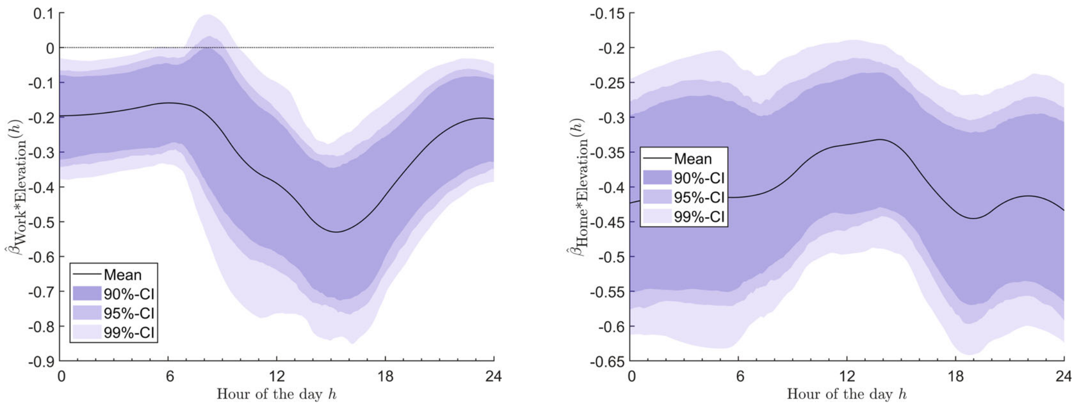


FIGURE 10 Elevation: estimated functional influences of $\hat{\boldsymbol{\beta}}_{\text{Work}*\text{Elevation}}(h)$ (left) and $\hat{\boldsymbol{\beta}}_{\text{Home}*\text{Elevation}}(h)$ (right) [Colour figure can be viewed at wileyonlinelibrary.com]

to the negative sign, the number of allocated bikes at a station decreases as the elevation of the station increases, showing empirically that cyclists prefer cycling downhill over cycling uphill. The change in $\hat{\boldsymbol{\beta}}_{\text{Work}*\text{Elevation}}(h)$ in the afternoon emphasises that cyclists might use the bikes even less for cycling uphill in their free time.

Most surprising are the results regarding the effect of precipitation as no effects could be identified from the data. However, bear in mind that only 8 % of all precipitation observations were non-zero. Including precipitation as a dummy variable could perhaps enable to see an effect for precipitation. Furthermore, public transport was included in the model, as Jäppinen et al. (2013) suggested that it was one of the major factors influencing bike-sharing usage. The influence of the distance of a bike-sharing station to public transport varies as shown in Figure 11. While the estimate $\hat{\boldsymbol{\beta}}_{\text{Home}*\text{Train}}(h)$ is not significantly different from zero, $\hat{\boldsymbol{\beta}}_{\text{Work}*\text{Train}}(h)$ shows significant effects from 6 to 8 in the morning and from 15 to 17 in the afternoon. In the morning, the number of bikes increases by about 3 bikes/km with increasing distance from the closest train station. The

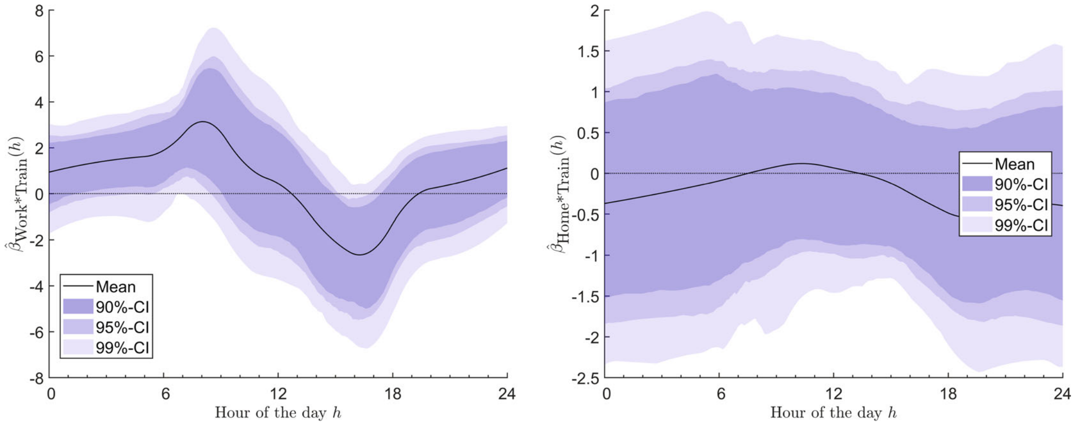


FIGURE 11 Distance to train station: estimated functional influence of $\hat{\boldsymbol{\beta}}_{\text{Work}*\text{Train}}(h)$ (left) and $\hat{\boldsymbol{\beta}}_{\text{Home}*\text{Train}}(h)$ (right) [Colour figure can be viewed at wileyonlinelibrary.com]

opposite is the case in the afternoon, with −2.2 bikes/km with increasing distance from the closest train station. The intercept of the cluster ‘Work’ increases during the respective morning period and decreases in the afternoon. This change is amplified by the interaction $\hat{\boldsymbol{\beta}}_{\text{Work}*\text{Train}}(h)$, meaning that more bikes are allocated to bike-sharing stations further away from the train stations. This finding supports the hypothesis that commuters use the public bike-share scheme to overcome distances from the train station to their destinations, also known as the last-mile problem. Moreover, the greater the distance of a bike-sharing station from the closest metro station, the more bikes are allocated to this station during the morning until the mid-afternoon. Thus, stations closer to the public transportation system are used more frequently (i.e. there are fewer bikes allocated during the day). Interestingly, the influence of the distance to metro stations is only less than half the effect of the distance to train stations.

5.2 | Random effects

Below, we discuss the results of random effects and the error term briefly. The random effects model explains the random variation in the data that is not explained by the fixed effects, that is, the mean of the process and the covariates. First, the temporal autoregressive dependence is estimated with the diagonal elements of the transition matrix $\hat{\mathbf{G}}$. The median values of the estimates of these diagonal elements range from roughly 0.45 to 0.50, where all elements are similar. The minimal and maximal values are in the ranges of [0.27, 0.38] and [0.54, 0.64] respectively. Consequently, about half of the random variation observed in a day is explained by the previous day. To analyse the spatial dependence, Figure 12 shows the estimated range parameters $\hat{\boldsymbol{\theta}}$. Their median values are around 160 m. Since the exponential covariance function declines rapidly and the covariance is about 0.37 at distance θ, the spatial dependence of the process is weak in most cases—the largest values are in the range [264, 415] m. Considering the distances between the bike-sharing stations in Helsinki, where the median distance is 3 km, the estimated range parameters $\hat{\boldsymbol{\theta}}$ reveal that the spatial dependence is constrained to stations that are very close together. Whereas, the diagonal elements of the matrix $\hat{\mathbf{V}}$ indicate a smaller random effect at night than

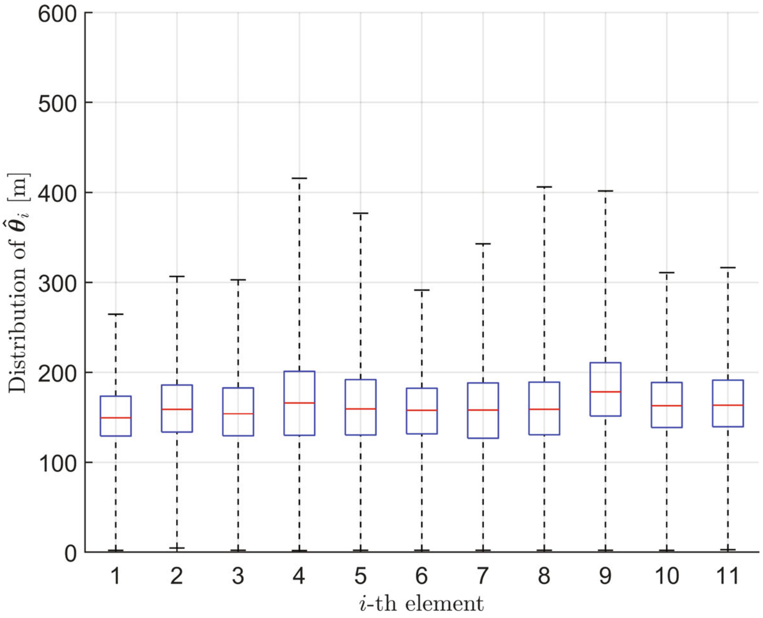


FIGURE 12 Distribution of the estimated range parameters $\hat{\boldsymbol{\theta}}$. The boxplots show the median in red, the 50%-central region is shown in blue and the whiskers denote the minimum and maximum values of the distribution [Colour figure can be viewed at wileyonlinelibrary.com]

during the day, it seems that the degree of the spatial dependence is mostly stable, with slightly more variability occurring during the day.

5.3 | Out-of-sample forecasts

Finally, one question remains open—how well does the model fit the data? To answer this question, an out-of-sample study was conducted for each of the 1000 subsamples. More precisely, 30 randomly chosen locations which were not included for estimation are used to compute the out-of-sample fit. The out-of-sample RMSE is depicted in Figure 13 with an orange line. For comparison, two alternative models (i.e. a simple intercept-only model, and a model with all regressors but no interactions with the clusters) are shown with blue and red curves. Interestingly, the three functions have roughly the same shape but are shifted. The RMSE is the lowest from midnight to 7 o'clock and increases drastically between 7 and 9 o'clock. After reaching a maximum at 9 in the morning, it decreases until 16 o'clock and has a local maximum between 17 and 19 o'clock. The range is ±1.5 bikes over the course of the day. Considering the daily out-of-sample RMSE, we observe median values varying between 5 and 9 bikes. Moreover, the estimated functional error variance $\hat{\hat{\sigma}}^2(h)$ is shown in Figure 13. Across the entire day, the maximal standard error is less than two bikes. During the night from about 1 o'clock to 5 o'clock, the variance is the smallest, showing that the variation in the data for all bike-sharing stations is low in the night. Also, from 22 in the evening until midnight, the standard deviation is less than one bike. In contrast, the variance is higher in the morning from 6 to 9 and in the afternoon from 15 to 19 o'clock. Here, the corresponding standard deviation is up to about 1.9 bikes.

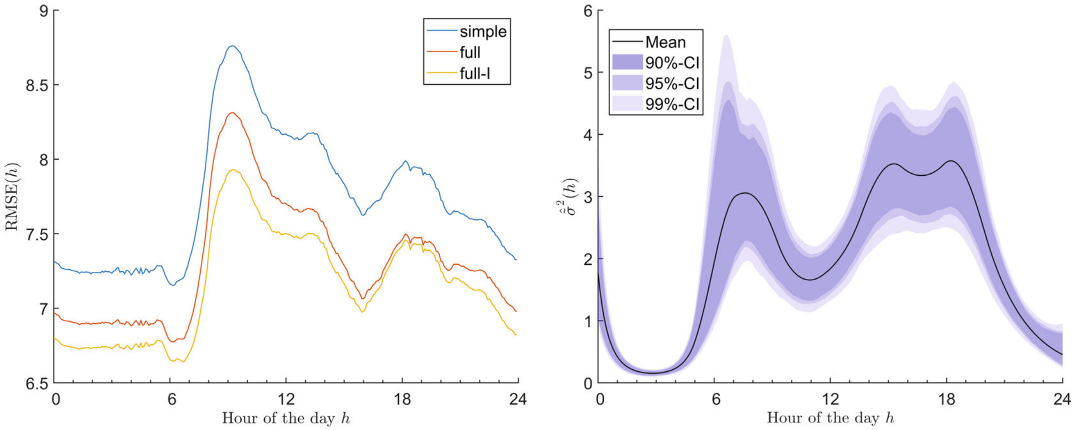


FIGURE 13 Comparison of the functional RMSE(h) (left) values for three selected models (blue: intercept-only, red: full model without interactions and orange: full model with interactions) and the estimated functional measurement variances $\hat{\hat{\sigma}}^2(h)$ (right) [Colour figure can be viewed at wileyonlinelibrary.com]

6 | CONCLUSION

This paper focused on station hire data from the bike-sharing system in Helsinki. With over 7 million observations from 140 bike-sharing stations taken at 5-min intervals, the analysis of bike-sharing station usage was brought to a new level, as the entire complex dataset was considered and knowledge about the changes in the influencing factors over the course of a day was inferred. We simultaneously accounted for spatial, temporal and spatiotemporal dependence by applying a geostatistical model in a functional framework. The model parameters were estimated using the implemented maximum-likelihood approach of the software package D-STEMv2 (see Wang et al., 2021). To supply computationally efficient estimated standard errors and guarantee a certain robustness against outliers, a subsampling approach was applied.

Most findings about the influencing factors were in line with the results from existing literature, although comparability is limited due to the use of a different methodology and data. We have shown that the bike-sharing stations can be divided into two clusters depending on the similarities in their spatiotemporal functional observations. It is important to note that these similar functional observations cluster together in space too. The estimated parameters have shown that the morning rush hour is particularly difficult to model and predict. There is a mountain-like shape to the daily available bikes for stations belonging to predominantly working areas. By contrast, we observe a valley-like shape in living areas. This behaviour is different on weekends, where the daily peaks are also shifted towards the afternoon. Furthermore, we examined which weather conditions could have an influence. According to Eren and Uz (2020), precipitation should affect bike-sharing station usage the most among the weather conditions. Here, however, the influence of precipitation was not significant.

A drawback of the model can be seen in the out-of-sample RMSE values, which ranged between seven and nine bikes, giving it a range similar to that for the random effects. The random effects covered the unexplained variation in the station hire data. Unfortunately, out-of-sample validation of models from the literature was not found, meaning that the error mentioned above could not be compared and evaluated. Moreover, a spatiotemporal integer-valued model might improve the prediction accuracy, especially for less frequently used stations. To better understand the implications of different types of the bike-sharing station usage, future studies could address spatially varying covariates, perhaps providing insights into the cause of the separation into clusters. On the other hand, bike-sharing system station usage is governed by the decisions made by individuals and maybe even pure coincidences in their behaviour. In general, analysing the relationship between the number of allocated bikes at the bike-sharing stations and the proposed covariates produces a correlation and does not necessarily imply causality.

It is worthwhile to develop and apply complex models to spatiotemporal functional data from bike-sharing systems, as detailed knowledge can be gained and perhaps lead to future improvements in implemented bike-sharing systems.

ACKNOWLEDGEMENT

We sincerely thank two anonymous reviewers for their helpful comments and suggestions. We would also like to thank Jouni Kuha for his detailed comments and suggestions, especially concerning the discussion of the specifics of Helsinki.

ORCID

Andreas Piter https://orcid.org/0000-0002-5566-3851
Philipp Otto https://orcid.org/0000-0002-9796-6682
Hamza Alkhatib https://orcid.org/0000-0002-4480-1067

REFERENCES

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S.I. et al. (2010) Space, time and visual analytics. *International Journal of Geographical Information Science*, 24, 1577–1600.

Brockwell, P.J. & Davis, R.A. (2016) *Introduction to time series and forecasting*. New York: Springer.

Buck, D. & Buehler, R. (2012) Bike lanes and other determinants of capital bikeshare trips. In: *91st transportation research board annual meeting*.

Calculli, C., Fassò, A., Finazzi, F., Pollice, A. & Turnone, A. (2015) Maximum likelihood estimation of the multivariate hidden dynamic geostatistical model with application to air quality in Apulia, Italy. *Environmetrics*, 26, 406–417.

Chastenet de Castaing, L. (2017) Cycling as a part of sustainable urban transport in Helsinki: assessing the influence of weather on cycling activity. Master's thesis.

City of Helsinki (a) Website, Online available from: https://www.hel.fi/hkl/en/by-bike/city-bikes/

City of Helsinki (b) Website. Online available from: https://kartta.hel.fi/

Cochran, W.G. (2007) *Sampling techniques*. New York: John Wiley & Sons.

Cooley, J.W. & Tukey, J.W. (1965) An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19, 297–301.

Eden, S.K., An, A.Q., Horner, J., Jenkins, C.A. & Scott, T.A. (2010) A two-step process for graphically summarizing spatial temporal multivariate data in two dimensions. *Computational Statistics*, 25, 587–601.

Efron, B. & Hastie, T. (2016) *Computer age statistical inference*, vol. 5. Cambridge: Cambridge University Press.

El-Assi, W., Mahmoud, M.S. & Habib, K.N. (2017) Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. *Transportation*, 44, 589–613.

Eren, E. & Uz, V.E. (2020) A review on bike-sharing: the factors affecting bike-sharing demand. *Sustainable Cities and Society*, 54, 101882.

Fassò, A. & Cameletti, M. (2009) The EM algorithm in a distributed computing environment for modelling environmental space–time data. *Environmental Modelling & Software*, 24, 1027–1035.

Fassò, A. & Finazzi, F. (2011) Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics*, 22, 735–748.

Fassò, A. & Finazzi, F. (2013) A varying coefficients space-time model for ground and satellite air quality data over Europe. *Statistica & Applicazioni, Special Online Issue*, 45–56.

Fassò, A., Finazzi, F. & Ndongo, F. (2016) European population exposure to airborne pollutants based on a multivariate spatio-temporal model. *Journal of Agricultural, Biological, and Environmental Statistics*, 21, 492–511.

Fassò, A., Finazzi, F. & Madonna, F. (2018) Statistical issues in radiosonde observation of atmospheric temperature and humidity profiles. *Statistics & Probability Letters*, 136, 97–100.

Ferraty, F. & Vieu, P. (2006) *Nonparametric functional data analysis: theory and practice*. Berlin: Springer Science & Business Media.

Finazzi, F. & Fassò, A. (2014) D-STEM: a software for the analysis and mapping of environmental space-time variables. *Journal of Statistical Software*, 62, 1–29.

Finazzi, F., Scott, E.M. & Fassò, A. (2013) A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62, 287–308.

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

Froehlich, J.E., Neumann, J. & Oliver, N. (2009) Sensing and predicting the pulse of the city through shared bicycling. In: *Twenty-first international joint conference on artificial intelligence*.

García-Palomares, J.C., Gutiérrez, J. & Latorre, M. (2012) Optimizing the location of stations in bike-sharing programs: a GIS approach. *Applied Geography*, 35, 235–246.

Gebhart, K. & Noland, R.B. (2014) The impact of weather conditions on bikeshare trips in Washington, DC. *Transportation*, 41, 1205–1225.

Gervini, D. & 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, 585–602.

Gupta, S., Hyndman, R.J., Cook, D. & Unwin, A. (2021) Visualizing probability distributions across bivariate cyclic temporal granularities. *Journal of Computational and Graphical Statistics*, 1–26.

Helsinki Region Transport (a) Website. Online available from: https://www.hsl.fi/en/news/2018/record-breaking-city-bike-season-2018-16277

Helsinki Region Transport (b) Website. Online available from: https://www.hsl.fi/en/news/2019/city-bike-season-vantaa-start-june-17462

Helsinki Region Transport (c) Website. Online available from: https://digitransit.fi/en/

Jäppinen, S., Toivonen, T. & Salonen, M. (2013) Modelling the potential effect of shared bicycles on public transport travel times in Greater Helsinki: an open data approach. *Applied Geography*, 43, 13–24.

Ji, Y., Ma, X., Yang, M., Jin, Y. & Gao, L. (2018) Exploring spatially varying influences on metro-bikeshare transfer: a geographically weighted poisson regression approach. *Sustainability*, 10, 1526.

Kainu, M. (2017) Kaupunkifillari17. Website. Online available from: https://gitlab.com/muuankarski/kaupunkifillari17

Lathia, N., Ahmed, S. & Capra, L. (2012) Measuring the impact of opening the London shared bicycle scheme to casual users. *Transportation Research Part C: Emerging Technologies*, 22, 88–102.

Lee, J. & Li, S. (2017) Extending Moran's index for measuring spatiotemporal clustering of geographic events. *Geographical Analysis*, 49, 36–57.

Li, Y., Zheng, Y., Zhang, H. & Chen, L. (2015) Traffic prediction in a bike-sharing system. In: *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems*, pp. 1–10.

Maranzano, P., Ascari, R., Chiodini, P.M. & Manzi, G. (2020) Analysis of sustainability propensity of bike-sharing customers using partially ordered sets methodology. *Social Indicators Research*, 1–16.

Martinez, L.M., Caetano, L., Eirò, T. & Cruz, F. (2012) An optimisation algorithm to establish the location of stations of a mixed fleet biking system: an application to the city of Lisbon. *Procedia-Social and Behavioral Sciences*, 54, 513–524.

Nair, R., Miller-Hooks, E., Hampshire, R.C. & Bušić, A. (2013) Large-scale vehicle sharing systems: analysis of Vélib. *International Journal of Sustainable Transportation*, 7, 85–106.

Ndongo, F.B. (2017) Spatio-temporal processes for functional data with application in climate monitoring. Ph.D. thesis, Università degli studi di Bergamo.

O'brien, O., Cheshire, J. & Batty, M. (2014) Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography*, 34, 262–273.

Ramsay, J. & Silverman, B. (1997) *Functional data analysis*. Berlin: Springer.

Ramsay, J.O. & Silverman, B.W. (2007) *Applied functional data analysis: methods and case studies*. Berlin: Springer.

Raninen, M. (2018) Spatiotemporal analysis of a bike sharing system in Helsinki. Master's thesis. Finnish Title: Helsingin kaupunkipyöräjärjestelmä: käytön alueelliset ja ajalliset rakenteet.

Rixey, R.A. (2013) Station-level forecasting of bikesharing ridership: station network effects in three US systems. *Transportation Research Record*, 2387, 46–55.

Rue, H., Martino, S. & Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71, 319–392.

Schuijbroek, J., Hampshire, R.C. & Van Hoeve, W.-J. (2017) Inventory rebalancing and vehicle routing in bike sharing systems. *European Journal of Operational Research*, 257, 992–1004.

Shi, Z., Zhang, N., Liu, Y. & Xu, W. (2018) Exploring spatiotemporal variation in hourly metro ridership at station level: the influence of built environment and topological structure. *Sustainability*, 10, 4564.

Shi, L., Zhang, Y., Rui, W. & Yang, X. (2019) Study on the bike-sharing inventory rebalancing and vehicle routing for bike-sharing system. *Transportation Research Procedia*, 39, 624–633.

Sun, Y. & Genton, M.G. (2011) Functional boxplots. *Journal of Computational and Graphical Statistics*, 20, 316–334.

Taghavi-Shahri, S.M., Fassò, A., Mahaki, B. & Amini, H. (2019) Concurrent spatiotemporal daily land use regression modeling and missing data imputation of fine particulate matter using Distributed Space-Time Expectation Maximization. *Atmospheric Environment*, 117202.

Tarnanen, A. (2017) Modelling cycling speeds and travel times in the Helsinki region. Master's thesis, University of Helsinki. Finnish Title: Pyöräilyn nopeuksien ja matkaaikojen paikkatietopohjainen mallinnus pääkaupunkiseudulla.

Tran, T.D., Ovtracht, N. & d'Arcier, B.F. (2015) Modeling bike sharing system using built environment factors. *Procedia CIRP*, 30, 293–298.

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

Wang, J.-L., Chiou, J.-M. & Müller, H.-G. (2016) Functional data analysis. *Annual Review of Statistics and its Application*, 3, 257–295.

Wang, Z., Cheng, L., Li, Y. & Li, Z. (2020) Spatiotemporal characteristics of bike-sharing usage around rail transit stations: evidence from Beijing, China. *Sustainability*, 12, 1299.

Wang, Y., Finazzi, F. & Fassò, A. (2021) D-STEM v2: a software for modeling functional spatio-temporal data. *Journal of Statistical Software*, 99, 1–29.

Wickham, H., Hofmann, H., Wickham, C. & Cook, D. (2012) Glyph-maps for visually exploring temporal patterns in climate data and models. *Environmetrics*, 23, 382–393.

Willberg, E., Toivonen, T. & Salonen, M. (2019) Bike sharing as part of urban mobility in Helsinki – a user perspective. Master's thesis.

Yang, Z., Hu, J., Shu, Y., Cheng, P., Chen, J. & Moscibroda, T. (2016) Mobility modeling and prediction in bike-sharing systems. In: *Proceedings of the 14th annual international conference on mobile systems, applications, and services*, pp. 165–178.

Yang, H., Zhang, Y., Zhong, L., Zhang, X. & Ling, Z. (2020) Exploring spatial variation of bike sharing trip production and attraction: a study based on Chicago's Divvy system. *Applied Geography*, 115, 102130.

Zhang, Y. & Mi, Z. (2018) Environmental benefits of bike sharing: a big data-based analysis. *Applied Energy*, 220, 296–301.

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

How to cite this article: Piter, A., Otto, P. & Alkhatib, H. (2022) The Helsinki bike-sharing system—Insights gained from a spatiotemporal functional model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(3), 1294–1318. Available from: https://doi.org/10.1111/rssa.12834