Spatial kriging for replicated temporal point processes

Daniel Gervini
Department of Mathematical Sciences
University of Wisconsin–Milwaukee

June 15, 2022

Abstract

This paper presents a kriging method for spatial prediction of temporal intensity functions, for situations where a temporal point process is observed at different spatial locations. Assuming that several replications of the process are available at the spatial sites, this method avoids assumptions like isotropy, which are not valid in many applications. As part of the derivations, new nonparametric estimators for the mean and covariance functions of temporal point processes are introduced, and their properties are studied theoretically and by simulation. The method is applied to the analysis of bike demand patterns in the Divvy bicycle sharing system of the city of Chicago.

Key words: Cox process; Poisson process; spline smoothing; tensor-product splines.

1 Introduction

Spatiotemporal point processes have been part of mainstream probability theory for a long time (Cox and Isham, 1980), but their range of applications to everyday data analyses used to be rather limited compared to other statistical fields. The increasing complexity and widespread availability of massive datasets in the last decade, however, has made spatiotemporal point process modelling more broadly applicable. For example, the Chicago Data Portal (data.cityofchicago.org) maintains a data set of taxi trips made in the city of Chicago since 2013. For each of the approximately 200 million taxi trips included in the dataset, exact time and place of passenger pick up and drop off are available. This level of detail allows more complex spatiotemporal modelling and inference than was possible when such data was available only in aggregates.

The exact interplay between the temporal and spatial aspects of a point process model depends, of course, on the type of applications that the researcher has in mind. In this paper we study temporal point processes that occur at predefined spatial locations. Therefore, the random events are only time events and the spatial aspect of the problem is non-random. However, the spatial distribution of the locations determines the correlations among the temporal intensity functions of the point process, and therefore the ability to predict the process at a new spatial location, which is the main goal of this paper.

This work was motivated by the analysis of bike trips in the Divvy bicycle sharing system of the city of Chicago. In recent years, urban bicycle sharing systems have become widespread around the world (Shaheen et al., 2010). These systems allow customers to pick up and return bicycles from automated stations distributed throughout a city. Although the stations are automated, the maintenance of the system requires active human involvement. An important management issue is the control of flow imbalance (Nair and Miller-Hooks, 2011). The demand for bikes is different in different areas of the city at different times of the day. For example, during the morning commute hours most trips tend to flow from the outskirts of a city towards downtown areas, and the opposite occurs in the afternoon. Therefore, bike stations in some areas would quickly run out of bikes while stations in other areas would quickly fill up (thus running out of parking docks) if they were not manually rebalanced by trucking bikes from full stations to empty stations. These management operations, to run efficiently, require knowledge of the spatio-temporal patterns of bike demand in a city. Other important aspects of system management include decisions to open, shut down or relocate bike stations. Before opening a new bike station, managers need to be able to forecast bike demand at the proposed location. This is not simple, because the patterns of bike demand may vary greatly within short distances (Gervini and Khanal, 2019).

From a mathematical point of view, bike demand can be modelled by spatio-temporal point processes (Gervini and Khanal, 2019). This can be done in several ways, depending on the researchers' goals. In this paper, we model bike check-out times at each station as a replicated temporal point process, where each day of the year is a replication of the process. Therefore, for a bike station located at a spatial location $\mathbf{s}_j$, we will have n intensity functions $\{\lambda_i^j(t)\}_{i=1}^n$, one per day. The availability of replications allows us to estimate the spatial mean and covariance of the intensity functions nonparametrically, without resorting to assumptions like isotropy which are not valid in this context (Gervini and Khanal, 2019). These estimators, in turn, can be used to predict the intensity functions $\lambda_i^0(t)$'s at a new spatial site $\mathbf{s}_0$, assuming only spatial smoothness of the mean and covariance functions.

In the literature, the usual approach to spatial prediction is kriging (Cressie, 1993). Functional data methods have been used in connection with spatial point patterns (Comas et al., 2008; Delicado at al., 2010) and kriging has been extended to continuous functional data (Giraldo et al., 2010, 2011; Menafoglio et al., 2013) and spatial point processes (Gabriel et al., 2016). These methods could be applied in our context if the intensity functions at each site were estimable by smoothing, but this is not always possible because of low daily counts at some bike stations. More importantly, these methods were developed for situations where only one observation per site was available; this lack of replications made simplifying modelling assumptions like isotropy unavoidable in order to obtain useful inferential tools. The problem of modelling spatial anisotropy was addressed by Bernardi et al. (2018) in the context of continuous spatial processes, so their method does not automatically extend to our point-process scenario, but it does underline the need for methods addressing spatial anisotropy.

In this paper we present a kriging method for spatial prediction of temporal intensity functions that can be applied in anisotropic situations, as long as several replications of the processes are available at the spatial sites. As part of our derivations we introduce new nonparametric estimators for the mean and covariance functions of temporal point processes and study their properties.

2 Background

2.1 Poisson point processes

A temporal point process X is a random countable set in some $\mathscr{S} \subseteq (0, \infty)$ (Møller and Waagepetersen, 2004, ch. 2). The process is locally finite if $\#(X \cap B) < \infty$ with probability one for any bounded set $B \subseteq \mathscr{S}$, where $\#$ denotes the cardinality of a set. In that case we define the count function $N(B) = \#(X \cap B)$ for each bounded set $B \subseteq \mathscr{S}$. In particular, we define $N(t) = \#(X \cap (0, t])$. A Poisson process is a locally finite process for which there exists a locally integrable function $\lambda : \mathscr{S} \to [0, \infty)$, called the intensity function, such that *(i)* $N(B)$ has a Poisson distribution with rate $\int_B \lambda(t)\ dt$ for any bounded set $B \subseteq \mathscr{S}$, and *(ii)* for disjoint sets $B_1, \ldots, B_k$ in $\mathscr{S}$ the random variables $N(B_1), \ldots, N(B_k)$ are independent. A consequence of *(i)* and *(ii)* is that the conditional distribution of the points in $X \cap B$ given $N(B) = m$ is the distribution of m independent and identically distributed observations with density function $\lambda(\cdot)/\int_B \lambda$. In this paper we will mostly consider temporal processes defined on a common bounded interval $\mathscr{S} = [a, b]$, for example $\mathscr{S} = [0, 24]$ for daily processes.

For replicated point processes, a single intensity function λ rarely provides an adequate fit for all replications; it is more reasonable to assume that λ itself is the realization of a random process Λ and thus changes from replication to replication. Such compound processes are called doubly stochastic or Cox processes (Møller and Waagepetersen, 2004, ch. 5). A doubly stochastic Poisson process is a pair (X, Λ) where $X|\Lambda = \lambda$ is a Poisson process with intensity function λ, and Λ is a random function that takes values on the space $\mathscr{F}$ of non-negative locally integrable functions on $\mathscr{S}$. Thus the n daily replications of the process can be modeled as n independent and identically distributed pairs $(X_1, \Lambda_1), \ldots, (X_n, \Lambda_n)$, where the latent process Λ is not directly observable; only X is observed.

Consider now several doubly stochastic temporal processes that are observed at d different spatial locations. This situation can be modeled as a multivariate doubly stochastic process $(\mathbf{X}, \mathbf{\Lambda})$ with $\mathbf{X} = (X^1, \ldots, X^d)$ and $\mathbf{\Lambda} = (\Lambda^1, \ldots, \Lambda^d)$, where the

X^js are conditionally independent given $\mathbf{\Lambda} = \boldsymbol{\lambda}$. The dependencies among the X^js are then determined by the dependencies among the Λ^js. Since each Λ^j is associated with a specific spatial location $\mathbf{s}_j$, we make this explicit in the notation by writing $\Lambda^j(t) = \Lambda(t, \mathbf{s}_j)$, but note that $\Lambda(t, \mathbf{s})$ is not a joint spatio-temporal intensity function, it is just a temporal intensity function in the variable t for each $\mathbf{s}$. Our goal is to predict the temporal intensity process at a new spatial site $\mathbf{s}_0$, $\Lambda(t, \mathbf{s}_0)$, using the existing $\Lambda(t, \mathbf{s}_j)$s.

2.2 Spatial kriging

The unbiased kriging predictor of $\Lambda(t, \mathbf{s}_0)$ based on $\Lambda(t, \mathbf{s}_1), \ldots, \Lambda(t, \mathbf{s}_d)$ is

$$\Lambda^*(t, \mathbf{s}_0) = \sum_{j=1}^{d} c_j^* \Lambda(t, \mathbf{s}_j), \tag{1}$$

where $\mathbf{c}^* = (c_1^*, \ldots, c_d^*)$ minimizes the squared prediction error

$$\text{SPE}(\mathbf{c}) = E\{\|\Lambda(\cdot, \mathbf{s}_0) - \sum_{j=1}^{d} c_j \Lambda(\cdot, \mathbf{s}_j)\|^2\} \tag{2}$$

subject to the unbiasedness constraint

$$\mu(t, \mathbf{s}_0) = \sum_{j=1}^{d} c_j \mu(t, \mathbf{s}_j), \tag{3}$$

where $\mu(t, \mathbf{s}) = E\{\Lambda(t, \mathbf{s})\}$ and $\|\cdot\|$ is the L_2 norm.

The squared prediction error (2), in view of the constraints (3), comes down to

$$\text{SPE}(\mathbf{c}) = \mathbf{c}^T \mathbf{\Sigma} \mathbf{c} - 2\mathbf{c}^T \boldsymbol{\sigma}_0 + \sigma_{00}, \tag{4}$$

where $\mathbf{\Sigma}$ has elements

$$\Sigma_{jk} = \int \operatorname{cov}\left\{\Lambda(t, \mathbf{s}_j), \Lambda(t, \mathbf{s}_k)\right\} \, dt, \tag{5}$$

$\boldsymbol{\sigma}_0$ has elements

$$\sigma_{0j} = \int \operatorname{cov}\left\{\Lambda(t, \mathbf{s}_j), \Lambda(t, \mathbf{s}_0)\right\} \, dt, \tag{6}$$

and $\sigma_{00} = \int \mathrm{var}\,\{\Lambda(t, \mathbf{s}_0)\}\, dt$. By multiplying both sides of (3) by the $\mu(t, \mathbf{s}_k)$'s and integrating t out, the constraints can be expressed as

$$\mathbf{M}\mathbf{c} = \mathbf{m}_0, \tag{7}$$

where $\mathbf{M}$ has elements

$$M_{jk} = \int \mu(t, \mathbf{s}_j)\mu(t, \mathbf{s}_k)\ dt \tag{8}$$

and $\mathbf{m}_0$ has elements

$$m_{0j} = \int \mu(t, \mathbf{s}_j)\mu(t, \mathbf{s}_0)\ dt. \tag{9}$$

When $\mathbf{M}$ is full rank, the minimization of (4) subject to (7) has a closed-form solution

$$\begin{bmatrix} \mathbf{c}^* \\ \ell \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{M}^T \\ \mathbf{M} & \mathbf{O} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\sigma}_0 \\ \mathbf{m}_0 \end{bmatrix}, \tag{10}$$

where $\ell \in \mathbb{R}^d$ is the Lagrange multiplier and $\mathbf{O}$ is the $d \times d$ zero matrix. To compute $\mathbf{c}^*$, then, it is necessary to obtain estimators of $\boldsymbol{\Sigma}$, $\boldsymbol{\sigma}_0$, $\mathbf{M}$, and $\mathbf{m}_0$, which will be introduced in Section 3.

Note that although the kriging prediction problem (1) was framed in terms of the unobservable intensity processes $\Lambda^j(t)$'s, the estimated kriging coefficients $\hat{\mathbf{c}}^*$ can be used for direct real-time prediction of the count functions: for a given day i, the predicted count function at the new site $\mathbf{s}_0$ would be $N_i^{0*}(t) = \sum_{j=1}^d c_j^* N_i^j(t)$, where the $N_i^j(t)$'s are the observed count functions at the sites $\mathbf{s}_j$'s. This is an unbiased predictor, in view of (3), since $E\{N^0(t)\} = \int_a^t \mu(u, \mathbf{s}_0)du$ and $E\{N^j(t)\} = \int_a^t \mu(u, \mathbf{s}_j)du$.

3 Mean and covariance estimation

3.1 Nonparametric estimators at the observed sites

Estimation of the mean functions $\mu_j(t) = \mu(t, \mathbf{s}_j)$ and the covariance functions $\rho_{jk}(t, t') = \mathrm{cov}\,\{\Lambda(t, \mathbf{s}_j), \Lambda(t', \mathbf{s}_k)\}$ at the spatial points $\mathbf{s}_j$ and $\mathbf{s}_k$ where data is available can be done as follows. Since $X^j \mid \Lambda^j = \lambda^j$ is a Poisson process with intensity function $\lambda^j(t)$, for any integrable functions $f(t)$ and $g(t)$ we have, as shown in the

Supplementary Material,

$$E\left\{\sum_{t\in X^j} f(t)\right\} = \int f(t)\mu_j(t)\ dt, \tag{11}$$

$$E\left\{\sum_{t\in X^j}\sum_{t'\in X^k} f(t)g(t')\right\} = \iint f(t)g(t')R_{jk}(t,t')\ dt\ dt',\ \text{ for } j\neq k, \tag{12}$$

and

$$E\left\{\sum_{t\in X^j}\sum_{t'\in X^j, t'\neq t} f(t)g(t')\right\} = \iint f(t)g(t')R_{jj}(t,t')\ dt\ dt', \tag{13}$$

where $R_{jk}(t,t') = E\left\{\Lambda(t,\mathbf{s}_j)\Lambda(t',\mathbf{s}_k)\right\}$. These identities suggest the following nonparametric estimators of $\mu_j(t)$ and $R_{jk}(t,t')$. Consider a B-spline basis (De Boor, 2001, ch. 9) $\boldsymbol{\beta}(t) = (\beta_1(t),\ldots,\beta_p(t))^T$ on $[a,b]$, and let $\mathbf{G} = \int \boldsymbol{\beta}(t)\boldsymbol{\beta}(t)^T\ dt$. Then, given independent and identically distributed replications $\mathbf{X}_1,\ldots,\mathbf{X}_n$ of the multivariate process $\mathbf{X}$, define

$$\hat{\mu}_j(t) = \boldsymbol{\beta}(t)^T\mathbf{G}^{-1}\frac{1}{n}\sum_{i=1}^{n}\sum_{u\in X_i^j}\boldsymbol{\beta}(u), \tag{14}$$

$$\hat{R}_{jk}(t,t') = \boldsymbol{\beta}(t)^T\mathbf{G}^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{u\in X_i^j}\sum_{v\in X_i^k}\boldsymbol{\beta}(u)\boldsymbol{\beta}(v)^T\right\}\mathbf{G}^{-1}\boldsymbol{\beta}(t'),\ \text{for } j\neq k, \tag{15}$$

and

$$\hat{R}_{jj}(t,t') = \boldsymbol{\beta}(t)^T\mathbf{G}^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{u\in X_i^j}\sum_{v\in X_i^j, v\neq u}\boldsymbol{\beta}(u)\boldsymbol{\beta}(v)^T\right\}\mathbf{G}^{-1}\boldsymbol{\beta}(t'). \tag{16}$$

The consistency of these estimators as the number of replications n goes to infinity is proved in Section 4, and confirmed by simulations in Section 5.

From the above $\hat{\mu}_j(t)$'s and $\hat{R}_{jk}(t,t')$'s we obtain $\hat{\rho}_{jk}(t,t') = \hat{R}_{jk}(t,t') - \hat{\mu}_j(t)\hat{\mu}_k(t')$. These are plugged into equations (8) and (5) to obtain $\hat{\mathbf{M}}$ and $\hat{\boldsymbol{\Sigma}}$ respectively. These, in turn, as plugged into (10) to obtain estimators of the kriging coefficients $\hat{\mathbf{c}}^*$. Note, however, that we have not yet explained how to obtain estimators of $\mathbf{m}_0$ and $\boldsymbol{\sigma}_0$; they will be introduced in the next section.

It is often the case that $\hat{\boldsymbol{\Sigma}}$, although of full rank, is ill-conditioned. We have found in simulation studies (not reported here) that truncating $\hat{\mathbf{\Sigma}}$ improves kriging accuracy. Let $\hat{\mathbf{\Sigma}} = \mathbf{V}\mathbf{H}\mathbf{V}^T$ be the spectral decomposition of $\hat{\mathbf{\Sigma}}$, where $\mathbf{H} = \mathrm{diag}(\eta_1, \ldots, \eta_d)$ are the eigenvalues of $\hat{\mathbf{\Sigma}}$ in decreasing order and $\mathbf{V}$ is the orthogonal matrix of eigenvectors. Take the smallest s such that $\sum_{j=1}^{s} \eta_j / \sum_{j=1}^{d} \eta_j \geq 0.9$, say. Let $\mathbf{H}_s = \mathrm{diag}(\eta_1, \ldots, \eta_s)$ and let $\mathbf{V}_s$ be the first s columns of $\mathbf{V}$. Then we solve

$$\begin{bmatrix} \hat{\mathbf{c}}_s \\ \hat{\ell} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_s & \mathbf{V}_s^T \hat{\mathbf{M}}^T \\ \hat{\mathbf{M}}\mathbf{V}_s & \mathbf{O} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{V}_s^T \hat{\boldsymbol{\sigma}}_0 \\ \hat{\mathbf{m}}_0 \end{bmatrix} \tag{17}$$

and take $\hat{\mathbf{c}}^* = \mathbf{V}_s \hat{\mathbf{c}}_s$.

Similarly, the $d \times d$ matrix $\mathbf{M}$ in (7) is often not of full rank. For example, if $\mu(t, \mathbf{s}_j) \equiv \mu(t)$ for all $\mathbf{s}_j$, then $\mathbf{M}$ has rank one. Even when it has full rank, $\mathbf{M}$ is often ill-conditioned, with many eigenvalues close zero. In such situations $\mathbf{M}$ (and its estimator $\hat{\mathbf{M}}$) can also be truncated: let $\mathbf{M} = \mathbf{U}\boldsymbol{\Delta}\mathbf{U}^T$ be the spectral decomposition of $\mathbf{M}$, where $\boldsymbol{\Delta} = \mathrm{diag}(\delta_1, \ldots, \delta_d)$ are the eigenvalues in decreasing order and $\mathbf{U}$ is the orthogonal matrix of eigenvectors. Take the smallest r such that $\sum_{j=1}^{r} \delta_j / \sum_{j=1}^{d} \delta_j \geq 0.9$. Then the d-dimensional constraints (7) are replaced by the r-dimensional approximation

$$\tilde{\mathbf{M}}\mathbf{c} = \tilde{\mathbf{m}}_0, \tag{18}$$

where $\tilde{\mathbf{M}} = \boldsymbol{\Delta}_r \mathbf{U}_r^T$, $\tilde{\mathbf{m}}_0 = \mathbf{U}_r^T \mathbf{m}_0$, $\boldsymbol{\Delta}_r = \mathrm{diag}(\delta_1, \ldots, \delta_r)$, and $\mathbf{U}_r$ are the first r columns of $\mathbf{U}$. The corresponding substitutions are made in equations (10) and (17).

3.2 Estimators at the new site

To estimate $\boldsymbol{\sigma}_0$ and $\mathbf{m}_0$, the mean and covariance estimators defined above are extended, by smoothing, to spatial points $\mathbf{s}_0$ where no data is available. Consider first the mean function $\mu(t, \mathbf{s})$. This function can be modelled as $\boldsymbol{\beta}(t)^T \mathbf{B} \boldsymbol{\gamma}(\mathbf{s})$, where $\boldsymbol{\gamma}(\mathbf{s}) = (\boldsymbol{\gamma}_1(\mathbf{s}), \ldots, \boldsymbol{\gamma}_q(\mathbf{s}))^T$ is a spatial basis on a region $R \subset \mathbb{R}^2$ that includes $\mathbf{s}_0$ and the $\mathbf{s}_k$'s, and $\mathbf{B}$ is a coefficient matrix. In this paper we use tensor-product splines as $\boldsymbol{\gamma}(\mathbf{s})$ (De Boor, 2001, ch. 17), but other alternatives are possible, such as thin-plate splines (Wahba, 1990) or radial basis functions (Buhmann, 2003).

To estimate $\mathbf{B}$, note that $\hat{\mu}_j(t)$ in (14) has the form $\hat{\mu}_j(t) = \boldsymbol{\beta}(t)^T \hat{\mathbf{a}}_j$, with

$$\hat{\mathbf{a}}_j = \mathbf{G}^{-1} \frac{1}{n} \sum_{i=1}^{n} \sum_{u \in X_i^j} \boldsymbol{\beta}(u),$$

so the penalized least squares estimator of $\mathbf{B}$ would be

$$\hat{\mathbf{B}} = \arg\min_{\mathbf{B}} \sum_{j=1}^{d} \|\hat{\mathbf{a}}_j - \mathbf{B}\boldsymbol{\gamma}(\mathbf{s}_j)\|^2 + \xi_B P_1(\mathbf{B}),$$

where $P_1(\mathbf{B})$ is a roughness penalty on the function $\mathbf{B}\boldsymbol{\gamma}(\mathbf{s})$ and ξ_B is a smoothing parameter. As explained in the Supplementary Material, the roughness of a bivariate function $f(s^1, s^2)$ can be measured by $\iint (\sum_{1 \le i,j \le 2} f_{ij}^2) ds^1 ds^2$, where $f_{ij} = \partial^2 f / \partial s^i \partial s^j$, and then $P_1(\mathbf{B}) = \operatorname{tr}\left(\mathbf{B}^T \mathbf{B} \mathbf{J}\right)$, where $\mathbf{J}$ is a matrix that depends only on $\boldsymbol{\gamma}(\mathbf{s})$. The closed form of $\hat{\mathbf{B}}$ is then

$$\hat{\mathbf{B}} = \mathbf{A}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} + \xi_B \mathbf{J})^{-1}, \tag{19}$$

where $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}(\mathbf{s}_1), \ldots, \boldsymbol{\gamma}(\mathbf{s}_d)]^T$ and $\mathbf{A} = [\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_d]$. Once $\hat{\mathbf{B}}$ is obtained, $\mu(t, \mathbf{s}_0)$ is estimated by $\hat{\mu}(t, \mathbf{s}_0) = \boldsymbol{\beta}(t)^T \hat{\mathbf{B}} \boldsymbol{\gamma}(\mathbf{s}_0)$ and plugged into (9) to obtain $\hat{\mathbf{m}}_0$.

The optimal smoothing parameter ξ_B can be chosen by cross-validation (Hastie et al., 2009, ch. 7). The leave-one-site-out cross-validation statistic would be

$$\begin{aligned} \mathrm{CV}(\xi_B) &= \frac{1}{d} \sum_{j=1}^{d} \|\hat{\mathbf{a}}_j - \hat{\mathbf{B}}_{(j)} \boldsymbol{\gamma}(\mathbf{s}_j)\|^2 \\ &= \frac{1}{d} \sum_{j=1}^{d} \frac{\|\hat{\mathbf{a}}_j - \hat{\mathbf{B}} \boldsymbol{\gamma}(\mathbf{s}_j)\|^2}{(1 - h_{B,jj})^2}, \end{aligned}$$

where $\hat{\mathbf{B}}_{(j)}$ is the $\mathbf{s}_j$-deleted version of $\hat{\mathbf{B}}$, and $h_{B,jj}$ is the jth diagonal element of the hat matrix $\mathbf{H}_B = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} + \xi_B \mathbf{J})^{-1}\boldsymbol{\Gamma}^T$. If $\mathrm{df}_B = \operatorname{tr}(\mathbf{H}_B)$, the degrees of freedom of the fit, then $h_{B,jj} \approx \mathrm{df}_B / d$ and the generalized cross-validation statistic is

$$\mathrm{GCV}(\xi_B) = \frac{1}{d} \sum_{j=1}^{d} \frac{\|\hat{\mathbf{a}}_j - \hat{\mathbf{B}} \boldsymbol{\gamma}(\mathbf{s}_j)\|^2}{(1 - \mathrm{df}_B / d)^2}.$$

The optimal $\hat{\xi}_B$ is the minimizer of $\mathrm{GCV}(\xi_B)$.

To estimate $\boldsymbol{\sigma}_0$ we also use spatial smoothing, modeling $\Sigma(\mathbf{s},\mathbf{s}') = \int \mathrm{cov}\left\{\Lambda(t,\mathbf{s}),\Lambda(t,\mathbf{s}')\right\} dt$ by $\boldsymbol{\gamma}(\mathbf{s})^T\mathbf{C}\boldsymbol{\gamma}(\mathbf{s}')$, with $\mathbf{C}$ symmetric. The penalized least squares estimator of $\mathbf{C}$ is

$$\hat{\mathbf{C}} = \arg\min_{\mathbf{C}} \sum_{j=1}^{d} \sum_{\substack{k=1\\k\neq j}}^{d} \left\{\hat{\Sigma}_{jk} - \boldsymbol{\gamma}(\mathbf{s}_j)^T\mathbf{C}\boldsymbol{\gamma}(\mathbf{s}_k)\right\}^2 + \xi_C P_2(\mathbf{C}), \tag{20}$$

where, as before, $P_2(\mathbf{C})$ is a roughness penalty on $\boldsymbol{\gamma}(\mathbf{s})^T\mathbf{C}\boldsymbol{\gamma}(\mathbf{s}')$ and ξ_C is a smoothing parameter. Note that we only use the off-diagonal elements of $\hat{\boldsymbol{\Sigma}}$ in (20) because, in most applications, there are intrinsic sources of variability at each spatial site that create a ridge along the diagonal $\mathbf{s} = \mathbf{s}'$ on the function $\Sigma(\mathbf{s},\mathbf{s}')$, making it discontinuous there (a plausible probabilistic model for this effect is given by equation (23) in Section 5). As before, the roughness of a function $f(s^1, s^2, s^3, s^4)$ can be measured by $\iint(\sum_{1\leq i,j,k,l\leq 2} f_{ijkl}^2)ds^1ds^2ds^3ds^4$, and then $P_2(\mathbf{C}) = \mathrm{tr}\{(\mathbf{CJ})^2\}$ with $\mathbf{J}$ as before. As shown in the Supplementary Material, the closed form for $\mathrm{vec}(\hat{\mathbf{C}})$ is

$$\mathrm{vec}(\hat{\mathbf{C}}) = \boldsymbol{\Omega}^{-1}(\boldsymbol{\Gamma}^T \otimes \boldsymbol{\Gamma}^T)\,\mathrm{vec}(\hat{\boldsymbol{\Sigma}} - \mathrm{diag}\,\hat{\boldsymbol{\Sigma}}), \tag{21}$$

where $\boldsymbol{\Omega} = \left\{(\boldsymbol{\Gamma}^T \otimes \boldsymbol{\Gamma}^T)(\mathbf{I} - \mathbf{E}^T\mathbf{E})(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}) + \xi_C(\mathbf{J} \otimes \mathbf{J})\right\}$ and $\mathbf{E}^T\mathbf{E} = \sum_{j=1}^{d} \mathbf{e}_j\mathbf{e}_j^T \otimes \mathbf{e}_j\mathbf{e}_j^T$ with $\mathbf{e}_j$ the j-th canonical vector in $\mathbb{R}^d$. Once $\hat{\mathbf{C}}$ has been obtained, the σ_{0j}'s in (6) are estimated by $\hat{\sigma}_{0j} = \boldsymbol{\gamma}(\mathbf{s}_j)^T\hat{\mathbf{C}}\boldsymbol{\gamma}(\mathbf{s}_0)$. Details of numerical implementation are important and discussed in the Supplementary Material, because the large dimensions of $\boldsymbol{\Omega}$ make straight computation of (21) very inefficient and time consuming.

The optimal smoothing parameter ξ_C can be found, as before, by generalized cross-validation. The leave-one-(j,k)-out cross-validation statistic is

$$\begin{aligned}\mathrm{CV}(\xi_C) &= \frac{1}{d(d-1)} \sum_{j=1}^{d} \sum_{\substack{k=1\\k\neq j}}^{d} \left\{\hat{\Sigma}_{jk} - \boldsymbol{\gamma}(\mathbf{s}_j)^T\hat{\mathbf{C}}_{(j,k)}\boldsymbol{\gamma}(\mathbf{s}_k)\right\}^2 \\ &= \frac{1}{d(d-1)} \sum_{j=1}^{d} \sum_{\substack{k=1\\k\neq j}}^{d} \frac{\left\{\hat{\Sigma}_{jk} - \boldsymbol{\gamma}(\mathbf{s}_j)^T\hat{\mathbf{C}}\boldsymbol{\gamma}(\mathbf{s}_k)\right\}^2}{(1 - h_{C,(j,k)})^2},\end{aligned}$$

where $\hat{\mathbf{C}}_{(j,k)}$ is the (j,k)-deleted version of $\hat{\mathbf{C}}$ and $h_{C,(j,k)}$ is the diagonal element of the hat matrix $\mathbf{H}_C = (\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\boldsymbol{\Omega}^{-1}(\boldsymbol{\Gamma}^T \otimes \boldsymbol{\Gamma}^T)$ corresponding to the location of $\hat{\Sigma}_{jk}$ in vec($\hat{\mathbf{\Sigma}}$). As before, if $\mathrm{df}_C = \mathrm{tr}(\mathbf{H}_C)$ then $h_{C,(j,k)} \approx \mathrm{df}_C/d(d-1)$ and the generalized cross-validation statistic is

$$\mathrm{GCV}(\xi_C) = \frac{1}{d(d-1)} \sum_{j=1}^{d} \sum_{\substack{k=1 \\ k \neq j}}^{d} \frac{\left\{\hat{\Sigma}_{jk} - \boldsymbol{\gamma}(\mathbf{s}_j)^T \hat{\mathbf{C}} \boldsymbol{\gamma}(\mathbf{s}_k)\right\}^2}{\{1 - \mathrm{df}_C/d(d-1)\}^2}.$$

The optimal $\hat{\xi}_C$ is the minimizer of $\mathrm{GCV}(\xi_C)$.

Programs implementing these estimators, written in Matlab language, are available from the author's website.

4 Asymptotics

In this section we establish the consistency of the nonparametric estimators introduced in Section 3.1. The convergence rates obtained are in line with the standard asymptotic results for regression splines (Agarwal and Studden, 1980; Zhou et al., 1998).

Let $\boldsymbol{\beta}(t)$ be a B-spline basis of order r defined by a regular knot sequence $\{\tau_1, \ldots, \tau_k\}$, that is,

$$\int_a^{\tau_i} g(t)\ dt = \frac{i}{k+1}, \quad i = 1, \ldots, k,$$

for a strictly positive density function $g(t)$ on $[a, b]$. The basis dimension is then $p = r + k$. The observed point processes $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are assumed to be independent and identically distributed replications of a d-variate doubly stochastic Poisson process $\mathbf{X}$ with latent intensity process $\mathbf{\Lambda}$, as explained in Section 2.1. The norm $\|\cdot\|$ below is the standard $L_2[a, b]$ norm, and $L_2^r[a, b]$ is the Sobolev space of functions f such that $D^{r-1}f$ is absolutely continuous on $[a, b]$ and $D^r f \in L_2[a, b]$, where D denotes differentiation. Proofs of the results in this section are given in the Supplementary Material.

Theorem 1 *Let $\hat{\mu}_j(t)$ be the estimator defined in (14), and assume $\mu_j \in L_2^r[a, b]$. Then*

$$E\|\hat{\mu}_j - \mu_j\|^2 = \frac{1}{n} O(k) + O\left(\frac{1}{k^{2r}}\right).$$

The fastest convergence rate is attained for $k = O\left(n^{1/(2r+1)}\right)$, in which case $E\|\hat{\mu}_j - \mu_j\|^2 = O\left(n^{-2r/(2r+1)}\right)$.

Theorem 1 shows that the optimal nonparametric convergence rate $O\left(n^{-2r/(2r+1)}\right)$ for functions in $L_2^r[a,b]$ (Stone, 1982) can be attained by $\hat{\mu}_j(t)$ if k is appropriately chosen. For example, for cubic splines, $r = 4$ and the optimal rates are $k = O\left(n^{1/9}\right)$ and $E\|\hat{\mu}_j - \mu_j\|^2 = O\left(n^{-8/9}\right)$. The number of knots k, then, should grow slowly with n; for example, $400^{1/9} \approx 2$.

The next theorem gives convergence rates for the $\hat{R}_{jk}(t,t')$'s. Now $\|\cdot\|$ is the $L_2([a,b] \times [a,b])$ norm and $L_2^{(r,r)}([a,b] \times [a,b])$ is the tensor Sobolev space of bivariate functions f such that $D_i^{r-1} f$ is absolutely continuous on $[a,b] \times [a,b]$ and $D_i^r f \in L_2([a,b] \times [a,b])$, where D_i denotes differentiation with respect to the i-th variable.

Theorem 2 *Let $\hat{R}_{jk}(t,t')$ be the estimator defined in (15), if $j \neq k$, or in (16), if $j = k$. Then, if $R_{jk} \in L_2^{(r,r)}([a,b] \times [a,b])$, we have*

$$E\|\hat{R}_{jk} - R_{jk}\|^2 = \frac{1}{n} O(k^2) + O\left(\frac{1}{k^{2r}}\right).$$

The fastest convergence rate is attained for $k = O\left(n^{1/(2r+2)}\right)$, in which case $E\|\hat{R}_{jk} - R_{jk}\|^2 = O\left(n^{-2r/(2r+2)}\right)$.

Once again, Theorem 2 shows that the optimal convergence rate $O\left(n^{-2r/(2r+2)}\right)$ for bivariate functions (Stone, 1994) is attained by $\hat{R}_{jk}(t,t')$ if k is suitably chosen. For cubic splines, the optimal k would have rate $O(n^{1/10})$ and the squared estimation error would have rate $O(n^{-8/10})$. According to Theorems 1 and 2, the optimal rates for k are different for the $\hat{\mu}_j(t)$'s and the $\hat{R}_{jk}(t,t')$'s. However, for simplicity we use the same spline basis $\boldsymbol{\beta}(t)$ in both cases.

5 Simulations

In this section we study the consistency and convergence rates of the proposed estimators by simulation. Specifically, we are interested in the effects of the sample size n, the spatial grid size d, and the grid spacing $\delta = \min_{j \neq k} \|\mathbf{s}_j - \mathbf{s}_k\|$ on estimation and prediction errors.

We simulated the following scenarios. Three spatial grids were considered: *(i)* $d = 16$ uniformly spaced $\mathbf{s}_j$'s on the square $[-0.5, 0.5] \times [-0.5, 0.5]$, *(ii)* $d = 16$ uniformly spaced $\mathbf{s}_j$'s on the square $[-0.2, 0.2] \times [-0.2, 0.2]$, and *(iii)* $d = 64$ uniformly spaced $\mathbf{s}_j$'s on the square $[-0.5, 0.5] \times [-0.5, 0.5]$. The respective grid spacings were *(i)* $\delta = 0.33$, *(ii)* $\delta = 0.13$, and *(iii)* $\delta = 0.14$. Grids *(i)* and *(iii)* cover the same range but a larger d makes *(iii)* denser, while grids *(i)* and *(ii)* have the same size d but *(ii)* is denser because it covers a smaller range. The kriging predictor was evaluated at the spatial point $\mathbf{s}_0 = (0, 0)$.

The latent processes $\Lambda(t, \mathbf{s}_j)$'s were generated according to the log-Gaussian model

$$\Lambda(t, \mathbf{s}_j) = \exp\{\nu(t) + U_j \phi(t)\} \tag{22}$$

for $t \in [0, 1]$, with $\nu(t) = \sin(\pi t) + \ln 20$ and $\phi(t) = \sqrt{2} \sin(\pi t)$. The U_j's, which determine the spatial correlations, were generated as

$$U_j = g(\mathbf{s}_j)W + E_j \tag{23}$$

with $W \sim N(0, 0.072)$, $E_j \sim N(0, 0.018)$ and E_js independent among themselves and of W. Two functions $g(\mathbf{s})$ were considered: Model 1, $g(\mathbf{s}) = 1/(1 + \|\mathbf{s}\|)$, and Model 2, $g(\mathbf{s}) = 1$. The common factor $g(\mathbf{s})W$ in (23) makes $\text{cov}\{\Lambda(t, \mathbf{s}), \Lambda(t, \mathbf{s}')\}$ a smooth function for $\mathbf{s} \neq \mathbf{s}'$, but the E_js create a ridge at $\mathbf{s} = \mathbf{s}'$, as often observed in practice. Explicit expressions for $\mu(t, \mathbf{s}_j)$ and $\text{cov}\{\Lambda(t, \mathbf{s}_j), \Lambda(t, \mathbf{s}_k)\}$ are given in the Supplementary Material. Four sample sizes n were considered: 50, 100, 200, and 400. This makes a total of 24 simulated scenarios. Each scenario was replicated 400 times.

For estimation, we used cubic B-splines with five equally-spaced knots as temporal basis $\boldsymbol{\beta}(t)$, and tensor-product cubic B-splines with six equally-spaced knots on each coordinate as spatial basis $\boldsymbol{\gamma}(\mathbf{s})$. The respective basis dimensions were $p = 9$ and $q = 100$. The optimal smoothing parameters ξ_B and ξ_C were chosen by generalized cross-validation, as explained in Section 3.

We are primarily interested in estimation of the quantities $\mathbf{M}$ and $\mathbf{m}_0$ in (7), of $\boldsymbol{\Sigma}$ in (5), and of $\boldsymbol{\sigma}_0$ in (6), because they are needed for kriging. For $\hat{\mathbf{M}}$ we define the relative error measures: $\text{bias}(\hat{\mathbf{M}}) = \|E \,\text{vech}\, \hat{\mathbf{M}} - \text{vech}\, \mathbf{M}\| / \|\text{vech}\, \mathbf{M}\|$, $\text{sd}(\hat{\mathbf{M}}) = \{E\| \,\text{vech}\, \hat{\mathbf{M}} - E \,\text{vech}\, \hat{\mathbf{M}}\|^2\}^{1/2} / \|\text{vech}\, \mathbf{M}\|$, and $\text{rmse}(\hat{\mathbf{M}}) = \{E\| \,\text{vech}\, \hat{\mathbf{M}} - \text{vech}\, \mathbf{M}\|^2\}^{1/2} / \|\text{vech}\, \mathbf{M}\|$, where vech denotes the vectorization of the lower triangular

		Model 1					Model 2				
Grid	n	$\mathbf{M}$	$\mathbf{m}_0$	$\mathbf{\Sigma}$	$\boldsymbol{\sigma}_0$	SPE	$\mathbf{M}$	$\mathbf{m}_0$	$\mathbf{\Sigma}$	$\boldsymbol{\sigma}_0$	SPE
(i)	50	.077	.074	.41	.43	.80	.110	.105	.41	.42	.48
	100	.057	.056	.29	.40	.51	.070	.066	.28	.34	.27
	200	.042	.043	.22	.34	.51	.048	.045	.19	.29	.41
	400	.027	.031	.14	.35	.61	.041	.040	.12	.27	.61
(ii)	50	.089	.080	.35	.28	.07	.102	.097	.43	.39	.05
	100	.065	.060	.27	.22	.05	.073	.070	.25	.23	.03
	200	.043	.041	.18	.17	.04	.048	.045	.19	.17	.02
	400	.032	.030	.13	.13	.03	.039	.037	.13	.12	.02
(iii)	50	.076	.068	.38	.27	.23	.110	.105	.38	.36	.18
	100	.054	.054	.29	.22	.21	.071	.067	.26	.23	.13
	200	.036	.042	.20	.18	.19	.053	.050	.18	.15	.09
	400	.027	.033	.15	.16	.17	.033	.031	.13	.11	.06

Table 1: Simulation Results. Relative root mean squared errors of parameter estimators.

part of a matrix and $\|\cdot\|$ the usual Euclidean norm. Analogous measures are defined for $\hat{\mathbf{m}}_0$, vech $\hat{\mathbf{\Sigma}}$, and $\hat{\boldsymbol{\sigma}}_0$. To assess the accuracy of the kriging predictor, we compared the best SPE (2) attained by the true parameters, SPE_0, with the SPE attained by the estimators, $\widehat{\text{SPE}}$. Since $\widehat{\text{SPE}} \geq \text{SPE}_0$, $\text{bias}(\widehat{\text{SPE}}) = \text{rmse}(\widehat{\text{SPE}})$. The rmse's for all parameters are reported in Table 1. Biases and standard deviations can be found in the Supplementary Material.

We see in Table 1 that the mean and covariance estimators $\hat{\mathbf{M}}$ and $\hat{\mathbf{\Sigma}}$ are consistent as n increases, and the magnitudes of the errors do not depend on the grid, as expected. The estimation errors of $\hat{\mathbf{m}}_0$ do not depend on the grids either, because, for these models, the mean functions $\mu(t, \mathbf{s}_j)$ are nearly identical for all $\mathbf{s}_j$'s. However, the situation is different for $\hat{\boldsymbol{\sigma}}_0$ because the true covariance function does vary with $\mathbf{s}$. The errors are larger for the sparser grid *(i)* and smaller for grids *(ii)* and *(iii)*, being of comparable size for the latter two. The behavior of $\hat{\boldsymbol{\sigma}}_0$, then, fundamentally depends on grid spacing, not on grid size. The accuracy of the kriging predictor, on the other hand, is higher under grid *(ii)* than under grid *(iii)*, showing that under comparable grid spacings a smaller and more parsimonious grid is preferable. However, this behavior is model dependent: for Model 2, where $g(\mathbf{s})$ does not decrease away from $\mathbf{s}_0$, spatial sites further away from $\mathbf{s}_0$ contribute more to prediction at $\mathbf{s}_0$ than they do under Model 1, so the SPE errors for grids *(ii)* and *(iii)* are not as different from one another under Model 2 as they are under Model 1.

6 Application: predicting bike demand

As an example of application, we report an analysis of bicycle usage data from the Divvy bicycle-sharing system of the city of Chicago. The data is publicly available at the Chicago Data Portal (data.cityofchicago.org). We analyzed bike trips that took place on working days of 2016, that is, weekdays that were not holidays, in the downtown area known as 'the Loop', which is delimited by Grand, Roosevelt, and Halsted Avenues, and the lake front on the east. There were 68 active bike stations in that area during that period. We studied bike check-out times, which can be modelled as replicated Poisson processes. There were 254 working days in 2016, which constituted the n replications of the processes.

We set aside two of the 68 bike stations for prediction: the one at the Union train station, on Adams and Canal Streets, and the one at Lasalle Avenue and Calhoun Street. These two bike stations exhibited very different usage patterns: the one at the Union station showed a bimodal pattern of demand, with peaks at 8am and 5pm corresponding to the morning and evening work commutes respectively, whereas the one at Lasalle showed a unimodal pattern with peaks at 5pm only. The respective daily count functions are shown in Figures 1(a) and 1(b). Kriging prediction for these two stations was then based on the remaining $d = 66$ stations.

For estimation we used cubic B-splines with five equally-spaced knots as temporal basis $\boldsymbol{\beta}(t)$ and tensor-product cubic B-splines with six equally-spaced knots on each coordinate as spatial basis $\boldsymbol{\gamma}(\mathbf{s})$; as suggested by the asymptotic results (Section 4), we did not use too many knots. The optimal smoothing parameters ξ_B and ξ_C were chosen by generalized cross-validation. The truncated versions of the estimators (17) were used, using 0.9 as threshold for the cumulative eigenvalue proportion. To assess prediction accuracy, we compared the observed count functions $N_i(t)$ with the predicted count functions $\hat{N}_i(t)$ and computed the root average squared error $\{\sum_{i=1}^{n} \|N_i - \hat{N}_i\|^2/n\}^{1/2}$ for each station (here $\|\cdot\|$ denotes the L^2 norm). We obtained an error of 184.2 for the Union station and 70.9 for the Lasalle station.

It is not surprising that the Union station was harder to predict. This bike station

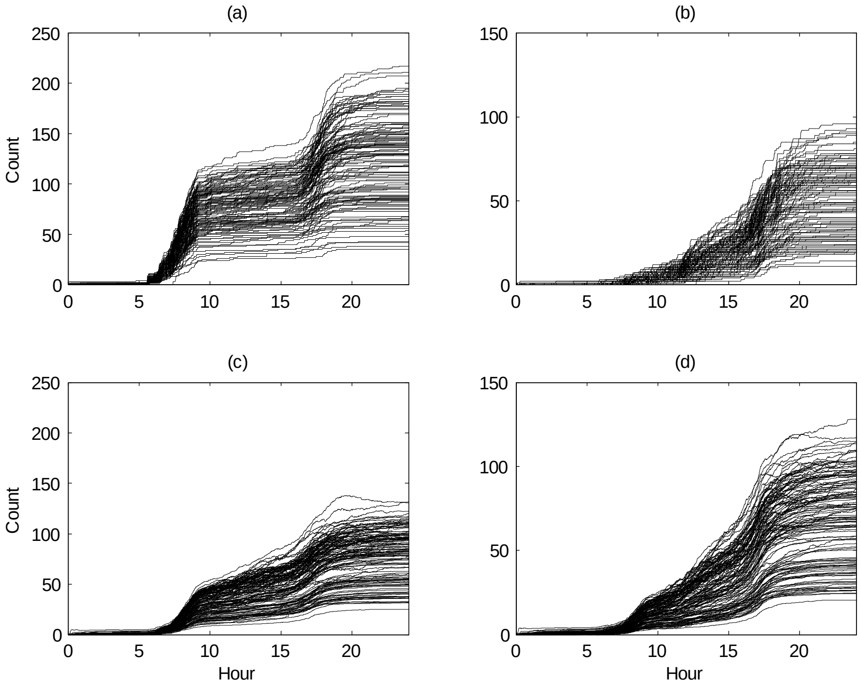

Figure 1: Divvy data analysis. Observed daily count functions for Union station [(a)] and Lasalle station [(b)], and respective predicted count functions for Union station [(c)] and Lasalle station [(d)].

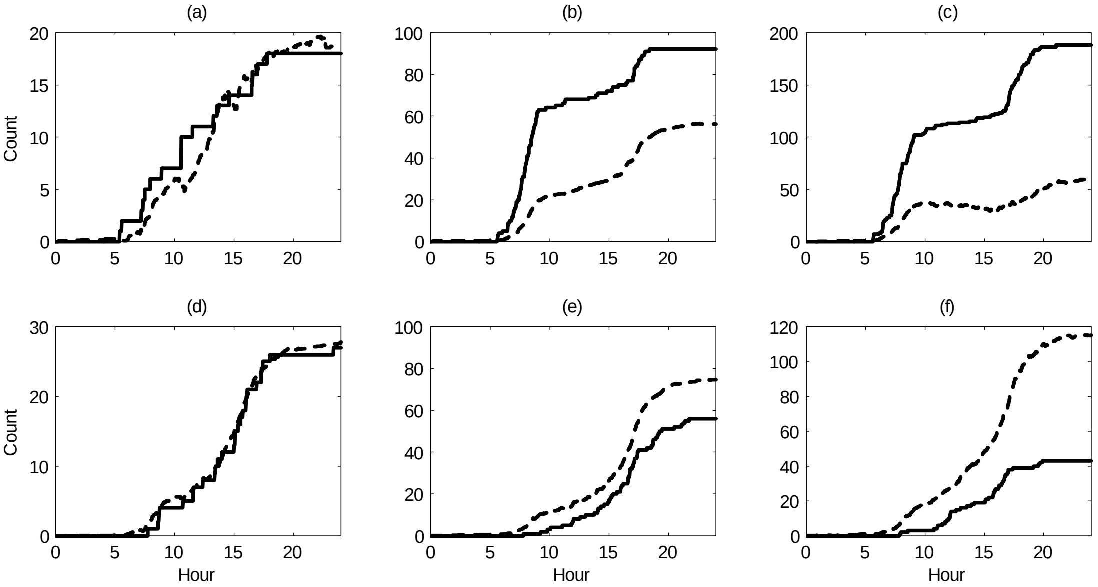

Figure 2: Divvy data analysis. Best fit [(a)], median fit [(b)] and worst fit [(c)] for Union station, and, respectively, best fit [(d)], median fit [(e)] and worst fit [(f)] for Lasalle station. (——) are the observed count functions and (- - -) are predicted count functions.

was situated at a train station and therefore had a peculiar pattern of demand, which was not shared by most other locations in the Loop. In contrast, the Lasalle station followed a more common pattern of demand. For instance, the average daily count for all Loop stations was 56.9, similar to the average daily count of 52.7 for Lasalle station, while the average daily count for the Union station was a much higher 124.4. This sharp change in $\mu(t, \mathbf{s})$ at the Union station makes $\mu(t, \mathbf{s}_0)$ hard to estimate accurately from neighboring spatial sites.

The daily predicted counts for these two stations are shown in Figures 1(c) and 1(d). We see that the predictors capture the overall patterns of demand at both stations, but the morning-commute peak is underestimated at the Union station. Figure 2 shows the best, median, and worst fits for each station. We see that the kriging predictor tends to underestimate the counts at the Union station and to overestimate them for the Lasalle station, but overall, prediction is reasonably accurate for the latter.

This example highlights both the possibilities and the limitations of spatial kriging. Prediction accuracy depends in part on the intrinsic variability at each station, which is independent of other stations and therefore cannot be predicted, and also on the degree of smoothness of the mean function $\mu(t, \mathbf{s})$ and the covariance functions $\Sigma(\mathbf{s}_j, \mathbf{s})$ at $\mathbf{s} = \mathbf{s}_0$. As long as the intrinsic variability is relatively low and there are no sharp peaks or troughs in $\mu(t, \mathbf{s})$ and $\Sigma(\mathbf{s}_j, \mathbf{s})$ at $\mathbf{s} = \mathbf{s}_0$, prediction will be reasonably accurate. But local landmarks, like train stations, theaters or stadiums, introduce spatial discontinuities that make prediction inaccurate for estimators that are solely based on spatial smoothing. We suspect that prediction in such situations can be improved by incorporating proximity to landmarks as covariates in the model, but that is a topic for future research.

References

Agarwal, G.G., and Studden, W.J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *The Annals of Statistics* **8** 1307–1325.

Bernardi, M.S., Carey, M., Sangalli, L.M., and Ramsay, J.O. (2018). Modeling spatial anisotropy via regression with partial differential regularization. *Journal of Multivariate Analysis* **167** 15–30.

Buhmann, M.D. (2003). *Radial Basis Functions : Theory and Implementations.* Cambridge University Press, Cambridge, UK.

Comas, C., Delicado, P., and Mateu, J. (2008). Analysing spatial point patterns with associated functional data. In *Statistics for Spatio-temporal Modelling. Proceedings of the 4th International Workshop on Spatiotemporal Modelling (METMA-4)* pp. 157–163.

Cox, D.R., and Isham, V. (1980). *Point Processes.* Chapman and Hall/CRC, Boca Raton.

Cressie, N. (1993). *Statistics for Spatial Data.* John Wiley & Sons, New York.

De Boor, C. (2001). *A Practical Guide to Splines, Revised Edition.* Springer, New York.

Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics* **21** 224–239.

Gabriel, E., Bonneu, F., Monestiez, P., and Chadœuf, J. (2016). Adapted kriging to predict the intensity of partially observed point process data. *Spatial Statistics* **18** 54–71.

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** 585–602.

Giraldo, R., Delicado, P., and Mateu, J. (2010). Continuous time-varying kriging for spatial prediction of functional data: An environmental application. *Journal of Agricultural, Biological, and Environmental Statistics* **15** 66–82.

Giraldo, R., Delicado, P., and Mateu, J. (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics* **18** 411–426.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition.* Springer, New York.

Menafoglio, A., Secchi, P., and Dalla Rosa, M. (2013). A universal kriging predictor for spatially dependent functional data of a Hilbert space. *Electronic Journal of Statistics* **7** 2209–2240.

Møller, J., and Waagepetersen, R.P. (2004). *Statistical Inference and Simulation for Spatial Point Processes.* Chapman and Hall/CRC, Boca Raton.

Nair, R., and Miller-Hooks, E. (2011). Fleet management for vehicle sharing operations. *Transportation Science* **45** 524–540.

Shaheen, S., Guzman, S., and Zhang, H. (2010). Bike sharing in Europe, the Americas and Asia: Past, present and future. *Transportation Research Record: Journal of the Transportation Research Board* **2143** 159–167.

Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics* **10** 1040–1053.

Stone, C. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics* **22** 118–184.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

Zhou, S., Shen, X., and Wolfe, D.A. (1998). Local asymptotics for regression splines and confidence region. *The Annals of Statistics* **26** 1760–1782.