A General Approach to Prediction and Forecasting Crime Rates with Gaussian Processes

Seth R. Flaxman
sflaxman@cs.cmu.edu

May 7, 2014

Carnegie Mellon University
Pittsburgh, PA 15213

Committee:
Daniel Neill, Chair
Alex Smola
Wilpen Gorr

Heinz College Second Paper

Abstract

We present a fully Bayesian spatiotemporal model for count data, which we use to forecast crime, in space and time, up to 12 weeks into the future. Our model fits a latent, smoothly varying relative risk surface using a Gaussian Process formulation. This relative risk surface is used as the mean in a Poisson likelihood for the observed weekly counts of crime by neighborhood. We use this model to assess the separate contributions of purely spatial and purely temporal predictors to our model's fit and forecast accuracy. We also consider the inclusion of a space/time interaction term. Our model is fully probabilistic, explicitly allowing us to characterize the uncertainty in all of our parameters, estimates, and forecasts. The main competitors for our method are univariate time series methods and heat maps (kernel-based intensity smoothing). As compared to time series methods, we model spatial dependence and variation through our relative risk surface. As compared to heat maps, our model enables temporal forecasts, with uncertainty intervals. We show that our model outperforms current methods. While we focus on the problem of forecasting, our model is equally suited to the problem of statistical inference and longer time periods; it could be used to answer questions like, how much did crime drop over the last decade in a city, and was this drop uniformly felt across all neighborhoods? Is there spatial or temporal variation in the amount of variance in crime rates? We conclude by discussing some practical approaches to speeding up inference with Gaussian Processes in moderately sized datasets.

1 Introduction

The last two decades has seen the collection and availability of large spatially and time-referenced crime datasets, and as a result researchers have focused on the possibility of short-term forecasting of crime [16]. The most widely used method in this area is to assume that "hot spots" (found with kernel intensity estimation) will persist in the short-term [11]. Other methods include extrapolating forecasts from univariate time series analysis [16], leading indicators models [6], and combinations, such as risk terrain modeling, in which kernel intensity estimates for various types of crime are combined [5]. Through commercial vendors and popular software packages, these models are being used in practice.

In the last few years, sophisticated modern spatiotemporal statistical models have been proposed for crime events [1, 18, 23]. While each of these methods was subjected to a small-scale evaluation on the problem for which it was designed, we are not aware of any larger evaluations or comparisons of these new methods to existing, deployed methods. This gap in the literature means not only that we do not know how promising these methods are, but also that there is little guidance for statisticians and computer scientists in developing new methods.

In this work, we draw on the geostatistical disease mapping literature to propose a new flexible framework based on Gaussian Processes for the modeling and short-term forecasting of crime in space and time with three goals in mind:

- Forecasting and evaluation: provide a formal statistical framework, fully characterizing uncertainty through forecast intervals, to allow comparisons with existing methods
- Spatial focus: produce visually interpretable heat maps forecasting crime intensity at a fine grained spatial and temporal resolution in the future
- Temporal focus: provide a modeling framework which is equally suited to long-term macro-level research on crime rates over time and short-term forecasting

In addition, our model is meant to be flexible and data-driven; it does not incorporate criminological theory on crime dynamics, rather focusing on the statistical problem of accurately modeling and forecasting crime counts in space and time. Thus, we argue that it is a reasonable starting point for evaluating future models which do incorporate crime dynamics or other sources of data. Our model is quite general and could be used to model the intensity of other spatiotemporal datasets.

Our model fits a latent, smoothly varying relative risk surface using a Gaussian Process formulation. This relative risk surface is used as the mean in a Poisson likelihood for the observed weekly counts of crime by neighborhood. We use this model to assess the separate contributions of purely spatial and purely temporal predictors to our model's fit and forecast accuracy. We also consider the inclusion of a space/time interaction term. Our model is fully probabilistic, explicitly allowing us to characterize the uncertainty in all of our parameters, estimates, and forecasts. Our model outperforms existing univariate time series and heat map-based methods.

2 Background

2.1 Kernel Intensity Estimation (Heat Maps)

Given the coordinates of the locations of crime incidents [2] treated as a point pattern, smoothing kernels can be used to estimate a spatially varying intensity. The technique is very similar to kernel density estimation, except that instead of an estimate of the density (which must be normalized), the estimate is of the (unnormalized) intensity—the number of crimes per square mile. Given points in space $\{s_1, \ldots, s_n\}$ the intensity function is defined as:

$$\lambda(s) = \lim_{ds \to 0} \frac{E(Y(ds))}{ds}$$

where $Y(ds)$ counts the number of points in a small region ds around s [12]. Given a smoothing kernel $k(r)$ and a bandwidth h, an estimate of the intensity function is:

$$\hat{\lambda}(s) = \sum_i \frac{1}{h} k\left(\frac{\|s - s_i\|}{h}\right)$$

Kernel intensity estimation is a popular technique because it is easy to understand and apply, and the resulting "heat maps" are appealing visual representations of a large, complex dataset. The heat maps of estimated intensity produced by kernel estimation are usually visually inspected for the presence of hot spots. A variety of choices must be made by the analyst in using the technique. The parametric form of the smoothing kernel and its bandwidth must be specified. The time period for which data is included is also an important choice. Methods have been proposed for data driven kernel selection, bandwidth selection [17], and also for performing edge corrections [9]. In comparing kernel intensity estimation to our methods, we use the implementation in the `density` function in the R package `spatstat` [4] and grid search to select our bandwidth.

Kernel intensity estimation is frequently used in a forecasting context, despite the fact that it contains no temporal information. The implicit assumption, then, is that current trends, especially the location of current hot spots, will continue in the near future. This might seem like a rather strong assumption, but at least two factors make it plausible: an important component of the spatial distribution of crime is chronic, with high-crime neighborhoods remaining high-crime for years if not decades [22] and as discussed in the next section the high degree of autocorrelation present in crime rates over time implies that a "no change" forecast is reasonably accurate in the short term.

2.2 Time Series Models

Gorr et al. [16] compared various univariate time series forecasting models, including random walk and a variety of exponential smoothing methods, to the naïve method in use by the police department: to forecast a certain month, use the observed counts from that month a year ago. The models considered had no spatial component, with each estimated separately for each location in the city. Seasonality was a major factor in most crime types considered. Time trends were only an important factor for simple assaults. In general, forecasting accuracy was higher in

precincts with higher observed crime counts; forecasting rare events was difficult. The main conclusion was that every univariate time series model outperformed naïve models. While the "no change" forecasting model discussed in the previous section did better than the naïve model based on predicting the counts from a year ago, all of the time series models were better than it as well. This suggests that there is much room for improvement over the method discussed in the previous section, of using heat maps as is for forecasting the short term. In terms of improving over existing time series methods, the fact that each time series was modeled separately seems to be a major drawback. Being able to appropriately "borrow strength" should improve forecasts especially for locations with low counts. Explicitly allowing for the interaction between space and time might also help, although this is by no means clear.

2.3 Gaussian Processes

We propose the use of Gaussian Processes (GPs) in a hierarchical Bayesian modeling framework as a spatiotemporal alternative to both time series and smoothing kernel models. In our framework, observations are counts modeled by a Poisson process whose intensity varies smoothly in space and time. This intensity surface has a natural interpretation as the relative risk. As compared to heat maps, our framework models temporal trends and thus naturally provides forecasts. As compared to univariate time series models, our framework models spatial trends. We provide a brief introduction to GPs. For a complete reference see [25].

A Gaussian process (GP) is a stochastic process where a realization of the process is a function $f(x)$. Thus a Gaussian process is a distribution over functions. We parameterize a GP by a mean function $\mu(x)$ and a covariance $k(x_i, x_j)$. Let us see how we draw a function:

$$f \sim \mathcal{GP}(\mu(x), k(x_i, x_j))$$

For a finite set of observation locations $x_1, \ldots, x_n$ we calculate a covariance matrix K where $K_{ij} = k(x_i, x_j)$. Then $(y_1, \ldots, y_n) \sim \mathcal{N}(\mu(\vec{x}), K)$, i.e. the observations y follow a multivariate Gaussian distribution with mean vector $\mu(\vec{x})$ and covariance K. Finally, we complete the specification by defining $f(x_i) := y_i$.

As an illustration we let $\mu = 0$ and $k(x_i, x_j) = exp(\|x_i - x_j\|^2)$ (the squared exponential covariance function). These parameters give a GP from which we can draw functions. In practice, here are the steps—we create a grid of points: $X = [-2, -1.9, \ldots, 1.9, 2]$ and calculate the covariance K. Now, draw Y from a multivariate Gaussian distribution with mean 0 and covariance K. This gives one draw of a function f from the GP. Three different draws are shown in Figure 1. In a Bayesian framework, these should be thought of as draws from the prior distribution over functions, before we've seen any data.

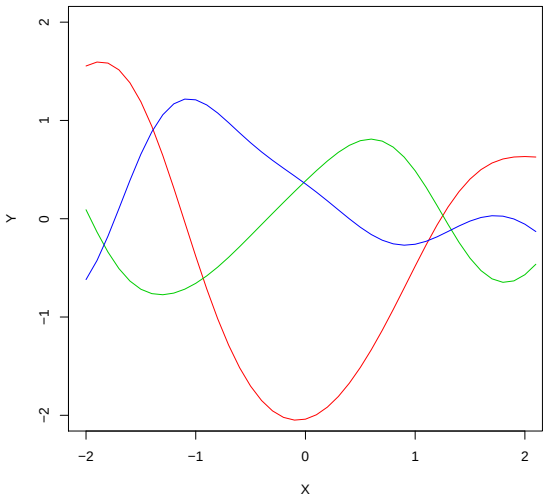


Figure 1: Three draws from a GP prior with mean 0 and RBF covariance function.

How do we update our prior given observations $Z = (X, Y)$? We start by specifying the joint distribution over both observed outputs (Y) and unobserved outputs (Y^*):

$$\begin{pmatrix} Y \\ Y^* \end{pmatrix} \sim \mathcal{N}(\mu(\vec{x}), K)$$

where we can calculate $K(x_i, x_j)$ for any pair of x's, observed or unobserved, i.e. :

$$K = \begin{pmatrix} K(X, X) & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{pmatrix}$$

Now, since we've observed (X, Y), we can find the conditional distribution using the properties of multivariate Gaussian distributions [25]:

$$Y^*|Y \sim \mathcal{N}(K(X^*, X)K(X, X)^{-1}Y, K(X^*, X^*) - K(X^*, X)K(X, X)^{-1}K(X, X^*)) \tag{1}$$

We give an illustration in Figure 2, where the observations $(-1, 1), (0, 0), (1, 1)$ are shown in black circles and 10 posterior function draws f^* are plotted. Notice that there is no uncertainty at the observed points.

In some cases, like modeling computer simulations, this noise-free behavior might be desirable, but for real data generated by nature we need to include an extra noise term. If we believe our noise is iid, we can use the following covariance function:

$$k(x_i, x_j) = \exp(\|x_i - x_j\|^2) + \sigma^2 I(i = j)$$

What does this extra variance σ^2 (called the "nugget" in geostatistics) do? It only appears when $i = j$, meaning that the diagonal of the covariance matrix has entries σ^2 instead of 1. If we use the same K as before, we have:

$$Y^*|Y \sim \mathcal{N}(K(X^*, X)(K(X, X) + \sigma^2 I)^{-1}Y, K(X^*, X^*) - K(X^*, X)(K(X, X) + \sigma^2 I)^{-1}K(X, X^*))$$

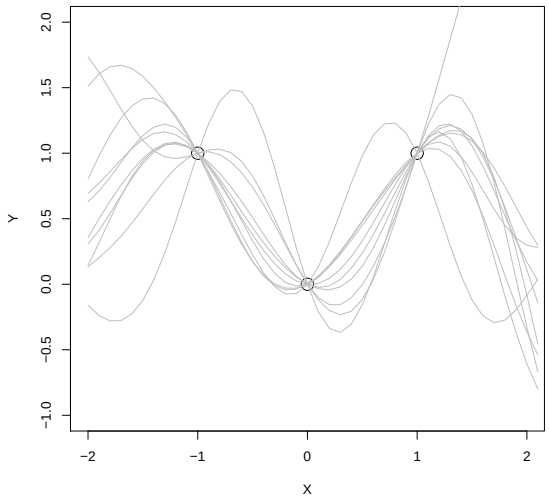


Figure 2: Draws from a noise-free GP posterior with mean 0 and RBF covariance function.

Because the noise term σ^2 is only used for observed data Y. If we use this prior, we can draw 10 posterior functions as before. In Figure 3 we have plotted these function draws. Notice that there is now uncertainty at the observed points.

2.4 Gaussian Processes for Time Series

GPs have been applied to time series data because they are well-suited to modeling non-iid observations. A variety of covariance functions are available to model various standard time series phenomena: e.g. trends, seasonality, periodic, and autoregressive components. The Matérn class of covariance functions encompasses the Ornstein-Uhlenbeck process and a continuous-time version of an AR(p) process, as discussed in [25]. Recent work has demonstrated the great flexibility of Gaussian Processes in handling time series data: [10] built a system to automate the search for combinations of covariance functions which performed quite well in modeling real datasets with non-stationary, periodic, and trend components.

2.5 Gaussian Processes for Spatial Data

The early development of Gaussian Process regression was in the context of geostatistics by Georges Matheron based on the work of Danie G. Krige and as a result the methods go under the name "kriging" (for a review of the history see [7]). Whereas the illustrative examples we have shown previously were one dimensional, i.e. we wished to predict the value of a time series at an index set of times, the extension to multiple dimensions is straightforward, provided a suitable covariance function can be specified. Common choices, which have been extensively studied in the geostatistics literature, include the squared exponential (called Radial Basis Function in the machine learning literature) and Matérn class of covariances functions.

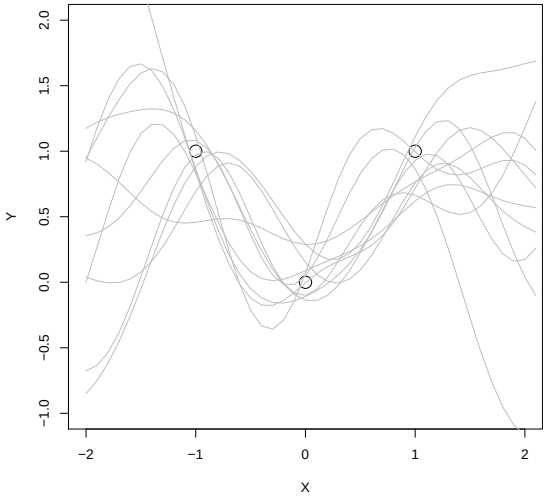


Figure 3: Draws from a GP posterior with mean 0 and RBF covariance function and noise variance $\sigma^2 = 0.05$.

2.6 Inference with Gaussian Processes

A variety of inference methods have been used with GPs. Once the hyperparameters of the covariance functions are specified, Equation 1 can be used to calculate the mean and variance of the predictions in closed form. Thus, inference consists in choosing the hyperparameters. With one or only a few hyperparameters, cross-validation is a reasonable approach. With more hyperparameters (corresponding a more richly parameterized covariance function) fully Bayesian sampling methods and MAP estimation are used.

3 Our Proposed Model

Our dataset consists of spatially and temporally referenced observations of crime counts, aggregated by neighborhood and week:

week (t)	neighborhood (s)	count
1	1	1
1	2	7
2	1	0
2	2	3
⋮	⋮	⋮

For convenience, we will refer to space-time regions $i = (s, t)$.

The key feature of a Gaussian Process is that it provides a prior for functions where any finite set of function values are distributed according to a multivariate Gaussian distribution. This means that it can be used to directly model continuous, real-valued data. We are dealing with count data y, which is positive and integer valued. This type of data is usually modeled with a

Poisson distribution:

$$p(y \mid \lambda) = \frac{\lambda^y}{y!} e^{-\lambda}$$

The only parameter that needs to be estimated is the underlying rate parameter λ. We borrow a successful approach from the disease mapping literature: we allow λ to vary in space and time. $\lambda(s,t)$ is thus a positive real-valued function, so we place a Gaussian Process prior not on the function itself, but on its log. Equivalently, we imagine a latent, real-valued function $f(s,t)$ with a Gaussian Process prior:

$$\lambda(s,t) = \exp(f(s,t))$$

This would complete our formulation, but again following the disease mapping literature we include a final fixed spatial term e_s giving the expected count at location s:

$$y_{s,t} \mid \lambda(s,t) \sim \text{Poisson}(\exp(f(s,t)) \cdot e_s)$$

This specification allows us to directly interpret $\exp(f(s,t))$ as the relative risk and $f(s,t)$ as the log-relative risk. When $f(s,t)$ is 0, $\exp(f(s,t)) = 1$, so $y_{s,t}$ has a Poisson distribution with mean equal to the fixed, expected count in location s, e_s. When f_i is positive, $\exp(f(s,t)) > 1$ so $y_{s,t}$ follows a Poisson distribution with an elevated mean, that is, greater than e_s, and when $f(s,t)$ is negative the mean is reduced. Conditional on the relative risk surface f, observed counts are independent, so the likelihood factors:

$$p(y \mid f) = \prod_{s,t} \text{Poisson}(y_{s,t} \mid \exp(f(s,t)) \cdot e_s)$$

We model the latent surface f by placing a Gaussian Process prior with mean 0 and covariance K on it:

$$f \sim \mathcal{GP}(0, K)$$

All of the modeling work is done with respect to the covariance function K, which should be interpreted as giving the dependence structure, in space and time, between observation locations (s,t). We combine spatial and temporal variation as follows: given a spatial covariance function $k_s(s,s')$ and a temporal covariance function $k_t(t,t')$ we can specify an additive covariance function:

$$K(i,i') = k_s(s,s') + k_t(t,t')$$

We might wish to make this simple additive model more complex by considering a joint covariance k_{st} over space and time. Separable space-time covariance functions are easily constructed by multiplying spatial and temporal covariance functions: $k_{st}((s,t),(s',t')) = k(s,s')k(t,t')$[1]

Next, we include a periodic temporal term $k_p(t,t')$, to account for seasonal variation. We adopt the following parameterization [25]:

$$k_P(t,t') = \exp\left(-\frac{2\sin^2\left(\frac{(t-t')\pi}{52}\right)}{\ell^2}\right)$$

[1] Recent research has focused on formulating non-separable space-time covariance functions [8, 15], but we do not consider these in this work.

where we use a fixed period of 52 weeks.

Our final covariance structure is as follows:

$$K((s,t),(s',t')) = k_s(s,s') + k_t(t,t') + k_{st}((s,t),(s',t')) + k_P(t,t')$$

where:

- $k_s(s,s')$ is a Matérn covariance function with $\nu = 3/2$, length-scale ℓ_s and variance σ_s^2.
- $k_t(t,t')$ is a squared exponential (Radial Basis Function) covariance function with length-scale ℓ_t and variance σ_t^2.
- $k_p(t,t')$ is a periodic covariance function with period 52 and parameterization as given above.
- $k_{st}((s,t),(s',t')) = k_s(s,s')k_t(t,t')$ is a separable space-time covariance function with periodic time component parameterized as $k_s \cdot k_p$ with a single variance σ_{st}^2 and separate length-scales for space and time.

Throughout we use a Student's t-distribution with mean $\mu = 0$, scale $\sigma^2 = 1$, and degrees of freedom $\nu = 4$ as the prior distribution for each parameter [24]:

$$p(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi\sigma^2}}\left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}$$

As compared a Normal distribution with mean 0 and large variance (one of the traditional choices of "uninformative" priors), the Student's t-distribution, which has heavier tails than a Normal distribution, has been shown to be a more widely useful, weakly informative prior [14, 20]. The shape of this distribution is shown in Figure 4.

To learn the hyperparameters we consider MAP estimation, grid integration, and MCMC.

3.1 Assessing Our Model

After fitting our model, we use it to make in-sample predictions (with credible intervals) and out-of-sample forecasts (with forecasting intervals). We calculate and report the mean squared error (MSE) of these predictions and forecasts. One of our motivating substantive questions is to understand the variation in crime rates, i.e. how much of our model's predictions are being driven by our predictors for space, time, space/time, and covariates? Using our probabilistic framework, we can assess the contributions of the various pieces of our model to the final fit. Recall that we are fitting a latent relative risk surface with an additive covariance structure. We can decompose the covariance structure and use the parameters we learned for each separate covariance function to make a prediction for the relative risk surface f_j corresponding to covariance function k_j [13, p. 506]. Notice that the separate log-relative risk surfaces f_j sum to the final predicted log-relative risk surface, so we can consider each surface as contributing additively to an increase or a decrease in log relative risk, which varies in space or time. Exponentiating, each surface contributes multiplicatively to an increase or decrease in relative risk; all the surfaces are multiplied together to obtain the final relative risk surface.

We can convert this relative risk prediction to a prediction for the counts, i.e.: we use each separate f_j to calculate the predicted number of counts at location (s,t) as $\hat{n}_{s,t} = \exp(f_j(s,t)e_s)$.

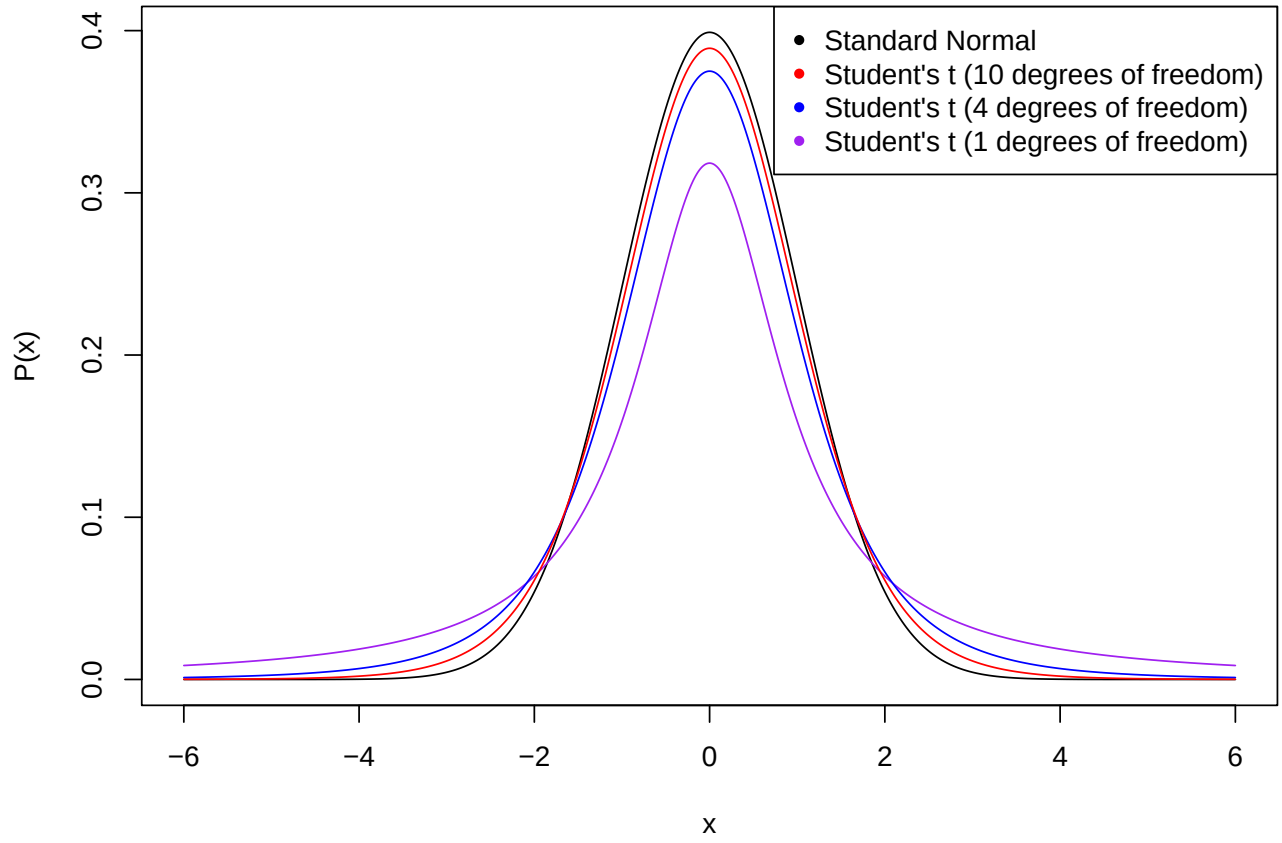


Figure 4: The distribution function of the Student's t distribution is compared to a Normal $\mathcal{N}(0, 1)$ distribution, varying ν, the number of degrees of freedom. As ν increases, the Student's distribution approaches the Normal distribution.

We performed graphical posterior predictive checks by inspecting the residuals $n_{s,t} - \hat{n}_{s,t}$ to look for remaining structure in the error. To address the question of what is driving our model's predictions, we can calculate the MSE for each separate f_j, and ask which terms of our model improve the model's predictions the most. An alternative way to assess which terms of our model are important is by focusing on forecast accuracy. We conduct the same analysis for out-of-sample counts of crimes.

3.2 Evaluating our Framework

We compare our results to the kernel intensity estimation approach described in Section 2.1, Holt exponential smoothing (the best univariate time series method in [16]) and to an AR(1) time series. The most widely used tool by police departments is Kernel Density Estimation (KDE). This tool is applied in a variety of ways in practice, but for evaluation we adopt the following straightforward approach: using the last W weeks of data (where W is a parameter chosen from the data) for the event of interest, smooth the locations of this event with KDE to obtain intensity estimates for each neighborhood, and predict that future crime counts by neighborhood will remain constant, up to 12 weeks into the future.

4 Experimental Results

The City of Chicago makes geocoded, date-stamped crime data publicly available through its data portal[2]. Chicago is divided into 77 community areas, which corresponds to a neighborhood or group of neighborhoods, as shown in Figure 5. Thefts during the first two weeks of January 2011 are shown on a map in Figure 6.

We downloaded the crime data, aggregated it into counts by type of crime, community area and week of the year. As an exemplar, we chose crimes coded as theft, a property crime which includes pick-pocketing, retail theft, etc. Burglary, a related property crime, implies breaking and entering, while robbery is theft accompanied by violence or the threat of violence (meaning it is categorized as a violent crime rather than a property crime) [3]. We wanted a crime type that was relatively frequent (very sparse events pose further modeling challenges, which we discuss in the conclusion) and showed interesting spatial and temporal patterns. Theft is very common in Chicago's central business district, the Loop, and has a marked seasonal pattern, peaking in the summer.

We use the following strategy to estimate e_s, the expected number of thefts in each neighborhood: we find the weekly average city-wide count of thefts in our entire dataset (1,387 thefts per week) and divide by the population of Chicago (2,718,590) to find a theft rate of 5.1 per 1,000 people. Then for each neighborhood with population p_s we calculate $e_s = 0.00051 \cdot p_s$.

All of our models were fit using the GPstuff package in matlab [24]. GPstuff implements MAP, grid integration, and MCMC. To reduce computational burden while we were developing our models, we fit various submodels and then expanded them, sometimes fixing the hyperparameters learned in the submodels. Another method we used to reduce computational burden

[2] http://data.cityofchicago.org

[3] http://www.ucrdatatool.gov/offenses.cfm

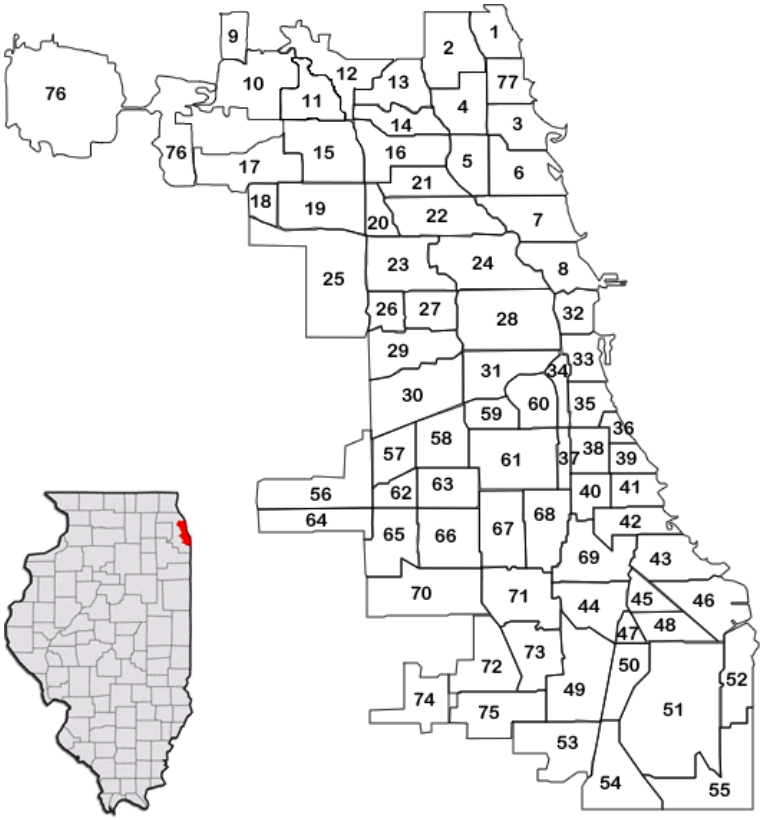


Figure 5: Community areas in Chicago. Source: [3]

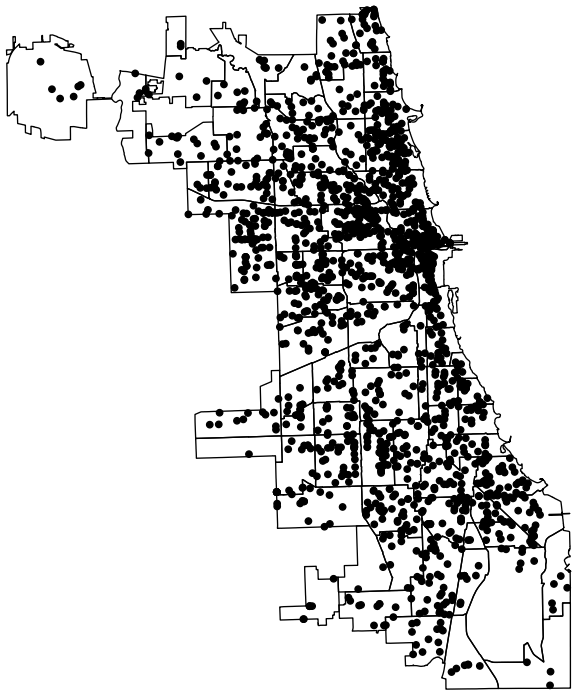


Figure 6: Map of thefts for the first two weeks of January 2011

was the Fully Independent Conditional (FIC) approximation [21] where latent inducing inputs are used and the very expensive covariance matrix updating is only performed at these pseudo inputs.

4.1 Long-term Time Trends

We started by considering the time period from January 2004 to December 2013. Ignoring spatial variation, we used a covariance function composed of the sum of periodic, exponential, and linear covariance functions. We fit our model to data from January 2004-December 2012 and predicted all of 2013. The fit, forecasts, and components of our model are shown in Figures 7 and 8. Our forecasts for an entire year fit the data quite well, suggesting that a large degree of variation in crime rates is explained by long-term trends composed of a linear and a periodic trend. These initial results suggest that our model is well-suited to modeling long-term trends. We achieve full coverage with our 95% uncertainty forecasting intervals. On an absolute scale, our forecasting errors range from -146 to 144. The mean relative forecasting error is 4.4%. Based on the linear component of our model (shown in Figure 8), there was an overall reduction in weekly thefts of 253 from January 2004 until December 2012.

4.2 Short-term Forecasting in Space and Time

For the remainder of our evaluation, we focus on short term forecasting in space and time. We consider data from 2011-2013 as training data, and leave out the last 12 weeks of data in 2013 for forecasting. First, we fit a purely spatial model to the average number of thefts across Chicago within the training data time period. In Figures 9a and 9b we compare the map of observed relative risk of theft (as compared to population) to our model's predicted relative risk of theft.

Next, we fit our full model, using MAP with scaled conjugate gradient descent. We show the results aggregated as a time series, predicting city-wide counts, in Figure 10. The fit of our model to the in-sample data is quite good, as are the forecasts for the last 12 weeks of 2013. Overall, the mean squared forecasting error of our model is 25.81, and the mean squared prediction error (in sample) is 23.32.

In Table 1, we report the MSE from making predictions and forecasts with the various components of our model. The baseline we compare to is assuming a constant relative risk, i.e. making predictions that are constant in time and are based solely on population density (we call this expected count e_s). We also calculate a version of R^2 which we call reduction in variance:

$$1 - \frac{\sum_{s,t}(n_{s,t} - \hat{n}_{s,t})^2}{\sum_{s,t}(n_{s,t} - e_s)^2}$$

The numerator calculates the sum of squared errors given predictions $\hat{n}_{s,t}$. The denominator calculates the sum of squared errors from assuming a constant relative risk. The MSE for the in-sample predictions from assuming a constant relative risk of 1 is 204 and the MSE for the out-of-sample forecasts is 179.

In Figure 11a and Figure 11b, we compare a map of our average weekly predictions of theft counts to the observed weekly counts for in sample data. In Figures 11c and 11d we show the

	Predictions		Forecasts	
Component	Reduction in variance	MSE	Reduction in variance	MSE
Periodic	-4.0%	212	1.0%	177
Time	0.3%	203	0.1%	178
Space-Periodic	1.7%	200	2.8%	174
Space	73.1%	55	56.4%	78
Combined	88.5%	23	85.5%	26

Table 1: Forecast and predictions of various components of our model as compared to the full model ("Combined"). The reduction in variance column is analogous to R^2 in a linear model: it is calculated as one minus the ratio between the residual sum of squares and the sum of squares from assuming a constant relative risk of 1, i.e. it is meant to give some idea of how much variance is "explained" by the component of the model. MSE stands for mean squared error. The MSE for the in-sample predictions from assuming a constant relative risk of 1 is 204 and the MSE for the out-of-sample forecasts is 179. Spatial variation accounts for most of the improvements in accuracy and reduction in variance.

same maps for the out-of-sample forecasts and the observed data. The same results are shown as scatterplots in Figures 12. In all cases, the predictions and forecasts are exceptionally accurate, with Spearman correlations above 0.97.

In Figure 13a we show the fit of our model to Austin, a neighborhood in Chicago (community area 25), which is the largest community area by population. In Figure 13b we show the fit of our model to the Loop, Chicago's central business district. In Figure 14a we show the temporal variation due to the various components of our model for Austin and in Figure 14b we show the Loop.

4.3 Comparison with other methods

For the comparison with kernel intensity estimation, we used grid search to evaluate various amounts of lagged data and bandwidths and selected the model with the best forecasting accuracy, a liberal approach to model selection which could overestimate accuracy. The best model used 5 weeks of previous data and a kernel with bandwidth 1000 feet and obtained a mean squared forecasting error of 47.70. In Figure 15 we compare the intensity as estimated by smoothing kernels to the intensity estimated with a Gaussian Process. Instead of aggregating to community areas, we created a regular grid, and aggregated counts to this grid, then fit these counts with a Poisson parameterized the same way.

We also compared to fitting an AR(1) model separately to each neighborhood. The mean squared forecasting error was 37.98. A Holt-Winters exponential smoothing performed even worse, achieving a mean squared forecasting error of 46.99. The comparison between our forecasting results and these baselines is shown in Figure 16, where we also show how the forecasts deteroriate over time. A paired t-test shows that the difference between MSE for each forecasted neighborhood-week for our method and its closest competitor AR(1) was statistically significant ($p \leq$ 1.3e-08) even after excluding the very poor forecast at the end of December.

5 Conclusion

We presented a general framework for the statistical modeling of spatiotemporal count data. There is nothing special about crime events, and early experiments using our methods to forecast 311 (calls for non-emergency services like potholes) have been promising. In the application of this framework, we made a series of modeling decisions which would be worth exploring in more detail in future work. For a predictive policing application, police beats or census blocks might be more appropriate than community areas. Shorter time windows, and even taking into account time of day would also be interesting. The use of spatiotemporal leading indicators—other crime types, weather patterns, other events measured online or offline—might provide measurable improvements in our forecasts. The use of a Poisson likelihood should be revisited for other types of crime or events: it is straightforward to use a Negative Binomial model to handle inflated variance. In the case of zero-inflation, that is, when counts are often zero, Gaussian Processes for zero-inflated Poisson, Binomial, and Hurdle models have been explored [24]. Our framework has much in common with Log Gaussian Cox Process models for point processes, and it would be quite useful to explicitly compare to that framework [19].

The overall forecasting performance of our model was much better than the competitors we considered, but a fuller evaluation, and more automatic model search techniques, would be needed before recommending that city governments replace heat maps with Gaussian Processes. This work does strongly support the literature suggesting that we can do better than simply assuming that patterns in heat maps will persist in the future—as is currently widely assumed in the field. A full comparison between our method and recent reported results using sophisticated statistical models [1, 18, 23] is also needed. Our framework does incur a significant computational burden. There is much room for the further development of approximation techniques for Gaussian Processes and new formulations of models for fitting spatiotemporal data, and our framework and evaluation, based entirely on publicly available data and source code, should serve in the future as a baseline for comparison.

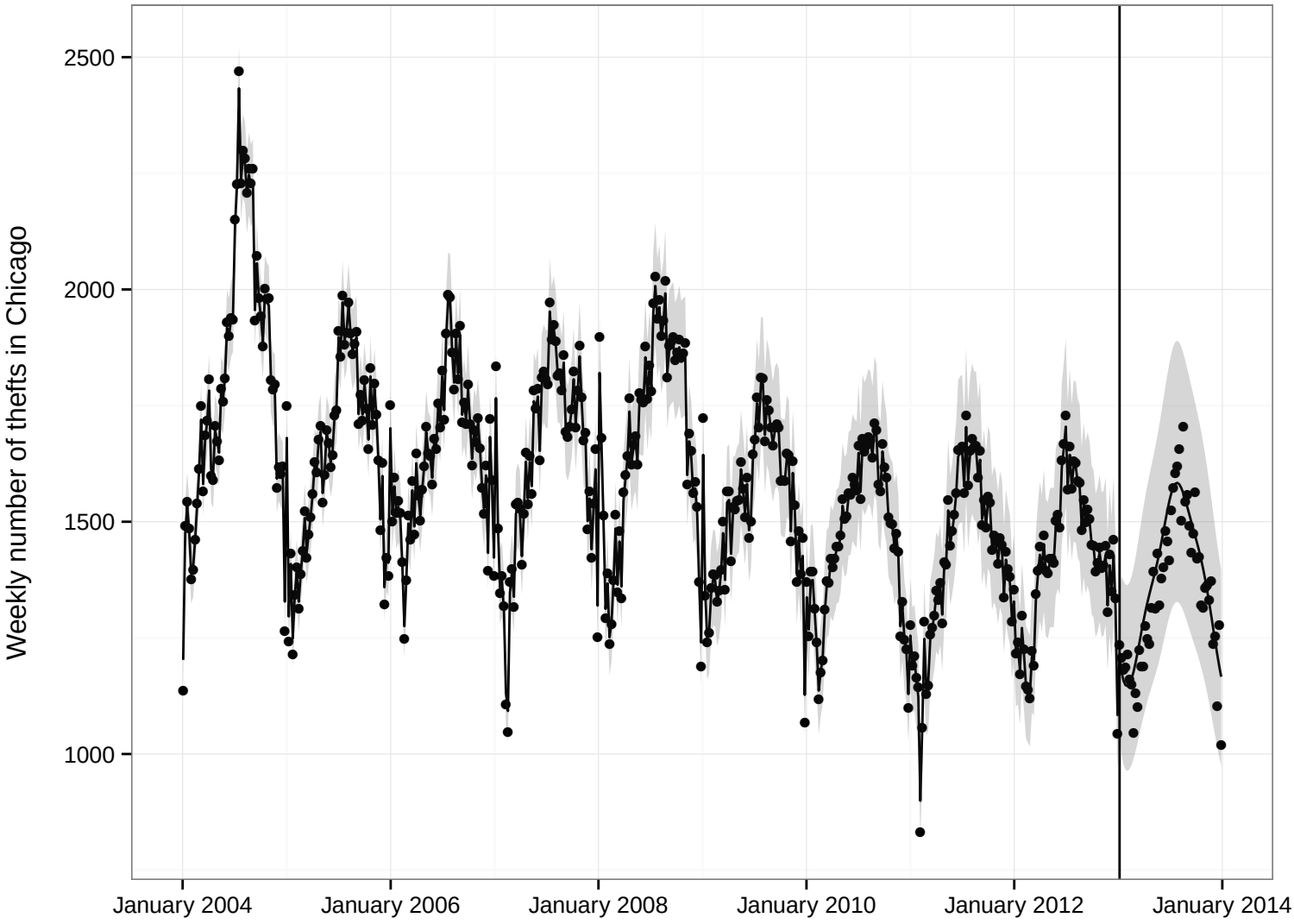


Figure 7: Weekly time series of total number of thefts in Chicago. Black dots are observed counts, and the black line is the model predictions. The model, consisting of a periodic term, linear term, and an exponential term, was fit to the aggregate data shown. Forecasts are shown for all of 2013 (to the right of the black vertical line). 95% uncertainty intervals are shown in gray.

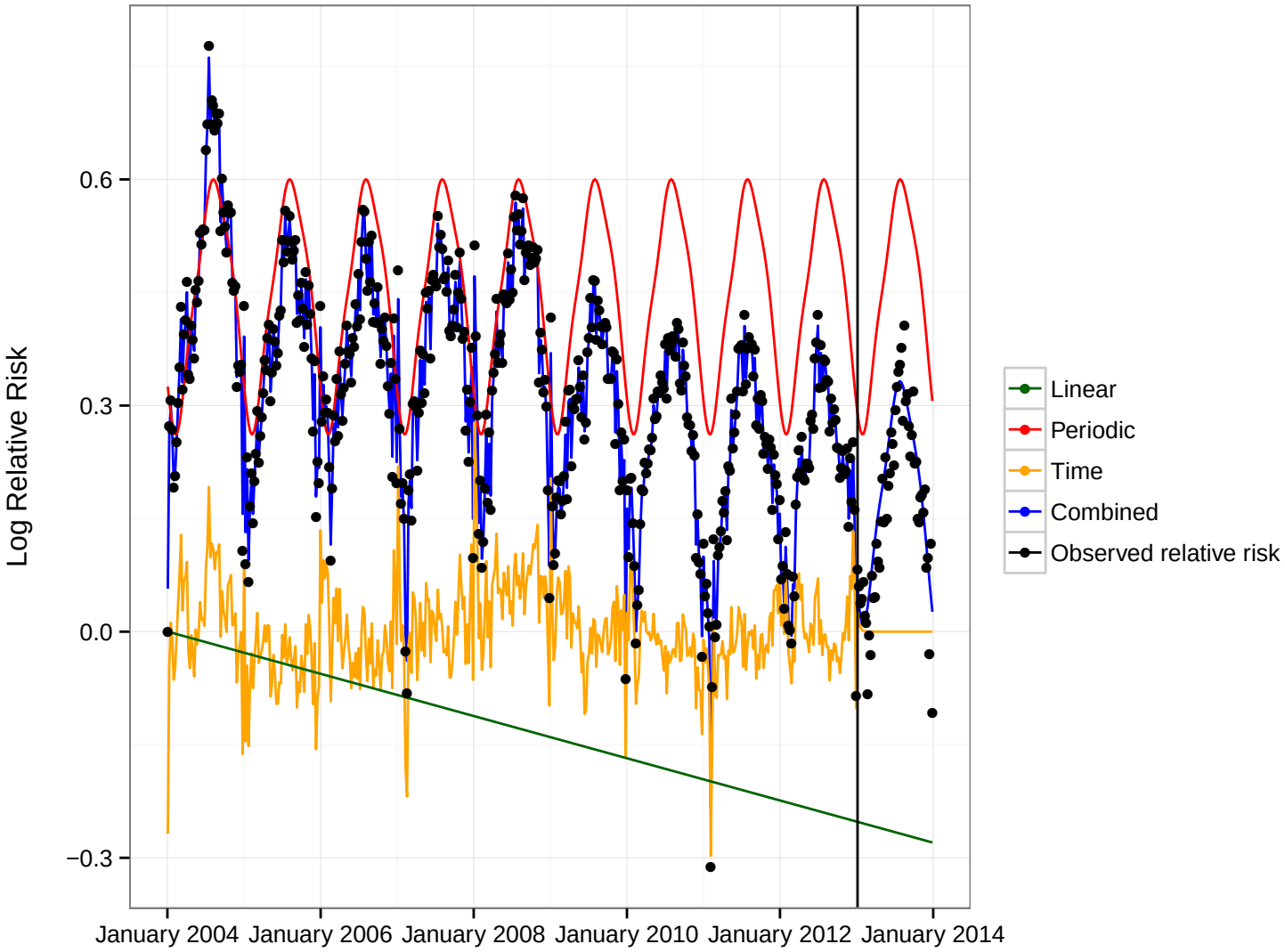
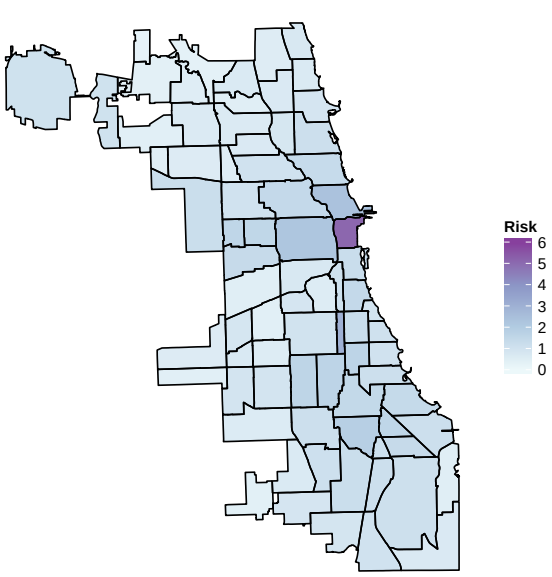
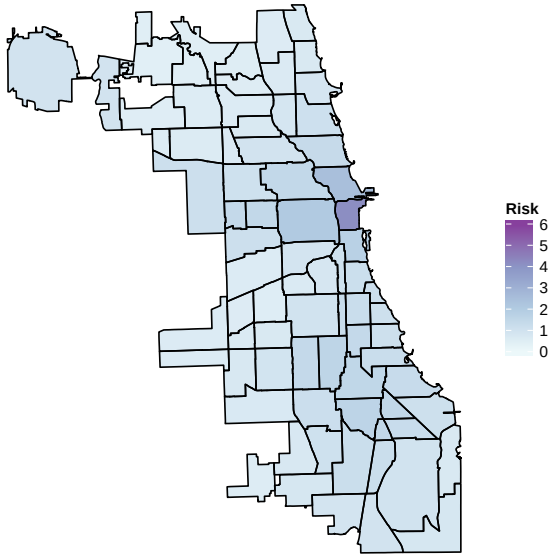


Figure 8: The time series version of our model fit to weekly citywide theft data from 2004 to 2012 (data is shown on a log relative risk scale), with forecasts made for 2013. The fit of the model is in blue, with the various contributions of the linear, periodic, and squared exponential (time) covariance functions shown.



(a) Observed relative risk of theft across Chicago



(b) Predicted relative risk from a purely spatial model for theft across Chicago

Figure 9

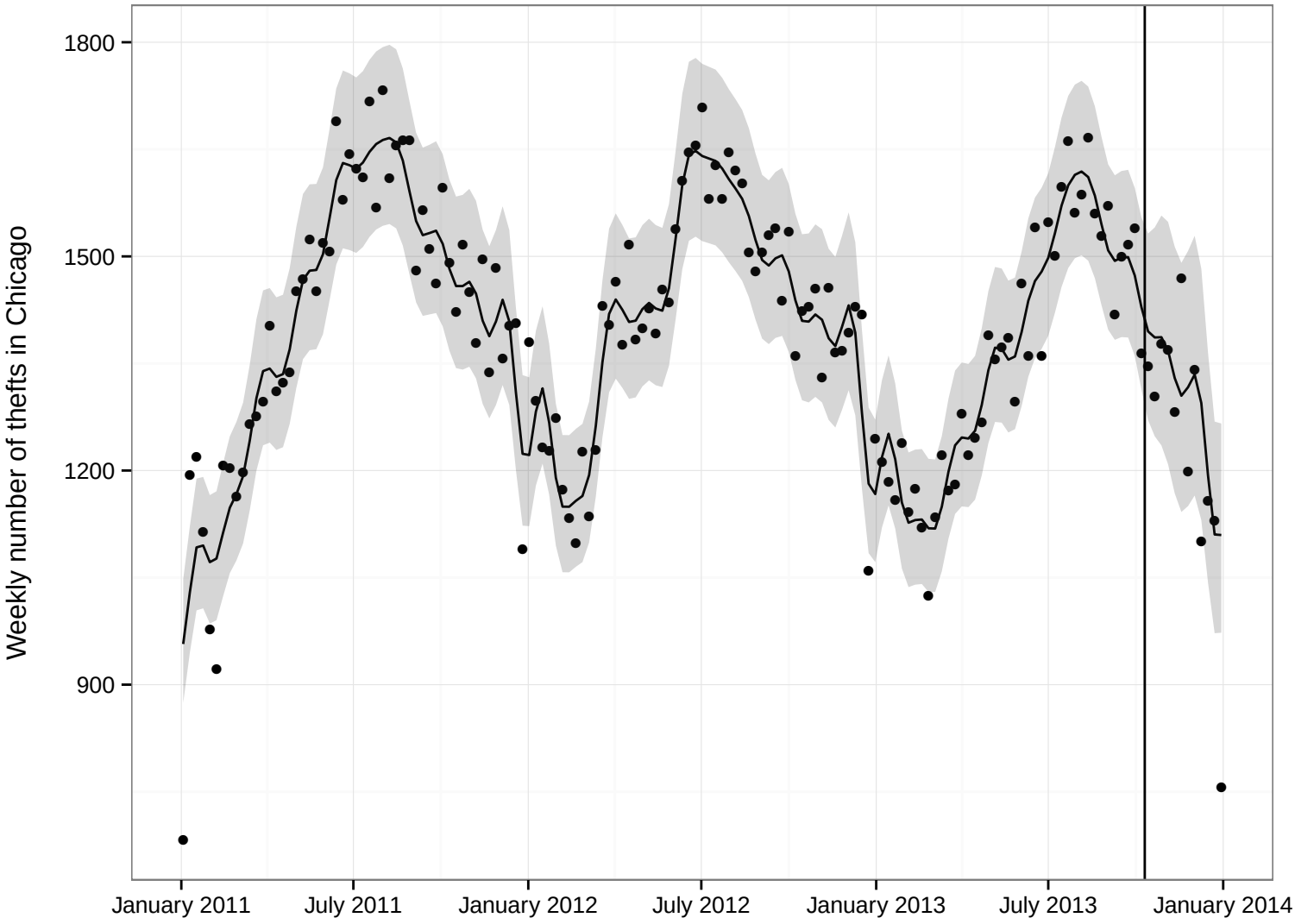
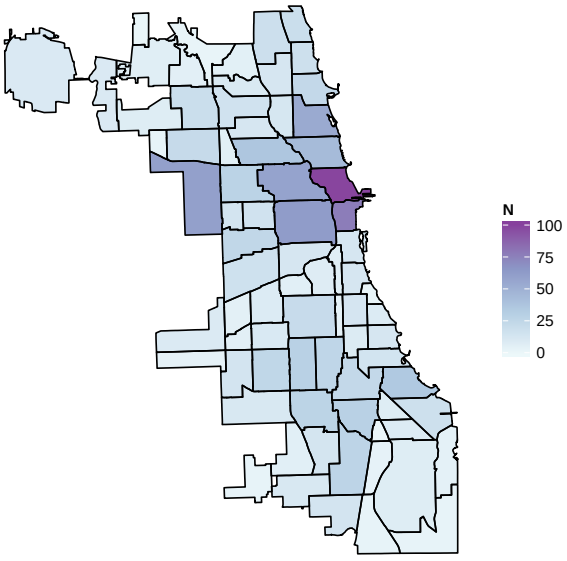
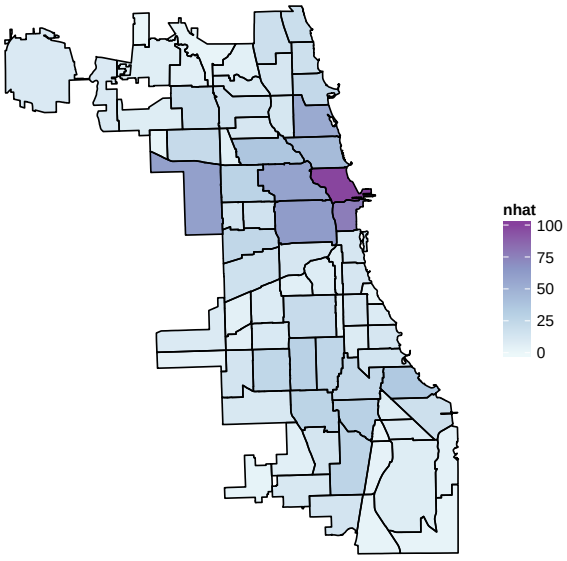


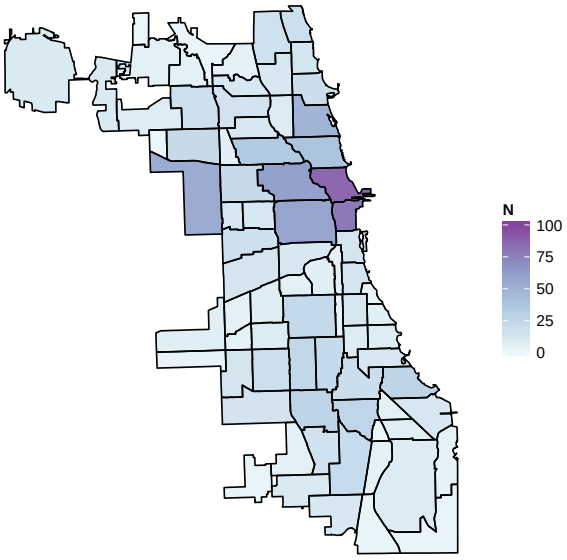
Figure 10: Weekly time series of number of thefts in Chicago for the full spatial/temporal model fit. Black dots are observations, the black line is the model predictions. Model forecasts are shown for the last 3 months of 2014 (to the right of the black vertical line). 95% uncertainty intervals are shown in gray.



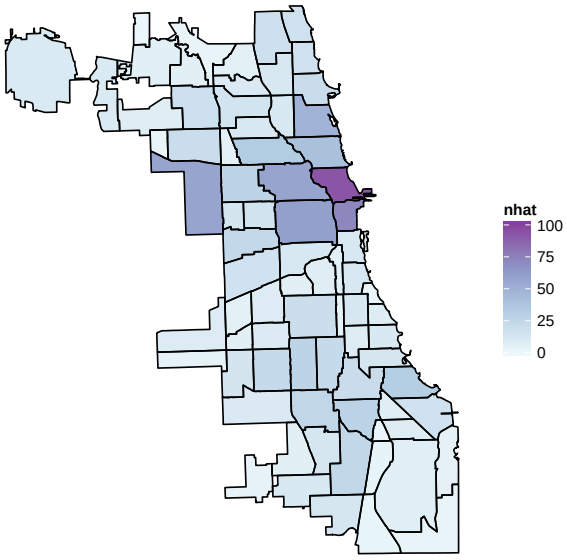
(a) Average weekly number of thefts for January 2011-September 2013



(b) Predicted weekly number of thefts for January 2011-September 2013

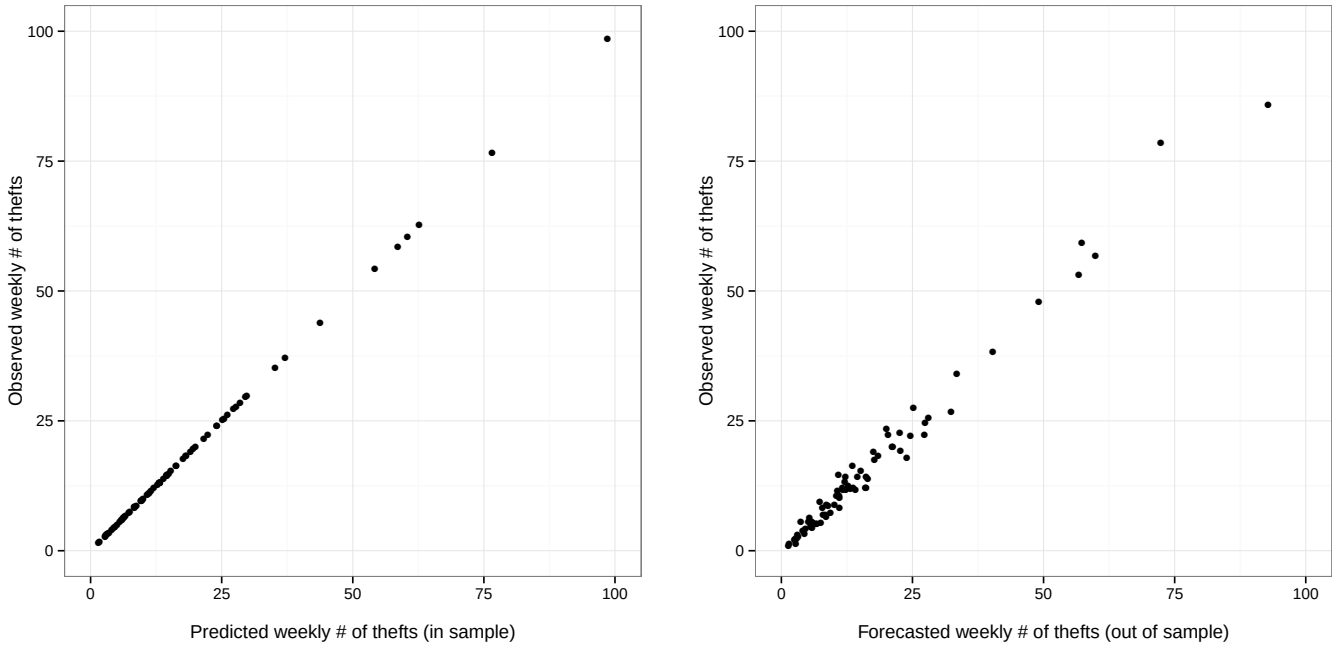


(c) Average weekly number of thefts for October 2013-December 2013



(d) Forecasted weekly number of thefts for October 2013-December 2013

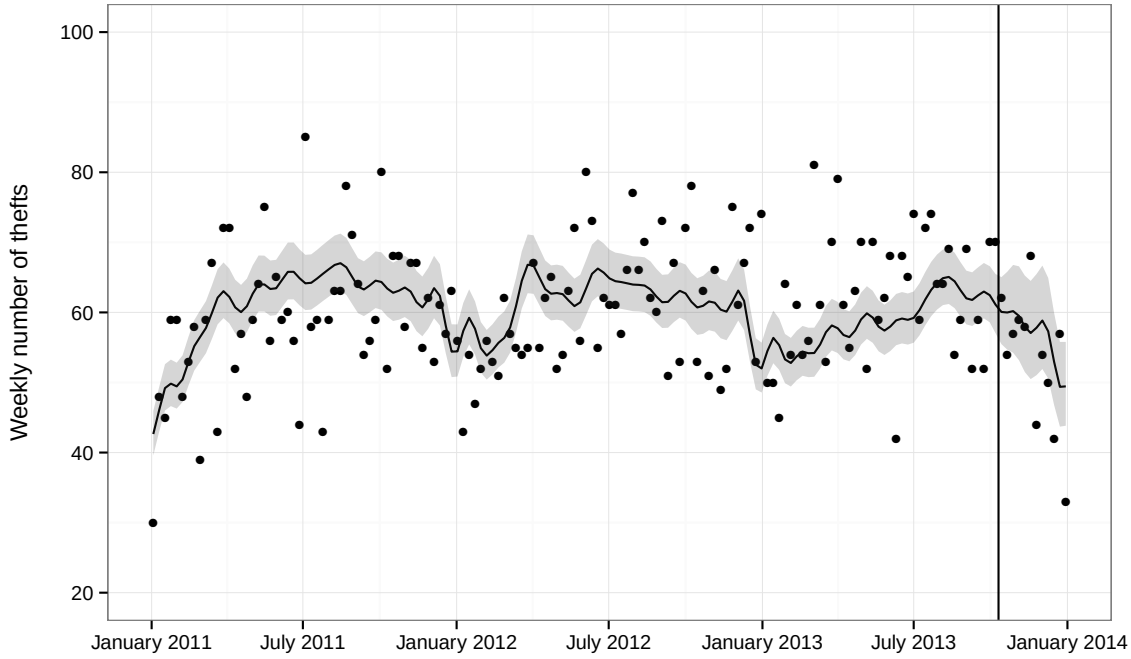
Figure 11: In sample predictions in 11b match in sample observations in 11a. The patterns do not change much for out of sample observations in 11c, suggesting that carrying forward static predictions will provide a reasonably good forecast.



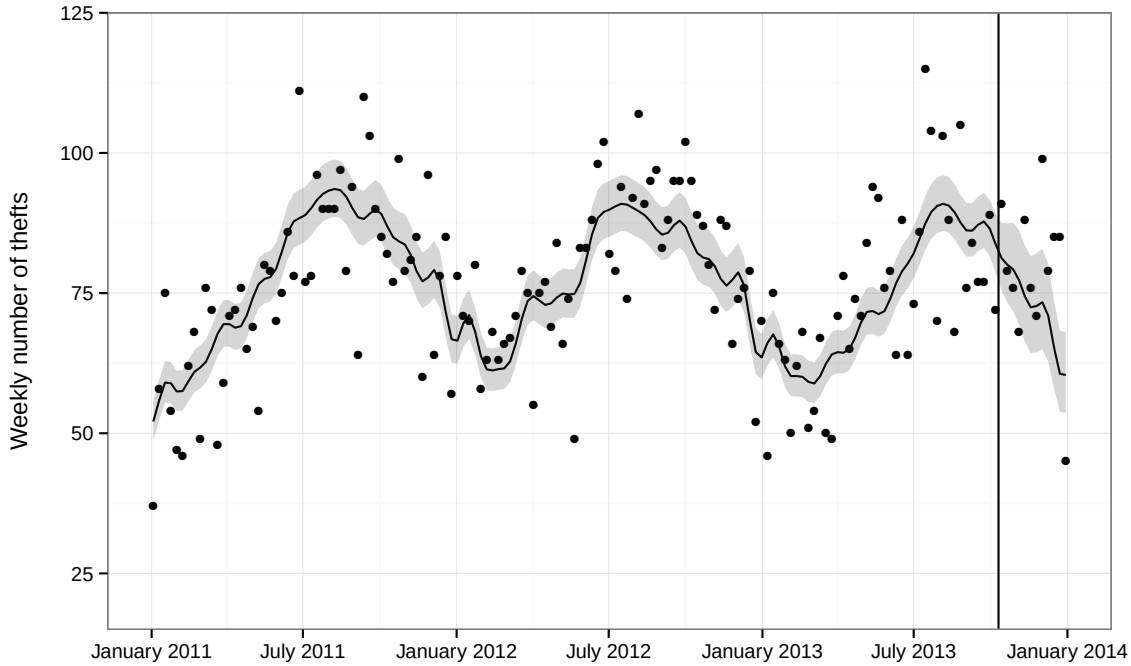
(a) Average weekly number of thefts for January 2011-September 2013

(b) Forecasted weekly number of thefts for October 2013-December 2013

Figure 12: The same data as in Figure 11. In 12a the Spearman correlation is 1.000 and in 12b the Spearman correlation is 0.977.

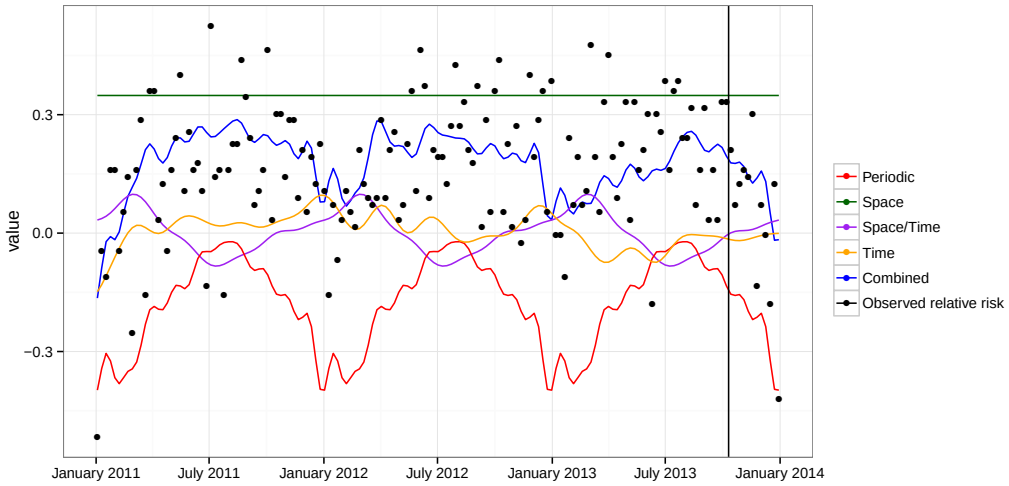


(a) Weekly number of thefts in the Austin neighborhood of Chicago

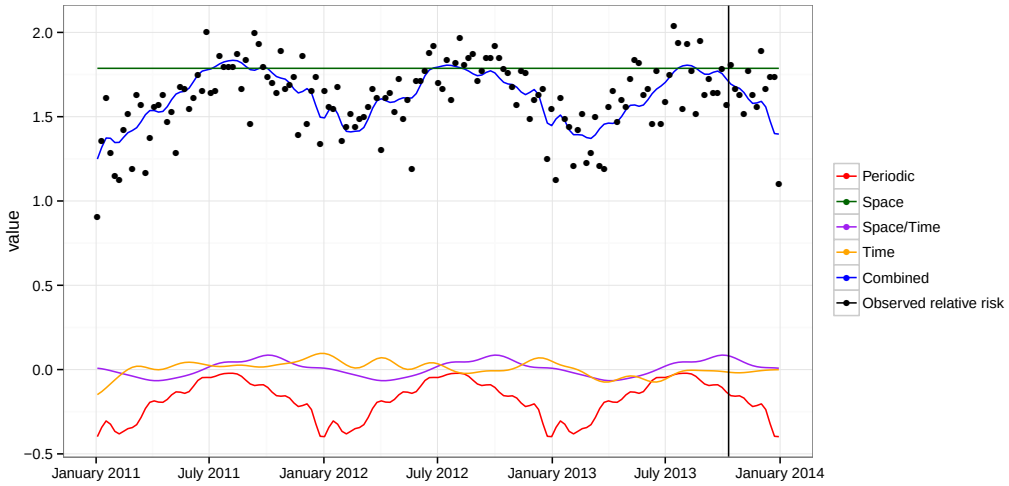


(b) Weekly number of thefts in the Loop (downtown business district) of Chicago

Figure 13: The out of sample forecasting accuracy of our model is quite good. At the neighborhood level, there is a lot of variance in the data. Our model could do a better job of capturing this variance, as observations currently fall outside of the 95% uncertainty intervals of our model.

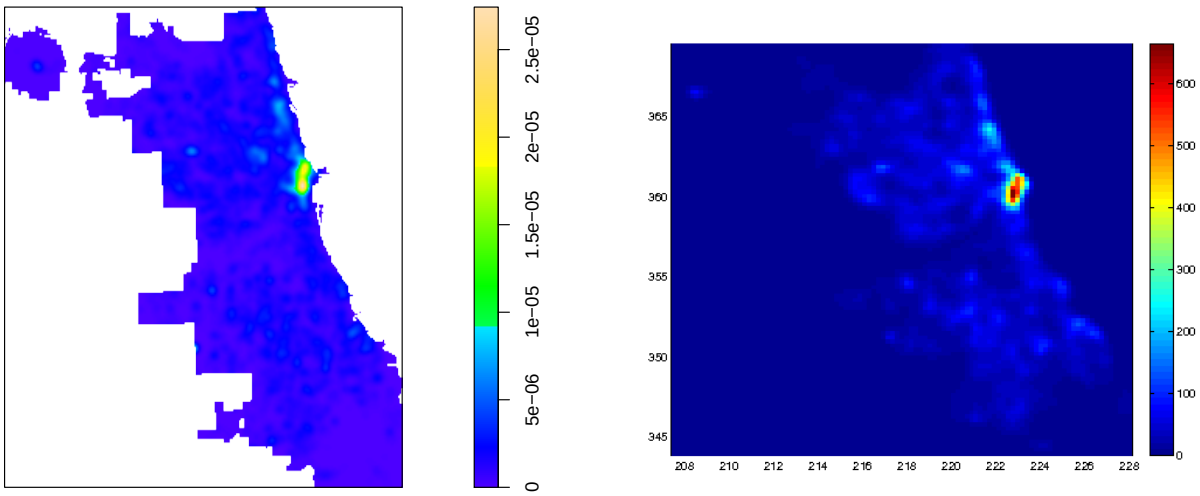


(a) The Austin neighborhood of Chicago



(b) The Loop (downtown business district) of Chicago

Figure 14: Predictions for the Austin neighborhood of Chicago 14a and the Loop 14b with each of the various additive components of our model shown on a log-relative risk scale. (On a relative risk scale the components would be composed by multiplication.) Since the plot is over time, the spatial component does not vary, but stays constant at relative risk of 1.4 for Austin and 6.0 for the Loop, meaning that compared to its population, the risk of theft in Austin is slightly elevated and it is very elevated (the most in Chicago) for the Loop, probably explained by the high concentration of targets during work hours. Another factor is the way in which the relative risk is calculated is relative to the expected number of thefts, which is based on population; the Loop has a low residential population. In both plots, the periodic component (which is non-spatial) reaches a maximum relative risk of 1.1 and a minimum relative risk of 0.9. The purely temporal component, which picks up unexplained temporal variation, reverts to a relative risk of 1 (log relative risk 0) as the forecasts move into the future. The space/time component, which is separable, is composed of a periodic time covariance function multiplied by a spatial covariance function, so the periodic nature of the trend varies across Chicago. Austin and the Loop are not nearby and the space/time trend is indeed different between the two.



(a)

(b)

Figure 15: In 15a we show the kernel-smoothed intensity estimate of thefts in Chicago for the 5 week period before September 1, 2013. In 15b, we use a Gaussian Process to fit the same data, fitting a Gaussian Process to the intensity surface by creating a fine grid and counting the number of thefts within each grid cell.

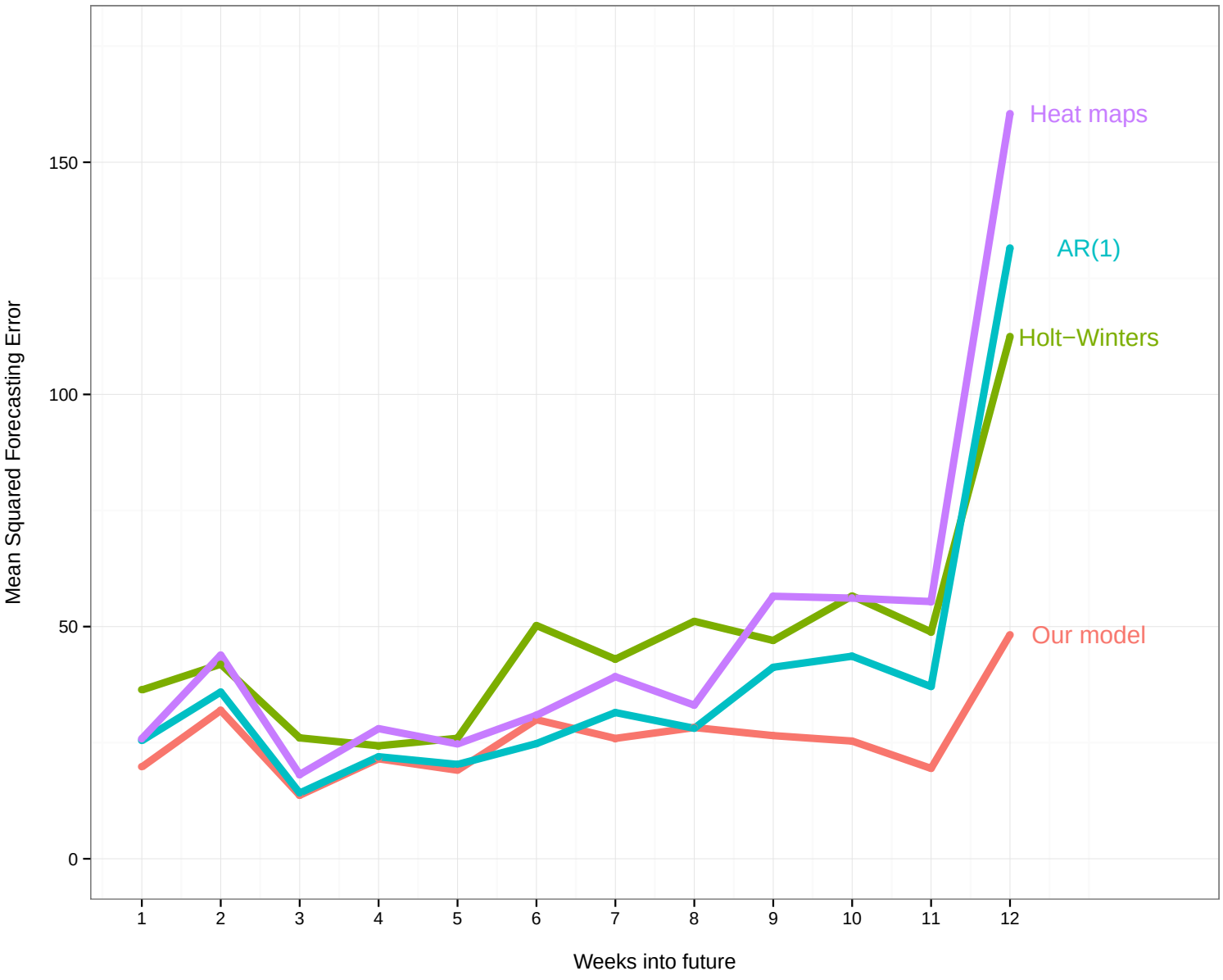


Figure 16: Forecasting accuracy of each of the methods we considered. The difference between our method and the others is statistically significant. The forecast accuracy for each method gets worse as we get farther out of sample. For the very last observation, our method is much better than the others.

Bibliography

[1] Sivan Aldor-Noiman, Lawrence D Brown, Emily B Fox, and Robert A Stine. Spatio-temporal low count processes with application to violent crime events. *arXiv preprint arXiv:1304.5642*, 2013. 1, 5

[2] Luc Anselin, Jacqueline Cohen, David Cook, Wilpen Gorr, and George Tita. Spatial analyses of crime. *Criminal justice*, 4:213–262, 2000. 2.1

[3] Jeremy Atherton, University of Chicago Library Map Collection, and Christopher Siciliano. Community areas of Chicago, Illinois. 2011. http://en.wikipedia.org/wiki/File:US-IL-Chicago-CA.svg. 5

[4] Adrian Baddeley and Rolf Turner. Spatstat: an r package for analyzing spatial point patterns. *Journal of statistical software*, 12(6):1–42, 2005. 2.1

[5] Joel M Caplan, Leslie W Kennedy, and Joel Miller. Risk terrain modeling: brokering criminological theory and gis methods for crime forecasting. *Justice Quarterly*, 28(2): 360–381, 2011. 1

[6] Jacqueline Cohen, Wilpen L. Gorr, and Andreas M. Olligschlaeger. Leading indicators and spatial interactions: A crime-forecasting model for proactive police deployment. *Geographical Analysis*, 39(1):105–127, 2007. ISSN 1538-4632. doi: 10.1111/j.1538-4632.2006.00697.x. URL `http://dx.doi.org/10.1111/j.1538-4632.2006.00697.x`. 1

[7] Noel Cressie. The origins of kriging. *Mathematical Geology*, 22(3):239–252, 1990. 2.5

[8] Noel Cressie and Hsin-Cheng Huang. Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448):1330–1339, 1999. 1

[9] Peter Diggle. A kernel method for smoothing point process data. *Applied Statistics*, pages 138–147, 1985. 2.1

[10] David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B Tenenbaum, and Zoubin Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint arXiv:1302.4922*, 2013. 2.4

[11] John Eck, Spencer Chainey, James Cameron, and R Wilson. Mapping crime: Understanding hotspots. 2005. 1

[12] Anthony C Gatrell, Trevor C Bailey, Peter J Diggle, and Barry S Rowlingson. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British geographers*, pages 256–274, 1996. 2.1

[13] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013. 3.1

[14] Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006. 3

[15] Tilmann Gneiting, M Genton, and Peter Guttorp. Geostatistical space-time models, stationarity, separability and full symmetry. *Statistical Methods for Spatio-Temporal Systems*, pages 151–175, 2007. 1

[16] Wilpen Gorr, Andreas Olligschlaeger, and Yvonne Thompson. Short-term forecasting of crime. *International Journal of Forecasting*, 19(4):579–594, 2003. 1, 2.2, 3.2

[17] Clive Loader. *Local regression and likelihood*. New York: Springer-Verlag, 1999. ISBN 0-387-9877. 2.1

[18] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493), 2011. 1, 5

[19] J. Møller, A.R. Syversveen, and R.P. Waagepetersen. Log gaussian cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998. 5

[20] Nicholas G Polson, James G Scott, et al. On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012. 3

[21] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6: 1939–1959, 2005. 4

[22] R.J. Sampson. *Great American city: Chicago and the enduring neighborhood effect*. University of Chicago Press, 2012. 2.1

[23] Matthew A Taddy. Autoregressive mixture models for dynamic spatial poisson processes: Application to tracking intensity of violent crime. *Journal of the American Statistical Association*, 105(492):1403–1417, 2010. 1, 5

[24] Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen, and Aki Vehtari. Gpstuff: Bayesian modeling with gaussian processes. *The Journal of Machine Learning Research*, 14(1):1175–1179, 2013. 3, 4, 5

[25] Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for machine learning, 2006. 2.3, 2.3, 2.4, 3