The Annals of Applied Statistics
2019, Vol. 13, No. 4, 2564–2585
https://doi.org/10.1214/19-AOAS1284

SCALABLE HIGH-RESOLUTION FORECASTING OF SPARSE SPATIOTEMPORAL EVENTS WITH KERNEL METHODS: A WINNING SOLUTION TO THE NIJ "REAL-TIME CRIME FORECASTING CHALLENGE"

BY SETH FLAXMAN[*,1], MICHAEL CHIRICO[†], PAU PEREIRA[‡] AND CHARLES LOEFFLER[§]

Imperial College London[*], Grab[†], Amazon, Inc.[‡] and University of Pennsylvania[§]

We propose a generic spatiotemporal event forecasting method which we developed for the National Institute of Justice's (NIJ) Real-Time Crime Forecasting Challenge (National Institute of Justice (2017)). Our method is a spatiotemporal forecasting model combining scalable randomized Reproducing Kernel Hilbert Space (RKHS) methods for approximating Gaussian processes with autoregressive smoothing kernels in a regularized supervised learning framework. While the smoothing kernels capture the two main approaches in current use in the field of crime forecasting, kernel density estimation (KDE) and self-exciting point process (SEPP) models, the RKHS component of the model can be understood as an approximation to the popular log-Gaussian Cox Process model. For inference, we discretize the spatiotemporal point pattern and learn a log-intensity function using the Poisson likelihood and highly efficient gradient-based optimization methods. Model hyperparameters including quality of RKHS approximation, spatial and temporal kernel lengthscales, number of autoregressive lags and bandwidths for smoothing kernels as well as cell shape, size and rotation, were learned using cross validation. Resulting predictions significantly exceeded baseline KDE estimates and SEPP models for sparse events.

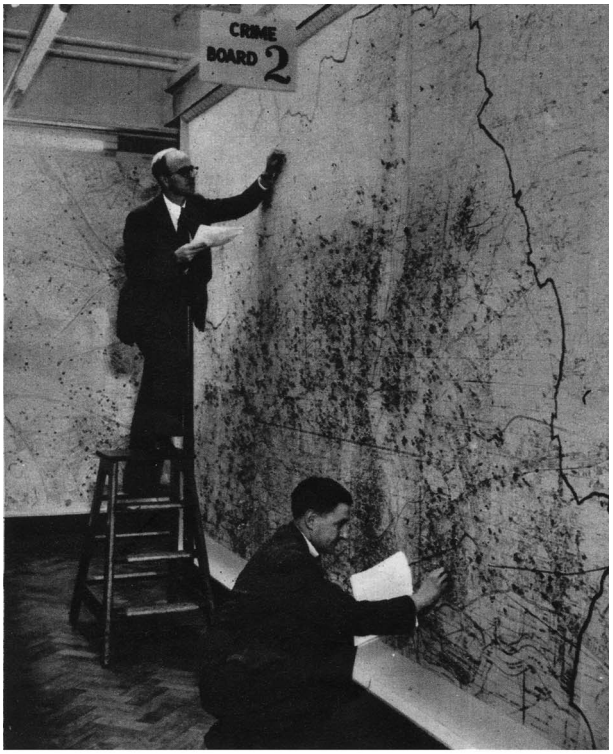
1. Introduction. Spatiotemporal forecasting of crime has been the focus of considerable attention in recent years as academic researchers, police departments and commercial entities have all sought to build forecasting tools to predict when and where crimes are likely to occur (Perry et al. (2013)). The earliest crime forecasting tools consisted of nothing more than pin maps (see Figure 1). Prior week's crimes were mapped and qualitative assessments of density, location, stability and significance were made (Schutt (1922)).

Subsequent tools have adopted a range of different smoothing techniques to augment this method with kernel density estimation the most commonly used approach (Gorr and Lee (2015), Porter and Reich (2012), Chainey, Tompson and

Received July 2018; revised July 2019.

[1]Supported by by the EPSRC (EP/K009362/1) and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013) ERC Grant agreement no. 617071.

Key words and phrases. Spatial statistics, time series, supervised learning, spatiotemporal forecasting, Cox process, RKHS.



CRIME AT A GLANCE: MARKING A MAP WITH COLOURED FLAGS IN THE NEW MAP ROOM AT SCOTLAND YARD.

The Map Room at Scotland Yard has recently been completed and is now in use to assist the police in their fight against crime. Every crime is allotted a coloured flag—red for burglary, yellow for housebreaking, and green for murder—so that a glance at the map is sufficient to show the areas where there is an increase in crime and the nature of the crimes committed. A recent outbreak of bag-snatching in one area was promptly countered by increasing the number of police officers on duty there.

FIG. 1. *Early use of crime pin maps at Scotland Yard*. 1947 ©*Illustrated London News Ltd/Mary Evans*.

Uhlig (2008), Johnson et al. (2009)). Many methods are model driven and based on theories of crime causation (Caplan, Kennedy and Miller (2011), Mohler et al. (2011)). Some use log Gaussian Cox Processes (LGCPs) (Rodrigues and Diggle (2012), Shirota and Gelfand (2017)), while others use self-exciting point process models (SEPPs) (Levine (2004), Liu and Brown (2003), Taddy (2010), Mohler et al. (2011), Rosser and Cheng (2019)) based on evidence of elevated levels of near-repeat victimization (Pease et al. (1998)). Some use additional information, such as weather, demographics and even social media (Wang, Gerber and Brown (2012)). Most simply use past events to forecast future events (Chainey, Tompson and Uhlig (2008), Kang and Kang (2017)), suggesting that methods that are effective at forecasting crime could readily be generalized to an increasing number of real-time spatiotemporal forecasting problems (Taddy (2010)). However, users of

these methods often confront the question of which method to adopt and how to ensure optimal performance across a wide variety of settings.

In 2016, the National Institute of Justice (NIJ) announced the Real-Time Crime Forecasting Competition to test which forecasting models could most accurately predict out-of-sample crime hotspots in the City of Portland. This solicitation drew in a wide range of competitors. Teams were given five years of historical calls for service data from the Portland Police Bureau (PPB) and asked to submit predictions for the locations of the largest crime clusters in the subsequent weeks and months.

Our team ("Team Kernel Glitches") tied for first place in the large organization category with wins across a range of categories. While our solution performed equally well on frequent and sparse crime forecasts and over short and long durations, it performed especially well, compared to competitors and contemporary methods, at forecasting sparse events over short durations. In describing our solution, we make the following contributions: we propose a flexible, generic and scalable spatiotemporal forecasting model, casting the problem of spatiotemporal forecasting explicitly as a supervised learning problem, while incorporating existing and highly successful modeling approaches from the spatiotemporal statistics literature: Gaussian processes, autoregressive terms, kernel smoothing and self-exciting point processes. This supervised learning setup provides a coherent framework for the time-consuming task of optimizing hyperparameters, while its modeling and inference scalability ensures that the model parameters themselves can be learned quickly enough to enable real-time forecasting. This approach achieves accuracy improvements well beyond those generated by existing best practices in crime prediction (Chainey, Tompson and Uhlig (2008), Johnson et al. (2009)).

The rest of this paper is laid out as follows. Section 2 describes our model. Section 3 describes the details of the NIJ competition. Section 4 reports competition performance. Section 5 concludes with a discussion of implications for future work on spatiotemporal prediction of crime and related phenomena.

2. Our model.

2.1. *Background*. Previous methods for spatiotemporal forecasting of crime have either focused on highly flexible but relatively simple kernel density estimation techniques (Johnson et al. (2009), Gorr and Lee (2015)) where crime events are aggregated over time, smoothed over space and used to predict crime patterns in the subsequent time period, or more complex and model-based approaches (Mohler et al. (2011), Rosser and Cheng (2019)). Recent work has demonstrated that Gaussian process modeling of crime data can produce highly accurate long-term forecasts by combining the benefits of nonparametric methods with the interpretability of additive methods (Flaxman (2014)). Subsequent work (Flaxman et al. (2015)) has proposed that, instead of specifying an additive kernel structure, it is is possible to learn it directly from the data, given enough data and a

rich enough class of kernels. This assumes, however, that it is possible to perform inference with very large datasets as the standard approach to Gaussian process inference requires matrix algebra to manipulate the multivariate Gaussian distribution, requiring $\mathcal{O}(N^3)$ time and $\mathcal{O}(N^2)$ storage. We therefore first present the hypothetical model we would use if computational constraints were not a concern, then our actual model which is an approximation to this model enabling application of this method to real-time rather than long-term forecasting problems.

2.2. *Model specification.* Our hypothetical model is a log Gaussian Cox Process (LGCP). The LGCP is a doubly stochastic point process model. Given an observation window W in spacetime, we place a GP prior on the log intensity $f(\boldsymbol{s})$ for any $\boldsymbol{s} \in W$. Let $N(\cdot)$ be a counting measure. For any spacetime region $S \subset W$, $N(S)$ is a Poisson distributed random variable counting the number of points in S. Our hierarchical parameterization is as follows:

$$
\begin{aligned}
f &\sim \mathcal{GP}\big(\mu, k_\theta(\cdot,\cdot)\big),\\
N(S)\,|\,f &\sim \text{Poisson}\bigg(\int_S \exp\big(f(\boldsymbol{s})\big)\,d\boldsymbol{s}\bigg).
\end{aligned}
\tag{2.1}
$$

We defer the specification of the mean μ and covariance kernel k_θ until later. For details on Gaussian processes, see Section A.1 of the Supplementary Materials (Flaxman et al. (2019a)).

Inference with the LGCP model is difficult because it is doubly intractable and existing approaches (Møller, Syversveen and Waagepetersen (1998), Brix and Diggle (2001), Cunningham, Shenoy and Sahani (2008), Adams, Murray and MacKay (2009), Teh and Rao (2011), Diggle et al. (2013)) are often limited to one dimension and small datasets. Lloyd et al. (2015) is a possible exception in that it points the way to a scalable stochastic variational inference approach.

To approximate this model, we discretize. We specify a spacetime grid partitioning W into N disjoints sets S_i, that is, $W = \bigcup_{i=1}^N S_i$. As described below, this approach leads to a tractable model. Also, it is consistent with the design of the forecasting competition motivating our approach. For simplicity, let each grid cell S_i be of equal volume $|S_i| = 1$. The centroid of each grid cell is a latitude/longitude/timestamp triple $\boldsymbol{s}_i = (x_i, y_i, t_i)$. The underlying point pattern is then represented as aggregate counts $o_i = N(S_i)$ of the number of crimes per cell. Given the grid, the integral in equation (2.1) is approximated with a sum. When considering the entire observation window W, the approximation takes the following form:

$$
\int_W \exp\big(f(\boldsymbol{s})\big)\,d\boldsymbol{s} \approx \sum_{i=1}^N \exp\big(f(\boldsymbol{s}_i)\big)|S_i| = \sum_{i=1}^N \exp\big(f(\boldsymbol{s}_i)\big).
\tag{2.2}
$$

In a Poisson process, conditional on the intensity, the random variables $N(S_1)$ and $N(S_2)$ are independent for $S_1 \cap S_2 = \varnothing$. Thus given the log intensity f, each grid

cell S_i can be considered independently, so combining equations (2.1) and (2.2) yields

$$o_i | f \sim \text{Poisson}\big(\exp\big(f(\boldsymbol{s}_i)\big)\big) \quad \forall i = 1, \ldots, N. \tag{2.3}$$

This produces an i.i.d. likelihood (observation model) over all cells i, yielding the so-called computational grid approximation to the log Gaussian Cox Process (Diggle et al. (2013), Flaxman et al. (2015)).

In the function-space view of GPs, inference is performed about the function f directly. Using the "kernel trick" (Schölkopf and Smola (2002)), all calculations can be carried out using a kernel k_θ, evaluated at all pairs of $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$. However, to do this requires storing and manipulating an $N \times N$ covariance matrix K at a cost of $\mathcal{O}(N^2)$ storage and $\mathcal{O}(N^3)$ computation (Rasmussen and Williams (2006)) which is infeasible for large N.

By contrast, the weight-space view of GPs (Rasmussen and Williams (2006), Ch. 2) requires an explicit feature map $\phi(\boldsymbol{s}) = k_\theta(\boldsymbol{s}, \cdot) \in \mathcal{H}$ where $\mathcal{H}$ is the Reproducing Kernel Hilbert Space corresponding to the kernel k_θ with $\phi(\boldsymbol{s})^\top \phi(\boldsymbol{t}) = k_\theta(\boldsymbol{s}, \boldsymbol{t})$. Instead of learning f directly (function space), for finite dimensional $\mathcal{H}$ a set of weights β can be learned by considering the vector $\phi(\boldsymbol{s})$ as a set of basis functions. Thus, we define $f(\boldsymbol{s}) := \phi(\boldsymbol{s})^\top \beta$ and observe that the weight-space view is equivalent to a linear model with a particular set of basis functions.

In practice, the weight-space view is not computationally tractable in the case of popular universal (Sriperumbudur, Fukumizu and Lanckriet (2011)) kernel choices like the Gaussian or Matérn kernel because the corresponding $\mathcal{H}$ is infinite dimensional. Unlike infinite-dimensional universal kernels, kernels corresponding directly to finite-dimensional RKHS are limited in their representational capacity, for example, polynomial kernels of order p only capture p moments of a distribution. A solution can be found, following recent trends in the literature (May et al. (2019)), using finite-dimensional approximations to universal kernels in the form of the random Fourier feature expansion (Rahimi and Recht (2007)) as described in Section A.2 of the Supplementary Materials (Flaxman et al. (2019a)). For any kernel this requires the selection of a dimension d, which determines the accuracy of the approximation $\hat{\phi} \in \mathcal{R}^{2d}$ where $\hat{\phi}(\boldsymbol{s})^\top \phi(\boldsymbol{t}) \approx \phi(\boldsymbol{s})^\top \phi(\boldsymbol{t}) = k_\theta(\boldsymbol{s}, \boldsymbol{t})$. An example of our approximation is illustrated in Figure A1 in the Supplementary Materials (Flaxman et al. (2019a)) where the Matérn-5/2 kernel is approximated using various values of d.

A finite dimensional $\hat{\phi}$ leads from the function-space view to the weight-space view (Rasmussen and Williams (2006), Ch. 2, Milton et al. (2019)). To make the connection explicit, we define a kernel $\hat{k}_\theta(\boldsymbol{s}, \boldsymbol{t}) = \hat{\phi}(\boldsymbol{s})^\top \hat{\phi}(\boldsymbol{t})$. Define a matrix Φ for observations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$ with each row $\Phi_i = \hat{\phi}(\boldsymbol{s}_i)^\top$. The function-space view on Gaussian process regression with covariance kernel $\hat{k}$ and a Gaussian likelihood is

$$\begin{aligned} f &\sim \mathcal{GP}\big(\mu, \hat{k}_\theta(\cdot, \cdot)\big), \\ y | f, \boldsymbol{s}_i &\sim \mathcal{N}\big(f(\boldsymbol{s}_i), \sigma^2\big). \end{aligned} \tag{2.4}$$

Equation (2.4) is equivalent to Bayesian linear regression with $\beta \in \mathcal{R}^{2d}$ (where the term weight-space view comes from considering the parameter vector β as "weights" to be learned):

$$\begin{aligned} \beta &\sim \mathcal{N}(0, I), \\ y|\beta, \boldsymbol{s}_i, \Phi &\sim \mathcal{N}\big(\mu(\boldsymbol{s}_i) + \Phi_i\beta, \sigma^2\big). \end{aligned} \tag{2.5}$$

For the present application, the data consists of count-valued observations, so we adopt a generalized linear modeling (GLM) framework and replace the Gaussian likelihood in equation (2.5) with the Poisson likelihood as in equation (2.3):

$$o_i|\beta, \boldsymbol{s}_i, \Phi \sim \text{Poisson}\big(\exp\big(\mu(\boldsymbol{s}_i) + \Phi_i\beta\big)\big). \tag{2.6}$$

It remains to specify the function μ. In the spatial statistics literature a linear model using spatially varying covariates is standard (e.g., Diggle et al. (2013)), while $\mu = 0$ is a common default choice in machine learning, though recent work has questioned this approach (Bhatt et al. (2017)). We consider a different approach, based on prior work that has shown that using historical crime rates can be very effective in crime forecasting. Expanding upon prior KDE-forecasting methods that search a limited number of possible values and in line with the supervised learning framework discussed above, μ is parameterized as follows for $\boldsymbol{s} = (x, y, t)$:

$$\mu(\boldsymbol{s}) = \sum_{j=1}^{p} \gamma_j \text{KDE}_{\lambda,j}(x, y, t), \tag{2.7}$$

where there are p autoregressive lagged terms, each representing a spatial KDE for a given time period in the past and regression coefficients γ_j are to be learned. $\text{KDE}_{\lambda,j}(x, y, t)$ is the kernel density estimator at location (x, y, t) using a spatial Gaussian kernel κ_λ with lengthscale λ,

$$\text{KDE}_{\lambda,j}(x, y, t) = \sum_{\{t_i | t - j\cdot D < t_i \le t - (j-1)\cdot D\}} \kappa_\lambda\big((x, y), (x_i, y_i)\big), \tag{2.8}$$

where D is the size of the temporal window in days.

Given the potential for a large number of parameters β (the more random frequencies d we choose for the random Fourier feature expansion, the better our approximation); the use of ℓ_1 and ℓ_2 regularization (as in the popular elastic net (Zou and Hastie (2005))) provides a useful simplification.

Finally, our objective is to maximize the penalized log likelihood of the Poisson distribution. Simplifying and dropping constant terms yields the following objective with parameters β and γ and regularization hyperparameters a and b:

$$\begin{aligned} &\sum_{i=1}^{N}\left[o_i\left(\sum_{j=1}^{p}\gamma_j \text{KDE}_{\lambda,j}(x_i, y_i, t_i) + \Phi_i\beta\right) - e^{\sum_{j=1}^{p}\gamma_j \text{KDE}_{\lambda,j}(x_i, y_i, t_i) + \Phi_i\beta}\right] \\ &\quad - a\big(\|\beta\|_1 + \|\gamma\|_1\big) - b\big(\|\beta\|_2^2 + \|\gamma\|_2^2\big). \end{aligned} \tag{2.9}$$

2.3. *Inference.* We learn the parameters β and γ by maximizing the objective in equation (2.9) using gradient ascent. The random Fourier feature approximation combined with linear regression leads to immediate speedups and memory savings whereas full GP regression is $\mathcal{O}(N^3)$ time and $\mathcal{O}(N^2)$ storage, calculating the random features for Φ is $\mathcal{O}(Nd)$ for both time and storage. Given a fixed design matrix Φ, ordinary linear regression requires calculating $\Phi^\top \Phi$ which is $\mathcal{O}(Nd^2)$ time and $\mathcal{O}(d^2)$ storage. Depending on how the lasso and ridge penalties are implemented, penalized linear regression can be very efficient, for example, cyclical coordinate descent takes $\mathcal{O}(Nd)$ time for each update of all of the parameters (Friedman, Hastie and Tibshirani (2010)). The important point is that the overall running time is linear in N rather than cubic, a significant savings in time. This approach is competitive with standard approaches to scalable inference in the spatial statistics literature (Sun, Li and Genton (2012), Milton et al. (2019)).

During competition, we performed optimization using the large-scale machine learning package Vowpal Wabbit (http://hunch.net/~vw). Vowpal Wabbit employs feature hashing (Weinberger et al. (2009)) and online learning, which is even faster than the standard $\mathcal{O}(Nd^2)$ approach to linear regression, allowing it to scale up to handle huge datasets. We fit the training dataset using default settings for the learning algorithm (a variant of online gradient descent), with at most 200 training passes (epochs) through the dataset. As a stopping criterion for convergence was applied, running times did not directly vary with dataset size. For any given set of hyperparameters (except the regularization parameters), a new dataset was produced and saved to disk, and then this model was fit across the full path of regularization parameters by repeatedly running Vowpal Wabbit. The entire process of dataset creation and multiple calls to Vowpal Wabbit usually took about half an hour, even with datasets as large as $N \approx 300$ k (one week time horizon). All of our computation was carried out in a parallel cluster computing environment, with 8 Dell PowerEdge R630 nodes. Each node consisted of 2 $\times$ Intel Xeon E5-2690 v4 2.6 GHz, 14 Core CPUs, and 256 GB memory.

After fitting the model, we made predictions in the form of counts for the test data, and then calculate PEI for each year of data, where PEI is a forecasting accuracy metric used in the crime forecasting literature. To learn the hyperparameters, we maximize average PEI. The hyperparameters related to our model are as follows: the number of random features d in our feature expansion, the number of lags p, the size of the temporal window D, the spatial lengthscale for KDE λ (with a Gaussian kernel), the lengthscale θ of the covariance kernel k_θ (we used a Matérn-5/2 kernel, a standard choice in spatial statistics (Guttorp and Gneiting (2005))) and the amount of ℓ_1 and ℓ_2 regularization a and b. In addition, there are competition-related hyperparameters that are learned, including cell size, shape, grid rotation and forecast area. We cross validated over a very large grid of hyperparameters, considering a range of values for each parameter and every possible combination of these values. As an alternative method to further explore

the entire space of hyperparameter choices, we separately performed a hyperparameter search using sequential Bayesian Optimization (O'Hagan (1992), Snoek, Larochelle and Adams (2012), Hennig, Osborne and Girolami (2015)). Having run both searches, we combined the results and chose the best sets of hyperparameters based on cross validated average PEI. Additional details are given in Section C of the Supplementary Materials (Flaxman et al. (2019a)).

2.4. *Relationship with prior work*. Supervised learning methods are widely used within nonspatiotemporal applications. However, they are less commonly used within the applied spatial (Heaton et al. (2019)), time series (Makridakis, Spiliotis and Assimakopoulos (2018)) and crime forecasting domains. In crime forecasting KDE-based forecasting approaches remain the most common forecasting techniques used (Gorr, Olligschlaeger and Thompson (2003), Gorr (2009), Chainey, Tompson and Uhlig (2008), Caplan, Kennedy and Miller (2011), Berk et al. (2018)). While small numbers of parameters may be user selected and modified, these methods are commonly implemented absent any framework for maximizing the objective function of forecasting accuracy. Instead, practitioners modify parameters on an ad hoc basis, assuming that the resulting forecasts are a reasonable implementation of KDE methods. For a recent exception to this approach, see Rosser and Cheng (2019).

When prior work has sought to improve upon the performance of these less-than-optimized KDE forecasts, the principle area of focus has not been on scalable hyperparameter optimization, but instead on implementing model-based characterizations of underlying crime intensities. Some work has focused on modeling spatial and temporal range of crime decays (Johnson et al. (2009)), but the Hawkes process has recently been the focus of significant attention in the crime forecasting literature (Ogata (1988), Møller and Rasmussen (2005), Mohler et al. (2011), Mohler (2013, 2014), Rosser and Cheng (2019), Loeffler and Flaxman (2018)). Both approaches seek to avoid a common feature of prior KDE methods which implicitly weight all prior events as equally informative with no attention to recency. However, the question of how to identify the optimal spatial and temporal range of crime decay is also not entirely addressed in these contributions.

The logic of our approach is that it combines state-of-the-art nonparametric spatiotemporal methods (Gaussian process regression), which fundamentally encode an assumption of spatial and temporal autocorrelation, with the most long-standing and widely used crime forecasting method (KDE surfaces) by defining sets of features for each. By placing these two sets of features into a penalized supervised learning framework for forecasting the intensity and considering a large set of hyperparameters and training data, we hope to combine the benefits of nonparametric modeling, principally accuracy in the absence of a known best model, with the benefits of parametric modeling, principally model simplicity, to obtain good predictive performance on unseen data. For a discussion of the similarities of optimized KDE features and so-called "Hawkes features," see Section A.3 of the Supplementary Materials (Flaxman et al. (2019a)).

3. The competition. The goal of the NIJ Real-Time Crime Forecasting Competition was to forecast hotspots for several categories of calls for service to the Portland Police Bureau (PPB) in Portland, Oregon. Contestants submitted forecasts on (or before) February 28, 2017, for various time horizons starting on March 1, 2017, and extending as late as May 31, 2017. The hotspot predictions were scored on two metrics related to their accuracy. Contest rules required that contestants predict which of the 62,500–360,000 square foot cells within PPB's 147.71 square mile service area would have the highest number of calls for service, with the total forecast area being no smaller than 0.25 square miles and no larger than 0.75 square miles, equivalent to forecasting 175–525 city blocks out of a total of 103,397 blocks. Prizes were given out for five different cumulative forecast periods (one week, two weeks, four weeks, eight weeks, 12 weeks), four different crime categories (burglary, street crime, theft of auto, all calls for service) and two different accuracy metrics.

3.1. *Data and setting*. The NIJ Real-Time Crime Forecasting dataset consists of 958,499 calls for service records from the Portland Police Bureau (PPB), representing calls to Portland's 911 system requesting police assistance from March 1st, 2012, through February 28th, 2017. As shown in Figure 2, the four categories of crime, which themselves varied in the degree of internal heterogeneity, included burglary (burglary and prowling), street crime (ranging from disturbance and threats up to armed robbery and assault with a Firearm), theft of auto and all calls for service.

3.2. *Metrics*. The simplest metric for evaluating the accuracy of crime forecasts is the "hit rate" (Chainey, Tompson and Uhlig (2008)),[2]

$$\text{Hit rate} = \frac{n}{T},$$

where n is the number of crimes predicted and T is the total number of crimes in that period in that area. Performance on this metric depends critically on the size of the forecasted area in addition to underlying crime densities and forecasting quality. In the case of the NIJ competition, this coverage area was between 0.2% and 0.5% of the City of Portland.

The NIJ competition focused on two alternatives metrics (Chainey, Tompson and Uhlig (2008), Hunt (2016)) with a goal of allowing for a comparison of hit rates across forecasts using different coverage areas. The first metric, the prediction accuracy index (PAI) (Chainey, Tompson and Uhlig (2008)), is the ratio of the hit rate to the fraction of area covered:

$$\text{PAI} = \frac{\frac{n}{T}}{\frac{a}{A}}.$$

[2]It is also known as sensitivity in the statistics literature. See Adepeju, Rosser and Cheng (2016) for a recent discussion of alternative evaluation metrics for crime forecasting.

FIG. 2. *The competition focused on four categories of crimes, ranging from the very abundant (all calls for service) to the very sparse (burglaries). Shown here are the locations of reported crimes in February* 2016.

This metric directly incorporates the trade-off between hit rate and coverage, as in an ROC curve, into the score weighting.

The second metric, the prediction efficiency index (PEI), is the ratio of PAI to the hypothetically maximum PAI that could have been obtained using the chosen coverage area and discretization of space. Since the forecasting area is the same in both the actual and hypothetical maximum cases, this reduces to

$$\mathrm{PEI} = \frac{n}{n^*},$$

where n is the number of crimes occurring in predicted hotspots and n^* is the maximum number of crimes that could have been captured for the forecasted area.

While optimizing either metric will produce similar results some of the time, optimizing for PEI incurs a PAI penalty proportional to the marginal change in forecasted area divided by the marginal change in correctly forecasted crimes. Therefore, the optimal cell selection for maximizing PEI will often fail to maximize PAI. For the competition we maximized the PEI metric. (For a result making the opposite choice, see Mohler and Porter (2018).)

3.3. *Data for training and hyperparameter selection*. For a given spatial grid size we restricted our temporal windows to match the corresponding forecasting window. For example, for a one week forecasting window the training data is aggregated to the weekly level. The training period consisted of each prior year's aggregated counts, excluding the corresponding time period, being forecasted. This excluded period formed the validation period. We then created a single dataset using data from the union of all of the training and validation periods. Using this dataset, we forecasted hotspot maps for the five different validation periods, corresponding to the five different years of pre-2017 data available and calculated PEI for each. The average of this heldout PEI was then maximized to select the hyperparameters of the model.

4. NIJ challenge results. In this section we describe the performance of our method according to the scoring metrics of the NIJ challenge, assess its robustness and investigate what features of the model contributed to its out-of-sample performance. Source code for replication is available in the Supplementary Materials (Flaxman et al. (2019b)) and online at https://github.com/MichaelChirico/portland.

There were a total of 40 prizes awarded, one for each of the highest PEI and PAI scores in each crime category and forecasting window. Our team won a total of nine prizes in the "Large Business" competition. As we focused on maximizing the forecasting performance on the out-of-sample PEI metric, most of our winning entries were in this category: all calls for service (one week, one month, three months), burglary (one week, two weeks), street crime (two weeks) and theft of auto (one week). In addition, we also had winning PAI entries for burglary (one week and two weeks).

At the heart of our model was a hyperparameter search strategy, in which final models were selected from the union of all models explored by an exhaustive grid search coupled with a Bayesian Optimization designed to optimize forecasting accuracy. In practice, there were no consistently chosen hyperparameter values: the grid cells were sometimes small squares, 250 ft $\times$ 250 ft (the minimum area), or large squares, 600 ft $\times$ 600 ft (the maximum area) or large rectangles, 800 ft $\times$ 450 ft (also the maximum area). The coverage fraction ranged from the minimum (0.25 sq miles) to the maximum (0.75 sq miles). The lengthscales for space and time were highly varied, as were the number of KDE lags and the KDE bandwidth. The number of random Fourier features went as low as $d = 5$, which

means that the surface was a very crude approximation to a Gaussian process consisting of the sum of 10 random sine and cosine functions, to as high as 362, a much better approximation. In a minority of cases, no ℓ_1 or ℓ_2 regularization was needed, but most final models used at least some ℓ_2 regularization. In a minority of cases (four out of 20) the best hyperparameters turned out to be those found by Bayesian Optimization, while in all other cases the best hyperparameters were those found by grid search. (See Table A1 in the Supplementary Materials for details (Flaxman et al. (2019a)).) The lack of overlap in optimal hyperparameter selection across competition categories both reinforces the importance of supervised learning optimization for forecasting accuracy and raises the question of whether other, possibly more uniform, hyperparameter choices might also exist.

We examine the distribution of all PEI values obtained in our grid search for each category/forecast window separately. For the one week theft of auto and burglary categories, 41% and 44%, respectively, of the possible hyperparameter combinations gave PEI scores of zero. This is strong evidence for the importance of an exhaustive hyperparameter search, at least for these sparse events. To further quantify this numerically, we calculate the z-score of the maximum PEI for the distribution of PEIs for each category/forecast window. Our winning theft of auto one week entry had a PEI z-score of 21, and our winning burglary entries had z-scores of 12.4 (one week) and 11 (two weeks), all results which are consistent with the idea that good forecasting accuracy requires an exhaustive hyperparameter search. The distributions for more abundant crime types did not yield such extreme z-scores: in the All Calls for Service category, the z-scores of the maximum PEIs ranged from 2.5 to 4.0. In the street crimes category the z-scores of the maximum PEIs ranged from 2.8 to 5.6. Thus, for more abundant crime types a range of hyperparameters could produce similar results.

A final question concerning the competition is the maximum achievable level of forecasting accuracy. As shown in Figure 3, which depicts the maximum achieved PEI for all competitors, for high volume crimes, such as all calls for service, even a week's worth of data is sufficient to achieve very high PEI scores (nearly 0.9) of the theoretical limit (one) for a one week prediction. Extending the cumulative forecast period leads to further improvements in forecasting accuracy, plateauing at 97%. Sizable subcategories, such as street crimes, share this basic trajectory as well, suggesting that for high volume crimes over both short and medium-term horizons near limit and unity performance can be expected. For some sparse crimes, such as theft of auto, despite lower starting values, similar improvements in predictive accuracy can be seen as the forecasting windows are expanded, even if these improvements are not strictly monotonically increasing. Whether a longer horizon would lead to further improvements is unknown. However, for other sparse crimes, such as burglary, adding additional weeks of data to the forecast period does little to improve maximum achieved forecast accuracy. Reinforcing the idea that crime forecasting is not a single problem but several, only some of which are more accurately solved through the addition of more data.

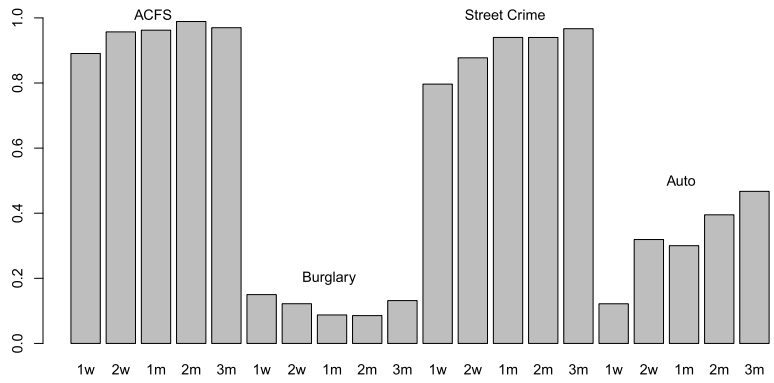


FIG. 3. *Competition Maximum PEI Performance Among All Competitors. Each column represents the highest level of forecasting accuracy achieved across all competitors for a particular combination of crime type and forecasting period. From left to right all calls for service, burglary, street crime and theft of auto for one week, two weeks, one month, two months and three months* (*also left to right*).

4.1. *Investigating method performance.* As discussed in Section 1, many crime forecasting implementations rely on KDE-type approaches. As our model included lagged KDE terms, we expected to always perform as well as a KDE-type baseline. As a post-competition check, we fit a model with just one KDE lag, corresponding to a KDE-type baseline, and fixed parameters according to common practice (Chainey and Ratcliffe (2005)) and found that our model was better than this baseline 90% of the time (18 cases out of 20) on the true out-of-sample forecasted data with an average absolute improvement of 0.16 for the PEI scoring metric. The improvements were most notable for sparse crimes and short time-horizons, as the baseline model often identified no correct theft of auto or burglary hotspots (Figure 4). Interestingly, for the two forecasts for which simple KDE outperformed our model (e.g., burglary 2 months and 3 months), hyperparameters for the model were selected using Bayesian Optimization (BO) rather than grid search, suggesting that BO will not always give the optimal set of parameters. Comparing the full method, which combines lagged KDE terms and a Gaussian process surface, to a model without the Gaussian process surface, the full model gave better PEI results 75% of the time. The average absolute improvement was 0.05. Thus, although the full model is an improvement, the improvements are not as dramatic as going from a simple KDE to a lagged KDE model with kernels optimized for forecasting accuracy. This result suggests that for many models, especially ones predicting sparse events (as depicted in Figure 5), the routine variation in performance is sufficient to swamp the benefits of using Gaussian process surfaces or other complex methods. Instead, considerable portions of achievable performance improvements can be realized by optimizing the parameters of simpler methods, such as lagged KDEs.

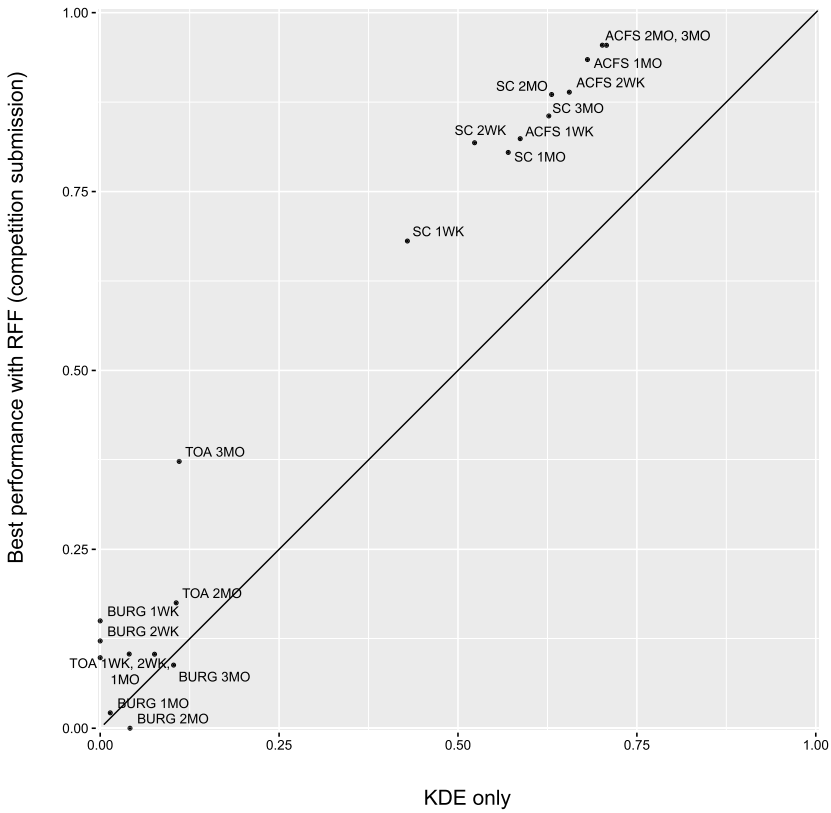


FIG. 4. *KDE Baseline* (*x-axis*) *compared to Full Model* (*y-axis*). *The Full Model outperformed the KDE Baseline Model in* 18 *out of* 20 *forecast problems. The average out-of-sample performance improvement of the Full Model over the KDE Baseline Model was* 0.16 *on the PEI scoring metric. BURG = burglary, SC = street crime, TOA = theft of auto, ACFS = all calls for service.*

Rosser et al. (2016) recently demonstrated that due to geocoding, noncardinal land use and other related factors, a nonstandard alignment could improve predictive accuracy in crime forecasting. In the present application, we explored altering the rotation of the entire tessellation and the dimensions of the cell rectangles. With improved performance of only 0.029 on the PEI scoring metric for a freely rotated model when compared to the best performing nonrotated model, rotation does not appear to be a major contributor to overall performance. However, certain crime categories and forecast windows can be observed to benefit more substantially. A similar result can be observed for altering cell dimensions, which only improves overall performance on the PEI scoring metric by 0.019, when compared to a conventionally used 600 ft $\times$ 600 ft rectangle. (See Figures A2–A4 in the Supplementary Materials (Flaxman et al. (2019a)) for more details.) These results parallel previous findings that showed the limited return on the inclusion of nonautoregressive information (Wang, Gerber and Brown (2012), Gerber (2014)).

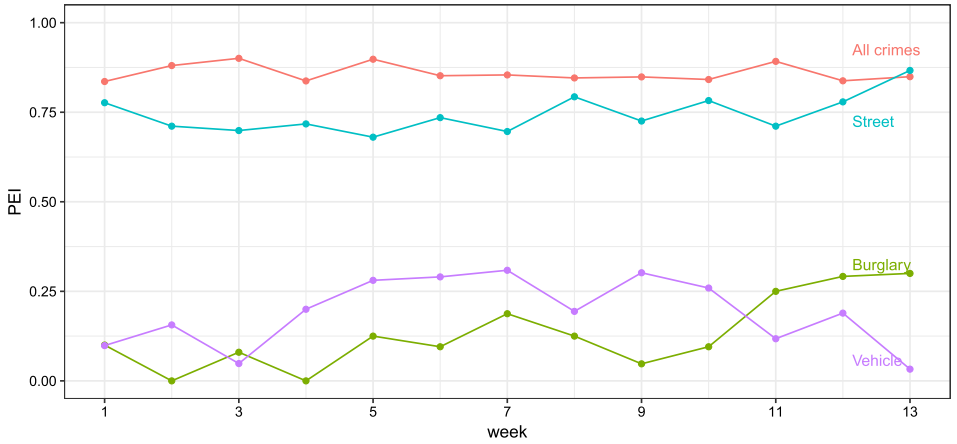


FIG. 5. *Rolling Forecast. The* 13 *week competition period* (*March through May* 2017) *was split into* 13 *one-week forecast periods and a one-week rolling forward prediction was made for each week using the competition model trained to predict only the first out-of-sample week.*

The sparseness of several of the forecasted incidents and recent findings on lack of robustness of forecasting models (Rosser and Cheng (2019)) suggests that it is worthwhile to examine the stability of the model's performance over multiple periods. To accomplish this, the 13 week competition period (March through May 2017) was split into 13 one-week forecast periods and a one-week rolling forward prediction was made for each week. The resulting predictions, as seen in Figure 5, manifest variability consistent with the stochastic events being predicted. However, these rolling-forward predictions provide little evidence of over-fitting to the first out-of-sample time period, even for the sparsest of incidents. They instead suggest, at least for settings like the competition, that the short-term accuracy improvements are robust and stable.

Alongside submodel component performance and model stability, a final area of interest is method error. Figure 6 (left) shows the actual performance of the full forecasting model for a high volume crime category (ACFS) and a middle-range forecasting period (one month). Polygons that were correctly forecast as the highest possible crime count polygons are in green. Polygons incorrectly forecast to not be hotspots are in red. And polygons that were incorrectly predicted to be the highest possible crime count polygons are depicted in blue. Crimes are black dots. The largest single cluster of hotspots for all calls for service can be seen downtown. However, the model slightly over invested in this section of Portland. As can be seen in the inset, hotspots just across the Willamette River had more crimes reported over the relevant forecast window. In practice, most of these misses were relatively small with "false negatives" only slightly "hotter" than the corresponding "false positive" cells (e.g., 44 crimes in a FN cell versus 39 crimes in a FP cell).

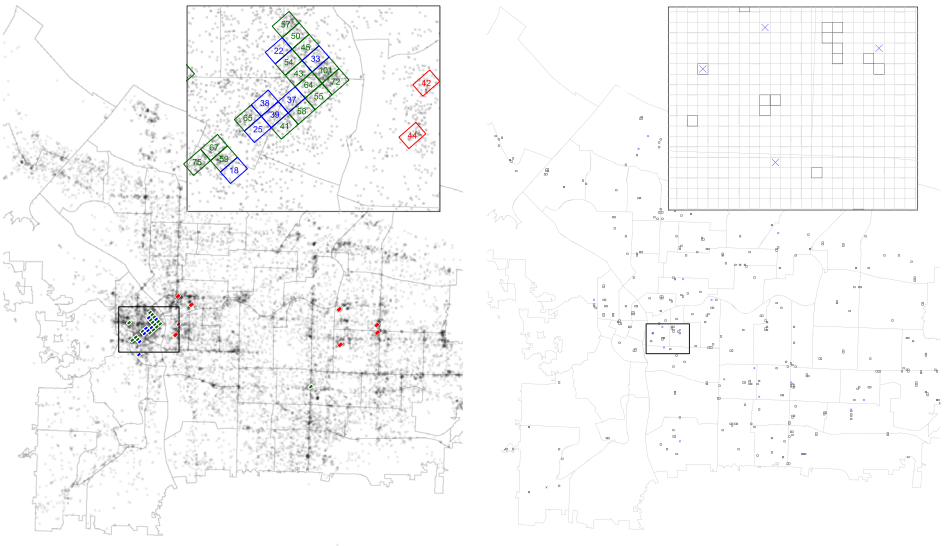


FIG. 6. *Left: all calls for service one month. Correctly forecast polygons are in green. "False negative" polygons are in red. "False positive" polygons are depicted in blue. Crimes are black dots. Right: burglary one week. Forecasted burglary cells are depicted with boxes. Actual burglaries are depicted by blue x's. Boxed x's indicate a successful prediction. Empty boxes indicate a "false positive" prediction.*

Figure 6 (right) show the actual performance of the model for a sparse crime category (burglary) and a short-range forecasting period (one week). Forecasted burglary cells are depicted with boxes and actual burglaries are depicted by blue x's. Boxed x's indicate a successful prediction, while empty boxes indicate a "false positive" prediction. The absence of large-scale clustering is quite visible in both the dispersion of the burglaries throughout Portland and in the similarly dispersed allocation of predictions. As can be seen in the inset, a successful prediction was accompanied by several near misses in the vicinity, including one near-miss off by only a single cell. Predicting sparse crimes, while more difficult than predicting concentrated crimes, is still achievable and with accuracy levels not previously seen with other conventional forecasting methods.

5. Discussion. Real-time spatiotemporal forecasting is an area of increasing interest. Yet many common approaches, such as kernel smoothing based on fixed bandwidths and cell sizes, can be quite limited in their out-of-the-box accuracy, especially for sparse events. Past work (Johnson et al. (2009)) has reported one-week burglary forecasting accuracy of 10% at 1.3% of coverage area and 25% of burglaries at 5% of coverage area using near-repeat models with baseline KDE models producing one-week accuracy of 10% at 2% coverage and 25% at 6.5% coverage. Mohler et al. (2011) report 5% accuracy for daily predictions at comparable coverage levels. By comparison, using the described methods, median one-

week burglary accuracy of 10% was achieved with a coverage area of 0.5% and 50% of the time 25% forecasting accuracy was achieved at 0.5% coverage.

These results build upon prior work exploring parameter tuning (Chainey (2013), Rosser and Cheng (2019)) and reinforce three points. First, it appears that simple but well-tuned models incorporating lagged kernel smoothing can achieve many of the benefits commonly associated with more complex methods. This conclusion stems from the recognition that parameter optimization, particularly in the case of kernel smoothing, is a reweighting of different spatiotemporal portions of an autoregressive process for forecasting accuracy. Second, the poor performance of conventional kernel estimators with parameters set based on rules-of-thumb suggests that many existing crime forecasting implementations are not as accurate as they could be. Third, while some parameters are more important than others, no one parameter is universally better and, as such, supervised learning will likely be a continuing feature of spatiotemporal crime forecasting.

While the results reported here suggest that forecasting the hottest high volume crime hotspots can be done with great accuracy using a variety of techniques, the same cannot be said for sparse events, at least not yet. This leaves as an open question whether rare crime events are intrinsically harder to forecast due to random error or are simply harder because of insufficient training data. The fact that some rare crime forecasts saw no improvement in forecasting accuracy, despite the addition of more training data and larger cumulative forecasting windows could be considered suggestive evidence that there may be a signal limit for this type of event. However, refitting our models in other settings would shed further light on this question, as would the inclusion of additional predictors. For example, μ based on the kernel density estimates of other types of crimes, inspired by criminology research on "leading indicators" of crime (Cohen, Gorr and Olligschlaeger (2007)).

Another question not answered by these results is why this method's performance was not more uniform. One possible answer is that the methods described in this paper simply do a better job at forecasting certain types of events over certain forecasting windows. Another possibility is that incomplete grid-search of hyperparamaters during competition led to the use of suboptimal parameters for certain forecasting subtasks. A final possibility is that the close performance of competitors, on at least some forecasting tasks, achieved near limit forecasting performance using known methods and data. In future work in other settings, these possibilities could more readily be teased out.

Pending completion of this research, the absolute performance of different methods in this competition also raises the policy question, "What is an acceptable level of accuracy for any crime forecasting method to be used?" In recent years crime forecasting tools have been a supplement or replacement for traditional crime analysis (Mohler et al. (2015)) with applications to police deployment, enforcement actions targeted at particular individuals or places (Lum and Isaac (2016), Perry et al. (2013)) as well as nonenforcement notification strategies (Groff and Taniguchi (2019)). These applications, especially those involving

law enforcement activity, have elevated concerns about fairness in criminal justice decision making, leading to a vigorous debate about definitions of algorithmic fairness (Berk et al. (2018), Corbett-Davies et al. (2017), Mitchell et al. (2018)). While fairness is an important debate, we have focused instead on accuracy, as this is a necessary precondition to considerations of fairness (Dressel and Farid (2018), Rudin and Ustun (2018)). As the results of our research suggest, opportunities for large gains in accuracy exist through the use of standard machine learning frameworks and spatial statistical methods.

Acknowledgments. Special thanks to our systems administrators: Tony Vo (University of Pennsylvania) and Stuart McRobert (Oxford).

SUPPLEMENTARY MATERIAL

Supplement to "Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ 'Real-Time Crime Forecasting Challenge'" (DOI: 10.1214/19-AOAS1284SUPPA; .pdf). Supplement on scalable Gaussian processes, additional results, hyperparameter choices.

Source code for "Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ 'Real-Time Crime Forecasting Challenge'" (DOI: 10.1214/19-AOAS1284SUPPB; .zip). R source code for the models described in this paper and data files from the NIJ competition.

REFERENCES

ADAMS, R. P., MURRAY, I. and MACKAY, D. J. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* 9–16. ACM, New York.

ADEPEJU, M., ROSSER, G. and CHENG, T. (2016). Novel evaluation metrics for sparse spatiotemporal point process hotspot predictions—A crime case study. *Int. J. Geogr. Inf. Sci.* **30** 2133–2154.

BERK, R., HEIDARI, H., JABBARI, S., KEARNS, M. and ROTH, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* 0049124118782533.

BHATT, S., CAMERON, E., FLAXMAN, S. R., WEISS, D. J., SMITH, D. L. and GETHING, P. W. (2017). Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. *J. R. Soc. Interface* **14** 20170520.

BRIX, A. and DIGGLE, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 823–841. MR1872069

CAPLAN, J. M., KENNEDY, L. W. and MILLER, J. (2011). Risk terrain modeling: Brokering criminological theory and GIS methods for crime forecasting. *Justice Q.* **28** 360–381.

CHAINEY, S. P. (2013). Examining the influence of cell size and bandwidth size on kernel density estimation crime hotspot maps for predicting spatial patterns of crime. *Bull. Geogr. Soc. Liege* **60** 7–19.

CHAINEY, S. and RATCLIFFE, J. (2005). *GIS and Crime Mapping*. Wiley, New York.

CHAINEY, S., TOMPSON, L. and UHLIG, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Secur. J.* **21** 4–28.

COHEN, J., GORR, W. L. and OLLIGSCHLAEGER, A. M. (2007). Leading indicators and spatial interactions: A crime-forecasting model for proactive police deployment. *Geogr. Anal.* **39** 105–127.

CORBETT-DAVIES, S., PIERSON, E., FELLER, A., GOEL, S. and HUQ, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the* 23*rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 797–806. ACM, New York.

CUNNINGHAM, J. P., SHENOY, K. V. and SAHANI, M. (2008). Fast Gaussian process methods for point process intensity estimation. In *Proceedings of the* 25*th International Conference on Machine Learning* 192–199. ACM, New York.

DIGGLE, P. J., MORAGA, P., ROWLINGSON, B. and TAYLOR, B. M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: Extending the geostatistical paradigm. *Statist. Sci.* **28** 542–563. MR3161587

DRESSEL, J. and FARID, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **4** eaao5580.

FLAXMAN, S. R. (2014). A general approach to prediction and forecasting crime rates with Gaussian processes. Technical report, Heinz College of Information Systems and Public Policy, Carnegie Mellon Univ., Pittsburgh, PA.

FLAXMAN, S., WILSON, A., NEILL, D., NICKISCH, H. and SMOLA, A. (2015). Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. In *International Conference on Machine Learning* 607–616.

FLAXMAN, S., CHIRICO, M., PEREIRA, P. and LOEFFLER, C. (2019a). Supplement to "Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ 'Real-Time Crime Forecasting Challenge'." DOI:10.1214/19-AOAS1284SUPPA.

FLAXMAN, S., CHIRICO, M., PEREIRA, P. and LOEFFLER, C. (2019b). Source code for "Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ 'Real-Time Crime Forecasting Challenge'." DOI:10.1214/19-AOAS1284SUPPB.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

GERBER, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decis. Support Syst.* **61** 115–125.

GORR, W. L. (2009). Forecast accuracy measures for exception reporting using receiver operating characteristic curves. *Int. J. Forecast.* **25** 48–61.

GORR, W. L. and LEE, Y. (2015). Early warning system for temporary crime hot spots. *J. Quant. Criminol.* **31** 25–47.

GORR, W., OLLIGSCHLAEGER, A. and THOMPSON, Y. (2003). Short-term forecasting of crime. *Int. J. Forecast.* **19** 579–594.

GROFF, E. and TANIGUCHI, T. (2019). Using citizen notification to interrupt near-repeat residential burglary patterns: The micro-level near-repeat experiment. *J. Exp. Criminol.* **15** 115–149.

GUTTORP, P. and GNEITING, T. (2005). On the Whittle–Matérn correlation family. National Research Center for Statistics and the Environment-Technical Report Series, Seattle, WA.

HEATON, M. J., DATTA, A., FINLEY, A. O., FURRER, R., GUINNESS, J., GUHANIYOGI, R., GERBER, F., GRAMACY, R. B., HAMMERLING, D. et al. (2019). A case study competition among methods for analyzing large spatial data. *J. Agric. Biol. Environ. Stat.* **24** 398–425.

HENNIG, P., OSBORNE, M. A. and GIROLAMI, M. (2015). Probabilistic numerics and uncertainty in computations. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **471** 20150142, 17. MR3378744

HUNT, J. M. (2016). Do crime hot spots move? Exploring the effects of the modifiable areal unit problem and modifiable temporal unit problem on crime hot spot stability. Ph.D. thesis, American Univ., Washington, DC.

JOHNSON, S. D., BOWERS, K. J., BIRKS, D. J. and PEASE, K. (2009). Predictive mapping of crime by ProMap: Accuracy, units of analysis, and the environmental backcloth. In *Putting Crime in Its Place* (D. Weisburd, W. Bernasco and G. Bruinsma, eds.) 165–192. Springer, Dordrecht.

KANG, H.-W. and KANG, H.-B. (2017). Prediction of crime occurrence from multi-modal data using deep learning. *PLoS ONE* **12** e0176244.

LEVINE, N. (2004). CrimeStat: A spatial statistics program for the analysis of crime incident locations, version 3.0. Technical report, Ned Levine and Associates/National Institute of Justice, Washington, DC.

LIU, H. and BROWN, D. E. (2003). Criminal incident prediction using a point-pattern-based density model. *Int. J. Forecast.* **19** 603–622.

LLOYD, C., GUNTER, T., OSBORNE, M. and ROBERTS, S. (2015). Variational inference for Gaussian process modulated Poisson processes. In *International Conference on Machine Learning* 1814–1822.

LOEFFLER, C. and FLAXMAN, S. (2018). Is gun violence contagious? A spatiotemporal test. *J. Quant. Criminol.* **34** 999–1017.

LUM, K. and ISAAC, W. (2016). To predict and serve? *Significance* **13** 14–19.

MAKRIDAKIS, S., SPILIOTIS, E. and ASSIMAKOPOULOS, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE* **13** e0194889.

MAY, A., BAGHERI GARAKANI, A., LU, Z. et al. (2019). Kernel approximation methods for speech recognition. *J. Mach. Learn. Res.* **20** Paper No. 59, 36. MR3960913

MILTON, P., COUPLAND, H. GIORGI, E. and BHATT, S. (2019). Spatial analysis made easy with linear regression and kernels. *Epidemics*. DOI:10.1016/j.epidem.2019.100362.

MITCHELL, S., POTASH, E., BAROCAS, S., D'AMOUR, A. and LUM, K. (2018). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. Available at arXiv:1811.07867.

MOHLER, G. (2013). Modeling and estimation of multi-source clustering in crime and security data. *Ann. Appl. Stat.* **7** 1525–1539. MR3127957

MOHLER, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecast.* **30** 491–497.

MOHLER, G. and PORTER, M. D. (2018). Rotational grid, PAI-maximizing crime forecasts. *Stat. Anal. Data Min.* **11** 227–236. MR3859025

MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. MR2816705

MOHLER, G. O., SHORT, M. B., MALINOWSKI, S., JOHNSON, M., TITA, G. E., BERTOZZI, A. L. and BRANTINGHAM, P. J. (2015). Randomized controlled field trials of predictive policing. *J. Amer. Statist. Assoc.* **110** 1399–1411. MR3449035

MØLLER, J. and RASMUSSEN, J. G. (2005). Perfect simulation of Hawkes processes. *Adv. in Appl. Probab.* **37** 629–646. MR2156552

MØLLER, J., SYVERSVEEN, A. R. and WAAGEPETERSEN, R. P. (1998). Log Gaussian Cox processes. *Scand. J. Stat.* **25** 451–482. MR1650019

O'HAGAN, A. (1992). Some Bayesian numerical analysis. In *Bayesian Statistics*, 4 (*Peñíscola*, 1991) 345–363. Oxford Univ. Press, New York. MR1380285

NATIONAL INSTITUTE OF JUSTICE (2017). Real-time crime forecasting challenge. Available at http://www.nij.gov/funding/Pages/fy16-crime-forecasting-challenge.aspx.

OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27.

PEASE, K. et al. (1998). *Repeat Victimisation*: *Taking Stock* **90**. Home Office Police Research Group, London.

PERRY, W. L., MCINNIS, B., PRICE, C. C., SMITH, S. C. and HOLLYWOOD, J. S. (2013). Predictive policing: The role of crime forecasting in law enforcement operations. Technical report, RAND Corporation, Santa Monica, CA.

PORTER, M. D. and REICH, B. J. (2012). Evaluating temporally weighted kernel density methods for predicting the next event location in a series. *Ann. GIS* **18** 225–240.

RAHIMI, A. and RECHT, B. (2007). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems* 1177–1184.

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

RODRIGUES, A. and DIGGLE, P. J. (2012). Bayesian estimation and prediction for inhomogeneous spatiotemporal log-Gaussian Cox processes using low-rank models, with application to criminal surveillance. *J. Amer. Statist. Assoc.* **107** 93–101. MR2949344

ROSSER, G. and CHENG, T. (2019). Improving the robustness and accuracy of crime prediction with the self-exciting point process through isotropic triggering. *Appl. Spatial Anal. Policy* **12** 5–25.

ROSSER, G., DAVIES, T., BOWERS, K. J., JOHNSON, S. D. and CHENG, T. (2016). Predictive crime mapping: Arbitrary grids or street networks? *J. Quant. Criminol.* **33** 569–594.

RUDIN, C. and USTUN, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces* **48** 449–466.

SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with Kernels*: *Support Vector Machines*, *Regularization*, *Optimization and Beyond*. MIT Press, Cambridge, MA.

SCHUTT, H. G. (1922). Advanced police methods in Berkeley. *Natl. Munic. Rev.* **11** 80–85.

SHIROTA, S. and GELFAND, A. E. (2017). Space and circular time log Gaussian Cox processes with application to crime event data. *Ann. Appl. Stat.* **11** 481–503. MR3693535

SNOEK, J., LAROCHELLE, H. and ADAMS, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems* 2951–2959.

SRIPERUMBUDUR, B. K., FUKUMIZU, K. and LANCKRIET, G. R. G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.* **12** 2389–2410. MR2825431

SUN, Y., LI, B. and GENTON, M. G. (2012). Geostatistics for large datasets. In *Advances and Challenges in Space-Time Modelling of Natural Events* 55–77. Springer, New York.

TADDY, M. A. (2010). Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime. *J. Amer. Statist. Assoc.* **105** 1403–1417. MR2796559

TEH, Y. W. and RAO, V. (2011). Gaussian process modulated renewal processes. In *Advances in Neural Information Processing Systems* 2474–2482.

WANG, X., GERBER, M. S. and BROWN, D. E. (2012). Automatic crime prediction using events extracted from Twitter posts. In *Social Computing Behavioral—Cultural Modeling and Prediction* 231–238. Springer, Berlin.

WEINBERGER, K., DASGUPTA, A., LANGFORD, J., SMOLA, A. and ATTENBERG, J. (2009). Feature hashing for large scale multitask learning. In *Proceedings of the* 26*th Annual International Conference on Machine Learning*. *ICML* '09 1113–1120. ACM, New York.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

S. FLAXMAN
DEPARTMENT OF MATHEMATICS
IMPERIAL COLLEGE LONDON
SOUTH KENSINGTON CAMPUS
LONDON SW7 2AZ
UNITED KINGDOM
E-MAIL: s.flaxman@imperial.ac.uk

M. CHIRICO
GRAB
9 STRAITS VIEW
MARINA ONE WEST TOWER #23-07/12
SINGAPORE 018937

P. PEREIRA
AMAZON, INC.
595 BURRARD STREET
VANCOUVER, BC, V7X 1L4
CANADA

C. LOEFFLER
DEPARTMENT OF CRIMINOLOGY
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: cloef@upenn.edu