Cities Are Already Smart:

urban spatiotemporal prediction with Gaussian processes

SIMON RIMMELE

ADVISOR: BENJAMIN GOODRICH

MAY, 2018

Columbia University | New York

Many thanks for your time and advice:

Benjamin Goodrich,
Jonathan Auerbach,
Seth Flaxman,
Jonah Gabry (also for the LateX template)

Drawing its modeling and evaluation framework from Flaxman (2014), this thesis investigates the use of Gaussian process models for spatiotemporal prediction of social phenomena in an urban setting. Inspired also by advances in timeseries forecasting by Facebook's Prophet tool ("Prophet", n.d.), this work privileges simplicity and parsimony in the user experience, accepting only location and timeseries data as inputs while producing results with easily interpretable parameters for audiences with varying technical knowledge. Models are evaluated at varying levels of spatial and temporal resolution on predictive accuracy and estimate reliability in comparison to standard autoregressive models in a case study of motor vehicle collisions in New York City. It is argued that this type of modeling could be further developed so decision-makers in urban environments may easily leverage existing spatiotemporal data in their daily work.

Chapter 1

Introduction

Urban areas produce vast amounts of data each day, much of it bound to specific location and time. Every bus signaling its location to an app, every complaint made about noisy neighbors, every police report about a traffic collision is a measurement. While it is easier than ever to put data on a map, making sense of it as an administrative or policy professional can be an exercise in frustration. Answers to simple questions about why things are happening and where they may happen next typically require statistical knowledge and access to even more explanatory data.

There is potential for explanatory or predictive modelling using solid statistical foundations and only the data at hand. Some progress in this direction has been made when considering timeseries; Facebook's Prophet is a successful forecasting tool with a Bayesian underpinning that nevertheless is accessible to analysts and researchers who bring their own skillset and expert knowledge (?, ?). To use Prophet all one has to do is bring a timeseries and potentially some prior guesses about trends or important events. The additive model underpinning Prophet is capable of neatly decomposing the timeseries into multiple interpretable trends and also offers forecasting under uncertainty.

Using space as a predictive dimension in the same way Prophet uses time has the potential to be immensely helpful for urban research and administration. There is obvious theoretical and practical appeal to using spatiotemporal methods in the context of the ’Smart City’, whose premise is found in rethinking urban areas as an intricate system monitored and managed at a previously impossible level of detail through the use of data (Kitchin, 2014). While Smart City proponents have pointed to the potential of technological data generating mechanisms such as sensor networks, there is already a vast amount of generated data available through administrative channels such as 311, public safety reporting, or existing networked sensor systems like traffic monitoring. While there have been limited surveys into the use of spatiotemporal models for urban applications, even these were restricted to forecasting at a low level of spatial granularity such as load demand for energy grids (Tascikaraoglu, 2018). Little work has been done to assess the viability of making use of hyperlocal data for forecasting outcomes in an urban environment, much less productionizing a model to better deliver social services exactly where they are needed and at the right time.

Chapter 2

Literature & Model Review

2.1 Spatiotemporal Modelling

Count-based spatial and spatiotemporal models are widely used in ecological fields e.g., for wildlife population and habitat studies (CARROLL & JOHNSON, 2008), as well as in the public health field for disease mapping (Schrdle & Held, 2011). Intrinsic Conditional AutoRegressive models (ICAR) are a spatial-only model with widespread application in public health context and produce a measure of relative risk similar to the one used for this project (Wakefield, 2006). Spatial and spatiotemporal models used in these fields are typically at a low to medium level of spatial granularity, such as the regional model of the northwestern United States used in (CARROLL & JOHNSON, 2008). For studies set in urban areas some research has been done at the level of census tracts in an individual county, like Chun et., al's study on disparities in environmental hazards in Maricopa County (CHUN, KIM, & CAMPBELL, n.d.).

Outside of the natural and environmental sciences, some forms of spatial and spatiotemporal modelling have found use in predictive policing research and industry use, given that ”space-time interactions are deeply embedded both empirically and theoretically into many areas of criminology” (Li, Haining, Richardson, & Best, 2014). The scope of police work also dictates most or all predictive modelling is done in urban areas and at a much finer scale of spatial granularity. Recently developed likelihood-based estimators for spatiotemporal counts have been fit on crime data from 138 census tracts in Pittsburgh, Pennyslyvania (Liesenfeld, Richard, & Vogler, 2017); Flaxman et. al., developed a forecasting algorithm for crime in Portland, Oregon at the granularity of 66,000 cells measuring a quarter of a square mile each (Flaxman, Chirico, Pereira, & Loeffler, 2018).

Other than in law enforcement, spatiotemporal models have also found some applications in urban social science through traffic modelling. Cheng et. al., used a non-Bayesian approach to model and forecast traffic dynamics in London (Cheng, Haworth, & Wang, 2012). In this study, the authors pose and attempt to grapple with one of the fundamental and ongoing challenges of using data in the context of a Smart City; the volumes of data available are often so massive thay they either overwhelm existing modelling methods and/or offer an almost endless number of ways to further segment and structure the data for use. Cheng et. al, use a variety of non-Bayesian methods including likelihood-based statistical models and gradient-based neural networks, but find these methods incapable of capturing the full autocorrelation structure of traffic networks. The tradeoff between the spatiotemporal complexity of a model and the computational feasibility of fitting the model in medium to large datasets is evident here.

2.2 Gaussian Processes

This project will assess the viability of using Gaussian Process models for spatiotemporal predictions in urban settings. Gaussian Processes have attractive properties for sparse data because a model can be interpreted as interpolating an unobserved point between known distributions. As mentioned in the literature review, spatiotemporal models have historically suffered from computational infeasibility when the amount of data as well as spatial and temporal dimensionality grows. Recent advances in Bayesian probablistic programming and implementations of Gaussian Processes that take advantage of the relative sparsity of data across spatial dimensions has reduced the computational burden of fitting these models.

A Gaussian Process (GP) is a generalization of a Gaussian - also known as the Normal - distribution. A Normal distribution is defined by a scalar mean μ and variance σ in the univariate case, and a n-length vector μ with a n-by-n dimensional covariance matrix Σ in the multivariate case. A Gaussian Process ”can be viewed as a potentially infinite-dimensional generalization of [the] Gaussian distribution” (Gelman et al., 2013) , p. 501. However any finite-dimensional marginal distribution from a Gaussian Process is also Gaussian, which makes a GP suited as a prior distribution for some unknown regression function $\mu(x)$ - where x is a vector with arbitary but finite dimensions. A generic Gaussian Process $\mu \sim GP(m,k)$ is a series of random functions drawn from an n-dimensional normal distribution (Gelman et al., 2013):

$$\mu(x_1),...,\mu(x_n)) \sim N((m(x_1),...,m(x_n)),K(x_1,...,x_n))$$

Thus a Gaussian Process is defined entirely by a mean function m and covariance function/s K. Sums and products of GPs are also GPs, which makes it simple to combine different variations of covariance functions.

2.2.1 Estimating Gaussian Processes

While it is not necessary to go into extensive detail about the theory of Gaussian Processes for this application, it is relevant to briefly describe what part of the estimation poses a challenge for high-dimensional datasets like those found in spatiotemporal forecasting. To generate samples from a finite-dimensional realization of a GP $\mu \sim N(\mathbf{m}, K)$, it is necessary to calculate the Cholesky decomposition of the covariance matrix: $K = LL^{\intercal}$. Then it easy to generate standard Normal draws $\mathbf{u} \sim N((0), I)$, and shift them using $\mu = \mathbf{m} + L\mathbf{u}$ (Rasmussen & Williams, 2005). Calculating L is quite computationally demanding as n rises. In practice the eigenvalues of K can also decay, causing the decomposition to fail. This is solved in practice by adding a small 'jitter' term to the diagonal of K. Ideally the jitter term is small enough to not influence the estimates, but this has to be assessed through trial and error.

2.3 The Log-Gaussian Cox Process

Log-Gaussian Cox Process (LGCP) models are a further extension of a Gaussian Process with particular applicability to prediction of count data. A LGCP is hierarchical in that the data are assumed to be drawn from a Poisson likelihood with intensity parameter λ. The Poisson likelihood function has theoretical properties suitable for relatively sparse count (integer) data which makes it appealing for use in modelling frequent events that are nevertheless sparse when segmented over space and time dimensions. In turn the log of λ is generated by a Gaussian process(Teng, Nathoo, & Johnson, 2017). This makes the LGCP quite flexible in inputs and dimensionality while also making model fitting generally computationally burdensome.

An alternative specification uses a Gaussian likelihood, which has the added appeal of making the likelihood and prior conjugate to each other and solvable in closed form. As the λ parameter of a Poisson distribution increases the distribution converges to a Normal distribution anyway, so this model may be more applicable to situations when counts are not sparse e.g., at lower granularities of space and/or time, which was not considered in detail here.

A zero-inflated Binomial (ZIB) likelihood would be another alternative to consider for this type of modelling. As the name implies, a ZIB distribution has a higher probability density around zero and may be better suited for extremely sparse data like gridded observations with very small grid dimensions.

The general model follows the specification and notation used by ”A General Approach to Prediction and Forecasting Crime Rates with Gaussian Processes” (Flaxman, 2014):

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

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

The outcome count $y_{s,t}$ at location s and time t is generated from a Poisson distribution whose scale parameter is a the function $f(s,t)e_t$. $f(s,t)$ is a function with a Gaussian process prior, while e_s is a fixed spatial expectation term $\mathbb{E}[y_s]$. The spatial expectation is a convenient way to incorporate prior information about the variable of interest. Again following Flaxman 2014, using an exponential link function gives a practical interpretation of $exp(f(s,t))$ as the relative risk function while $f(s,t)$ itself is the log-relative risk. When the log-relative risk is 0, the relative risk is 1 and the expected value of of the outcome $y(s,t)$ is just the prior spatial expectation e_s. Finally, every count outcome $y_{s,t}$ is assumed independent conditional on f, so the joint conditional likelihood of all y factors as a product.

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

2.3.1 f(s,t)

f is modeled as following a generic Gaussian process with a mean of zero and a covariance matrix K:

$$f \sim GP(0,K)$$

Spatiotemporal elements enter the Gaussian process model through K, which is used to capture the relationships of interest and entirely determines the model. Since variance/covariance is additive, any variance term can be decomposed at least theoretically into arbitrarily many additive components. Purely spatial and temporal elements as well as interactions can be incorporated using different kernel functions, and different combinations of kernels may produce more accurate results. Cross-validating models with different types and combinations of kernel functions will help identify the specification with the best predictive properties. Since the cross-validation data are correlated in this model the cross-validation set will have to be drawn from spatially contigous representative subsets of the data.

2.3.2 Kernels

There are a wide variety of existing kernel functions documented for use in Gaussian process models (Rasmussen & Williams, 2005) and it is also possible to define custom functions as long as they follow certain properties. The entirety of a GP is determined by its kernel/s, which makes them versatile and open to many different specifications.

An important consideration is whether the model is considered to be stationary. In spatiotemporal models the case for assuming stationarity weakens as the time-period being considered lengthens, and this should in turn affect the kernel choice. The long term model considered includes a non-stationary linear kernel component intended to capture any long-term trends:

- $k_t(t)$: a temporal kernel
- $k_p(t)$: a periodic kernel
- k_l: a linear kernel

The short-term models considered for this project are all stationary and loosely follow Flaxman 2014 by consisting of up to four individual kernels k , where $K = \sum_i k_i$. The four basic kernel types were:

- $k_s(s)$: a spatial kernel

- $k_t(t)$: a temporal kernel
- $k_p(t)$: a periodic kernel
- $k_{st}(s,t)$: a space-time interaction kernel

Radial Basis Function (RBF) and Matern family kernels are both candidates for spatial and temporal modelling. Their point of differntiaion lies in the extent to which they 'smooth' the input. The most appropriate kernel will vary based on the characteristics of the observations being modeled.

The RBF kernel - also called the squared exponential - is widely used as a default stationary kernel:

$$k_t(t) = exp\big(-\frac{(|t-t'|)^2}{2\ell^2}\big)$$

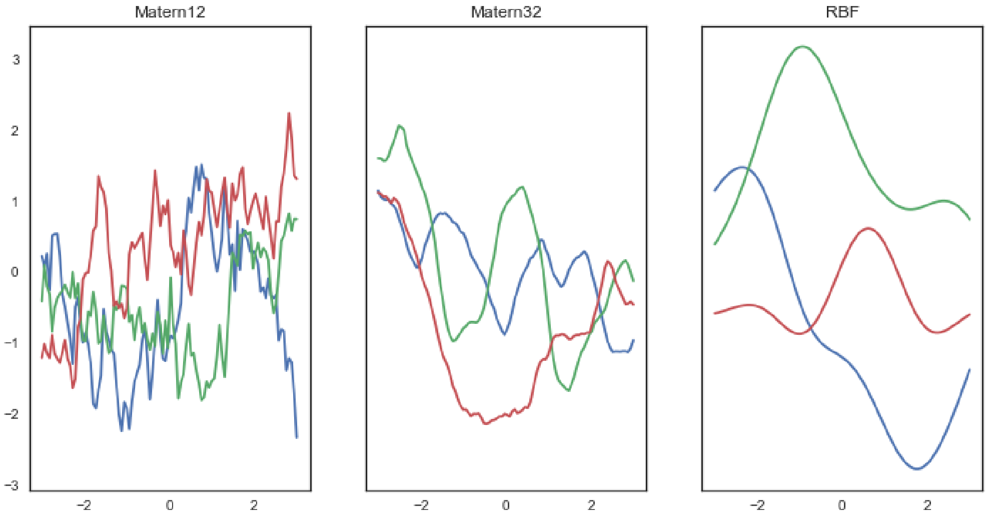
The L1 norm $|t-t'|$ of the input determines the kernel, along with ℓ, a lengthscale parameter that is common to kernel functions.

The Matern family kernels are defined:

$$k_t(t) = \frac{1}{2^{\nu-1}\Gamma(\nu)}\big(\frac{\sqrt{2\nu}}{\ell}(|t-t'|)\big)^{\nu}K_{\nu}\big(\frac{\sqrt{2\nu}}{\ell}(|t-t'|)\big)$$

The parameter ν determines the degree of smoothness and the kernel converges to a Gaussian kernel as ν goes to inf. Typical values of ν are fractions,with $\nu = \frac{1}{2}$ or $\frac{3}{2}$ being quite common. The name of a specific kernel is often expressed by referencing the parameter value, so in this paper a 'Matern32' refers to a Matern kernel with $\nu = \frac{3}{2}$. Figure 2.1 shows a comparison of the degree of smoothing a RBF,Matern12, and Matern32 kernel have on the same multivariate random input:

Figure 2.1 *A comparison of RBF with Matern family kernels.*



The periodic is another class of stationary kernel. The variant used in this project follows MacKay 1998 (MacKay, 1998):

$$k_p(t) = exp\left[-\frac{1}{2}\sigma_i \left(\frac{sin(\frac{\pi}{\Delta}(t_i - t_i'))}{\ell}\right)^2 \right]$$

Δ is a periodicity paramter that can either be fit or set to a fixed anticipated value such as 52 for annual weekly data. ℓ is again the lengthscale parameter.

2.3.3 Priors

Gelman 2006 suggests using a student-t prior distribution restricted to be positive (a 'half-t') as a prior for variance parameters in hierarchical models (Gelman, 2006). Since variance is always positive it is acceptable and desirable to restrict the prior distribution to be non-negative as well. Flaxman 2014 used a standard student-t prior with 4 degrees of freedom on all parameters (not just variance). Both standard Normal $N \sim (0,1)$ and Student-T priors were considered here. The normal prior has less probability mass in its tails than the Student-t but is otherwise a close substitute. The Normal prior did sometimes lead to the optimization algorithm failing to find a credible set of parameter estimates, while the Student-T with degrees of freedom between 2 and 10 performed better. Increasing the variance on the prior distributions was another way to lower the probability of the optimizer failing, but it also led to dramatically increased runtimes in some cases.

2.4 Methods for fitting LGCPs

Solutions to the computational challenges in fitting LGCPs have advanced rapidly over the past five years. As mentioned earlier, the computational bottleneck for Gaussian Process models is the the n-by-n covariance matrix, which requires an $O(n^3)$ computation. Fitting a GP model as n grows beyond a few thousand points is quite challenging (Gelman et al., 2013).

As recently as 2012 the best methods for fitting LGCP models involved variations of the Metropolis-Hastings sampling algorithm which were considered to be slow and highly inefficient in generating acceptable draws from the posterior distribution of the model (Murray, Ghahramani, & MacKay, 2012). Advances in statistical computing have opened the door for several new model fitting methods, each with their own advantages and drawbacks.

2.4.1 Variational Inference

Variational Inference (VI) methods attempt to recover parameter estimates for a posterior distribution of interest by specifying a more tractable family of distributions and optimizing the resulting approximation using closed-form or computational methods. Usually the family of approximating distributions is denoted Q, and the optimization seeks to minimize the Kullback-Leiber divergence $KL[(\theta), p(\theta|y)]$ between $q(\theta)$ and the true posterior $p(\theta|y)$.

In the case of the LGCP with Poisson likelihood there is no direct closed-form solution because an intractable integral is involved. VI methods have in the past been limited by needing to make simplifying assumptions about the approximating distributions. Usually VI makes the 'mean-field' assumption (borrowed from physics), that the posterior can be approximated by the product of some number of existing distributions: $q(\theta_1, ..., \theta_n) = \prod_j q_j(\theta_j)$. The mean-field assumption is especially unsuited for high-dimensional models like a Gaussian Process, but Tran and Blei characterize a Variational Gaussian Process method that avoids the assumption (Tran, Ranganath, & Blei, 2015). GPflow - the package used for this project - uses the Variational Gaussian for fitting (Matthews et al., 2017).

The major drawback of variational methods is there are no theoretical justifications for the accuracy of estimates produced because it is unclear and often unknowable how close the optimized approximation is to the true posterior distribution, where the KL divergence term is a unitless and uninterpretable difference between the two that cannot be compared across models. Yao et., al. recently proposed Pareto-Improved-Importance-Sampling (PSIS) and Variational-Simulation-Based-Calibration (VSBC) methods for assessing whether a VI approximation "worked"

(Yao, Vehtari, Simpson, & Gelman, 2018).

2.4.2 MCMC

Markov Chain Monte Carlo (MCMC) sampling methods by contrast do have theoretical properties that ensure consistent estimation of the posterior distribution - at least as the numbers of samples increases asympotically (Teng et al., 2017). MCMC methods are not a new development in themselves, but there have been breakthroughs both in MCMC algorithm design and computing power required to fit more complex models. The probablistic programming language Stan has been used for spatiotemporal modelling of causes of mortality using Gaussian Markov Random Field models, another potential alternative to LGCPs that may be appropriate for urban forecasting but are not considered here (Stan Development Team, 2018) (Foreman, Li, Best, & Ezzati, 2017). In contrast to VI, MCMC also is capable of producing estimates for full posterior distributions for parameters of interest, rather than point estimate approximations. In the case of urban prediction and forecasting, access to full posterior estimates would offer much more probalistic information from which to draw uncertainty-based conclusions in a policy or administrative context.

2.4.3 LaPlace Approximation

Integrated Nested LaPlace Approximation (INLA) is an alternative to MCMC capable of fitting a LGCP (Illian, Sørbye, & Rue, 2013). INLA works by approximating the distribution of each parameter around the mode of its posterior (Lindgren & Rue, 2015). The marginal posteriors are calculated by numerically integrating over the parameters, followed by another approximation of the marginal posterior (hence the ’Nested’ component of INLA). When the number of hyperparameters is relatively small, INLA is capable of quickly fitting a latent Gaussian field model - of which the LGCP is a special case (Rue, Martino, & Chopin, 2009). LGCP models usually assume a square grid structure for the spatial component in order to fit the model, but Simpson 2016 also fits LGCPs without relying on explicit grid structure (Simpson, Illian, Lindgren, Sørbye, & Rue, 2016).

2.5 Method Choice

MCMC methods clearly offer desirable properties superior to VI methods, and full MCMC models similar to the ones considered here have been explored (Flaxman, Gelman, Neill, Smola, & Vehtari, 2015). However, their practical limitations made them somewhat burdensome to consider for this project. The most unfortunate drawbacks were that MCMC is very slow to fit in comparison with VI. In an attempt to draw a compromise between ease of experimentation and model reliability this project used VI methods under the knowledge that they may not produce the best results for practical use. Simple posterior checks were done to assess the results.

Chapter 3

Experiments

Both long- and short-term prediction and forecasting experiments were considered, as well as both neighborhood level spatial granularity and a larger city level scheme. A summary of each is provided in the table below:

Table 3.1 *Models Considered*

Outlook	Spatial Feature	Frequency	Fitting Period	Look-Ahead Period	Stationary
Long-Term	None	Weekly	3 years	52 Weeks	No
Medium-Term	Neighborhoods	Weekly	1 year	26 Weeks	Yes
Short-Term	Grid Squares	Weekly	6 weeks	12 Weeks	Yes

3.1 Data

The experiments used observational administrative data gathered as part of New York City's Open Data civic reporting system (*"New York City Open Data: Motor Vehicle Collisions"*, n.d.). There are many recorded urban events outside of the aforemented policing context where predictive and forecasting knowledge could be helpful to resarch, policy, or operations. Quality-of-life programs such as noise abatement, pest control, and traffic reduction are all natural candidates where data is already collected from the public; in New York City it is done through the 311 civil reporting system. Each also involves allocating scarce public resources e.g., public safety officers, exterminators, and health inspectors over space and time.

Although the aforementioned urban events are ex-ante all topic of interests with some potential for spatiotemporal forecastability and would benefit from more accurate modelling, this project instead focused on another heavily spatially dependent public safety issue: vehicle collisions. There have been over 200 fatalities annually resulting from vehicle collisions in New York City over the past five years while Mayor Bill de Blasio's Vision Zero plan has publicly stated a goal of lowering the number of fatalities to zero (of Transportation, n.d.).

While approximately 200 fatalities a year is not a 'rare' event, the likelihood of observing a fatality at any given time and place in New York City is exceedingly rare. What's more, public safety interventions in the form of increased enforcement or street redesign can be deployed reactively to the scene of a fatality with little or no information about the relative risk of that location.

By contrast, non-fatal injuries resulting from traffic collisions are much more common, with about 59,000 in 2017 (of Transportation, 2018). By thinking of injuries as a potential fatality observing and attempting to forecast these injures reliably could provide a much better estimate of where the city is particularly at risk for traffic collisions. These estimates could in turn be used to evaluate the efficacy of any public safety intervention after the fact instead of relying on observing only the future fatality count, which may be subject to mean reversion in a situation where the fatalities at any given place and time are very close to zero (Auerbach, 2017). Since vehicle collisions are reported directly by responding public safety personnel they are not subject to the reporting bias it would be plausible to find in complaint/report data like 311, which would have added an additional source of potential bias.

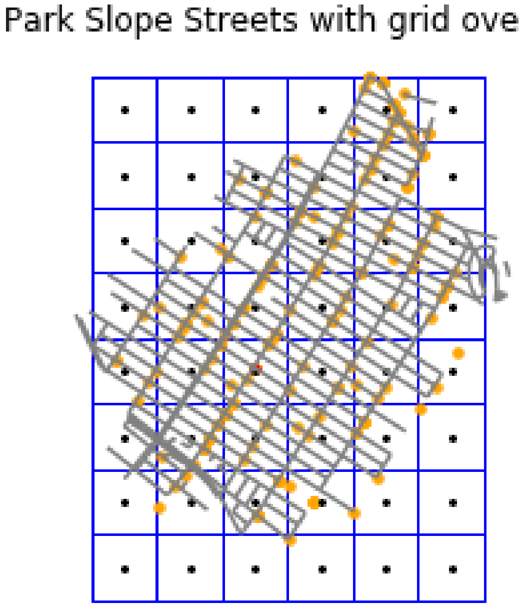
3.1.1 Data Processing

Organizing spatiotemporal count data into a suitable form for modelling presents a challenge in itself. While every count observation occured at a specific point and time, aggregating to a form appropriate both for modelling and capable of providing useful granular forecasts is a balancing act. Traffic collisions in particular are difficult to aggregate into consistent spatial locations that also have a coherent interpretation of the distance between each location necessary for fitting a spatiotemporal model.

There are many administrative spatial aggregation bounds which are of potential use, such as Neighborhoods, Zip Codes, Census Blocks, or even unique street intersection IDs. While all may be perfectly viable depending on the situation, gridding provides a more generalizable spatial aggregation method which is already used in many spatial statistics applications, see for example the case studies in (Blangiardo & Cameletti, 2015). Gridding simply overlays a square grid of chosen size on the area of interest and sums all the count data within each grid cell into a single data point, a flexible method that can be easily, if imperfectly, applied to many spatial contexts. Being able to flexibly choose the size of grid squares adds an additional layer of modelling flexibility.

Finally gridding offers clear benefits in the specific case of Gaussian process models. Since the spatial kernel k_s calculates the distance between any point s_i and all other points in s the computational cost of adding an additional point in s is high. By specifying the distance calculation only between the centroid of each grid square this method at least offers an explicit choice between higher spatial resolution and computational complexity.

Figure 3.1 *A grid of 1250ft grid squares overlaid on the Park Slope street grid.*



3.2 Kernel Choice

The kernel types specified for each k differed slightly from Flaxman 2014 after experimenting with various configurations. Table 3.2 describes the kernel combinations used for each model. Other combinations were considered, including using variations of the Matern class and

Table 3.2 *Kernel Specifications*

	Temporal	Spatial	Periodic	Product	Linear
Long-Term	RBF	No	Yes	No	Yes
Medium-Term	RBF	Matern32	Yes	Spatial x Temporal	No
Short-Term	RBF	RBF	Yes	Spatial x Temporal	No

varying the interaction term, but none exhibited better performance.

3.3 GPflow

GPflow is a Python package developed for fitting Gaussian Process models that takes advantage of the gradient methods of Google's TensorFlow to generate Maximum-A-Posteriori (MAP) estimates of the posterior distribution (Matthews et al., 2017) (Abadi et al., 2015). All models are converted into tensors and passed to Tensorflow's optimization algorithms that can run on Graphics Processing Units (GPUs) for fast and efficient computation. This cuts the time needed for each iteration significantly compared to running on a CPU and also doesn't restrict the user to using conjugate distributions when specifying priors and likelihoods - a prerequisite for the LGCP model. GPflow is also capable of full Bayesian inference using Hamiltonian Monte Carlo (HMC), but is currently only available experimentally.

3.3.1 Custom Modifications

While GPflow offers most of the settings required 'out-of-the-box', a few additional features had to be added independently. The base package does not offer the option for a Student-T prior, which was relatively easy to add by writing a new custom Student-T class to GPflow. The package also by default does not easily allow for decomposition of the different kernels' contribution to the parameter estimates. Finally, the GPflow implementation of Matern kernels often exhibit unstable behavior, even after re-centering data. A small jitter was added to the Matern32 kernel in in order to reduce the chance of the Cholesky decomposition failing.

Chapter 4

Results

Borrowing again from Flaxman 2014, the models were evaluated on a variant of R^2 that captures the reduction in total variance from each kernel component, which in turn can be interpreted as a measure of what components are most important in explaining the model. The error from predicting solely on the prior spatial expectation e_s serves as a baseline for comparison:

$$Reduction-in-Variance = 1 - \frac{\sigma_{s,t}(n_{s,t} - \hat{n_{s,t}})^2}{\sigma_{s,t}(n_{s,t} - e_s)^2}$$

Mean Squared Error (MSE) was also calculated for each of the predictions from each $f_i(s,t)$ and compared to an AR(1) model that only relies on temporal predictions from the prior period.

A Pareto-Smoothed-Importance-Sampling Leave-One-Out score (PSIS-LOO) was used to check the fit of the variational approximation of the posterior density (Vehtari, Gelman, & Gabry, 2015). The PSIS-LOO score, also referred to as k, is a measure for estimating the pointwise out-of-sample prediction accuracy of a model. Generally k scores of less than 0.5 are desirable, with scores between 0.5 and 1 necessitating caution.

4.1 Long Term Predictions

Weekly traffic collision and injury data was aggregated to one weekly series for all of New York City. While data is available for the past five years, there appears to have been a reporting error for part of 2016 which made the year of data problematic for both model fitting and testing 5.6. It would have been ideal to fit the model based on multiple years of data leading up to the latest available data, but given the unreliable data the best alternative was fitting the LGCP on three years of data from July 2012 to July 2015, and leaving the following 52 weeks for out-of-sample prediction. Crashes without latitude and longitude were excluded, as well as any crashes that happened on highways and other high-speed motorways. The prior spatial expectation e_s was taken from New York State Department of Health statistics for injuries stemming from motor vehicle collisions during the in-sample period (of Health, n.d.).

Figure 4.1 shows the model fit for the citywide data:

4.1.1 Kernel Components

Since the kernel components are all additive they can provide an interpretable decomposition of the model components. There is clear annual periodicity in the model, which is also confirmed by the period parameter fitted being almost exactly 52. The linear kernel component also suggests a small downward long-term trend in pedestrian injuries.

Figure 4.1 *A Gaussian Process fit to weekly pedestrian injuries in New York City from 2012 to 2015. The dotted line shows where out of sample prediction begins. The 95% credible interval - in light grey - contains less than 95% of the observed data, which suggests the parameters may not be optimal.*

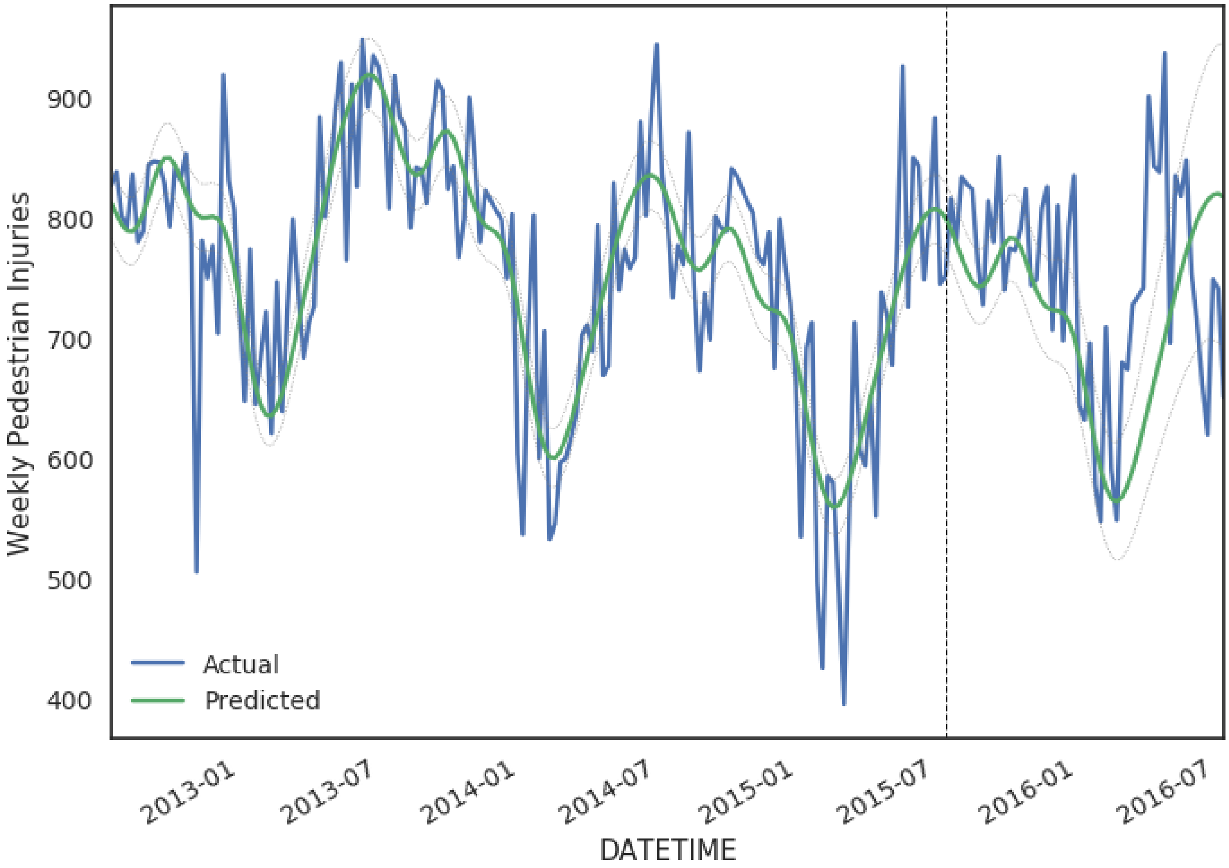


Table 4.1 *Models Considered*

kernel	prior	value
PartialVGP/kern/periodic/period	None	0.52
PartialVGP/kern/periodic/variance	student-T([0.],[1.][4.])	0.03
PartialVGP/kern/periodic/lengthscales	student-T([0.],[1.][4.])	0.40
PartialVGP/kern/rbf/variance	student-T([0.],[1.][4.])	0.38
PartialVGP/kern/rbf/lengthscales	student-T([0.],[1.][4.])	0.89
PartialVGP/kern/linear/variance	student-T([0.],[1.][4.])	0.54

The Reduction-in-Variance for each kernel component shows the temporal RBF kernel captures almost all the explanatory benefit of the model by itself. Note that each kernel component was evaluated independently so the RIV of the individual components will not add up to the Combined total.

Table 4.2 *Models Considered*

Kernel	Reduction-in-Variance
Linear	43.2%
Time (RBF)	85.2%
Periodic	13.5%
Combined	95.1%

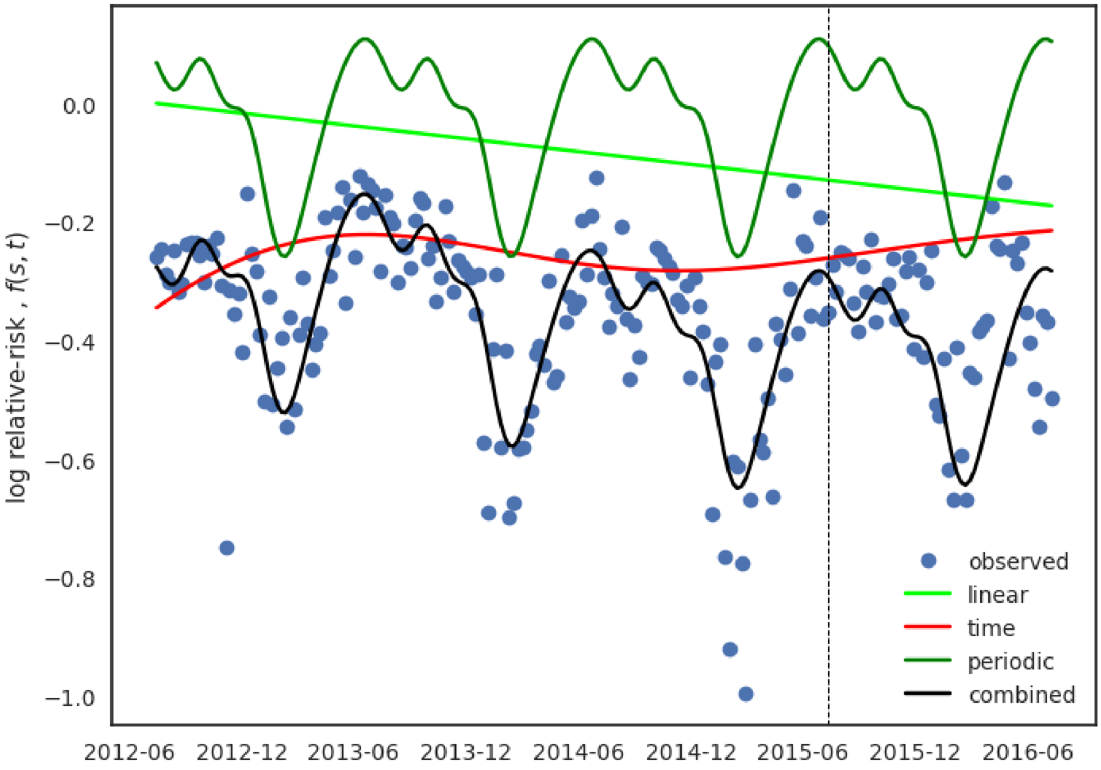
It is possible to plot the additive components of the $log(f)$ latent risk function in order to explain the final Combined total. As seen in 4.2, the long-term model suggests a modest downward trend in risk from traffic collisions as reflected by the linear kernel, independent of any seasonal or time-trend components.

The PSIS-LOO k score for the citywide model was 0.8, which is problematically high but not necessarily disqualifying according to (Vehtari et al., 2015).

4.2 Neighborhood Predictions

A neighborhood level model adds a spatial component to the existing temporal model. New York City defines official Neighborhood Tabulation Areas (NTAs), which were used as the geographic areas of interest. The model was fit on the 29 official NTAs in the borough of Manhattan using 52 weeks of weekly pedestrian injury data starting in January 2013. The out-of-sample period was the first 26 weeks of 2014. Again the reporting gaps in collision data limited

Figure 4.2 *Additive components of the latent risk function.*



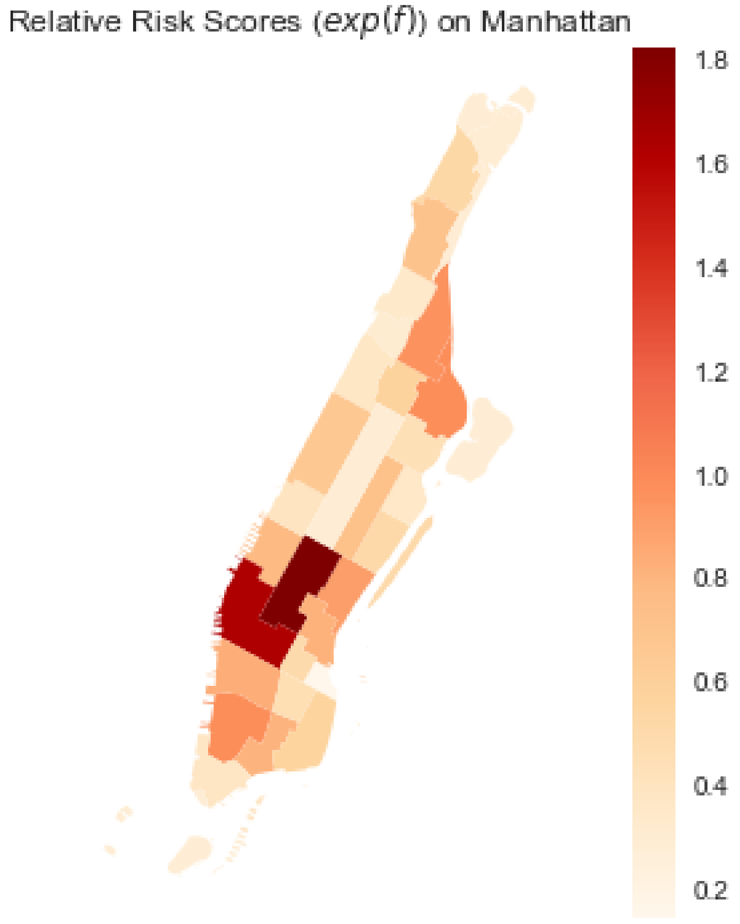
the periods where reliable data was available continuously for all neighborhoods. e_s was calculated using the average number of weekly neighborhood pedestrian injuries prior to January 2013. The spatial distance was calculated from the distance between centroids for each NTA (in feet) on the NAD83 New York State Plane projection.

Table 4.3 *NTA data example.*

	COUNT	NTAName	x_point	y_point	DATETIME
4	0.0	Battery Park City-Lower Manhattan	0.0814095	0.0972240	2012-07-01
20	0.0	Central Harlem North-Polo Grounds	0.1006435	0.1373956	2012-07-01
21	1.0	Central Harlem South	0.0977342	0.1323207	2012-07-01
23	5.0	Chinatown	0.0857386	0.0999914	2012-07-01
24	0.0	Clinton	0.0863560	0.1176865	2012-07-01

The LGCP predictions performed as well as AR(1) models trained on each neighborhood

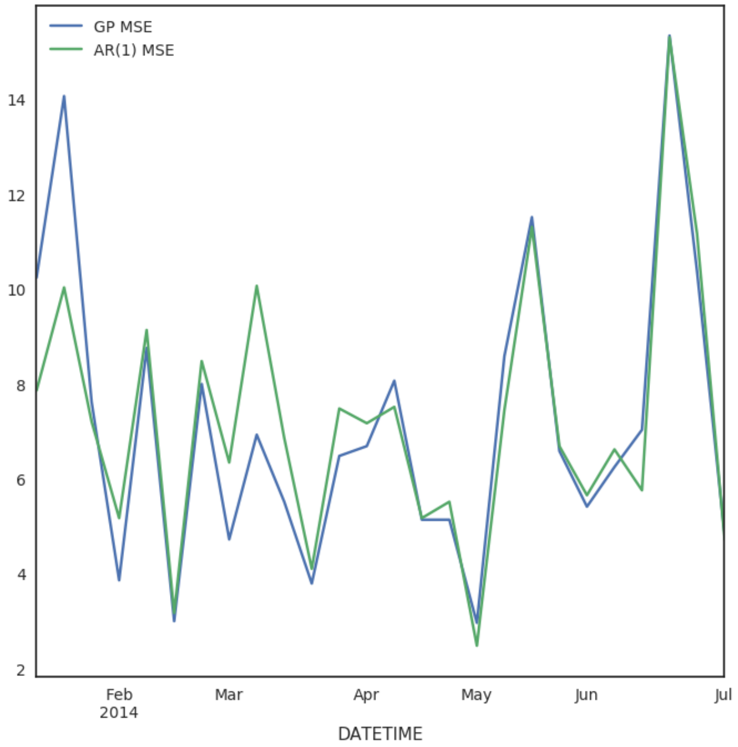
Figure 4.3 *Relative risk scores for Manhattan neighborhoods. Midtown is roughly twice as dangerous for pedestrians than Manhattan as a whole.*



seperately, Although the MSE for the spatiotemporal model during the out-of-sample period was slightly lower than the AR(1) model average, the difference is negligible.

The PSIS-LOO score for the neighborhood model was close to 1 , which is worryingly large and undermines the predictive potential of the model.

Figure 4.4 *Average Mean Squared Error for each neighborhood during the out-of-sample period for the LGCP model and a AR(1) comparison.*



4.2.1 Kernel Components

The neighborhood model assumes a stationary kernel in contrast to the citywide model. The two-dimensional spatial distance Matern32 kernel contributes noticeably to the Reduction-in-Variance score, while the RBF time kernel and the multiplicative interaction between space and time are also large contributors to the model's explanatory power. The Periodic component makes a smaller contribution than in the long-term model. The total Reduction-in-Variance of 65% is less than the long term model, but still a large improvement over defaulting to neighborhood averages.

Since the exponentiated latent function $exp(f(s,t))$ is directly interpretable as a measure

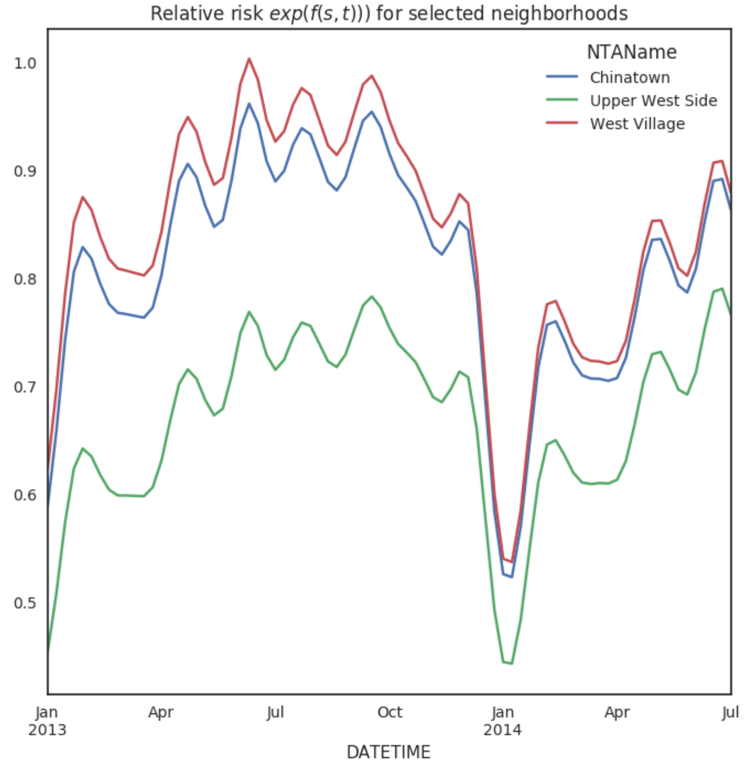
Table 4.4 *Neighborhood Reduction-in-Variance*

	Reduction-in-Variance
Time (RBF)	27.2%
Spatial (Matern32)	22.3%
Periodic	2.0%
Product	19.2%
Combined	64.6%

of relative risk it can be used for direct comparisons between neighborhoods

A similar assessment of relative risk can be done across periods, A comparison between three Manhattan neighborhoods in 4.5 shows varying levels of latent risk while all three are subject to seasonality and periodicity.

Figure 4.5

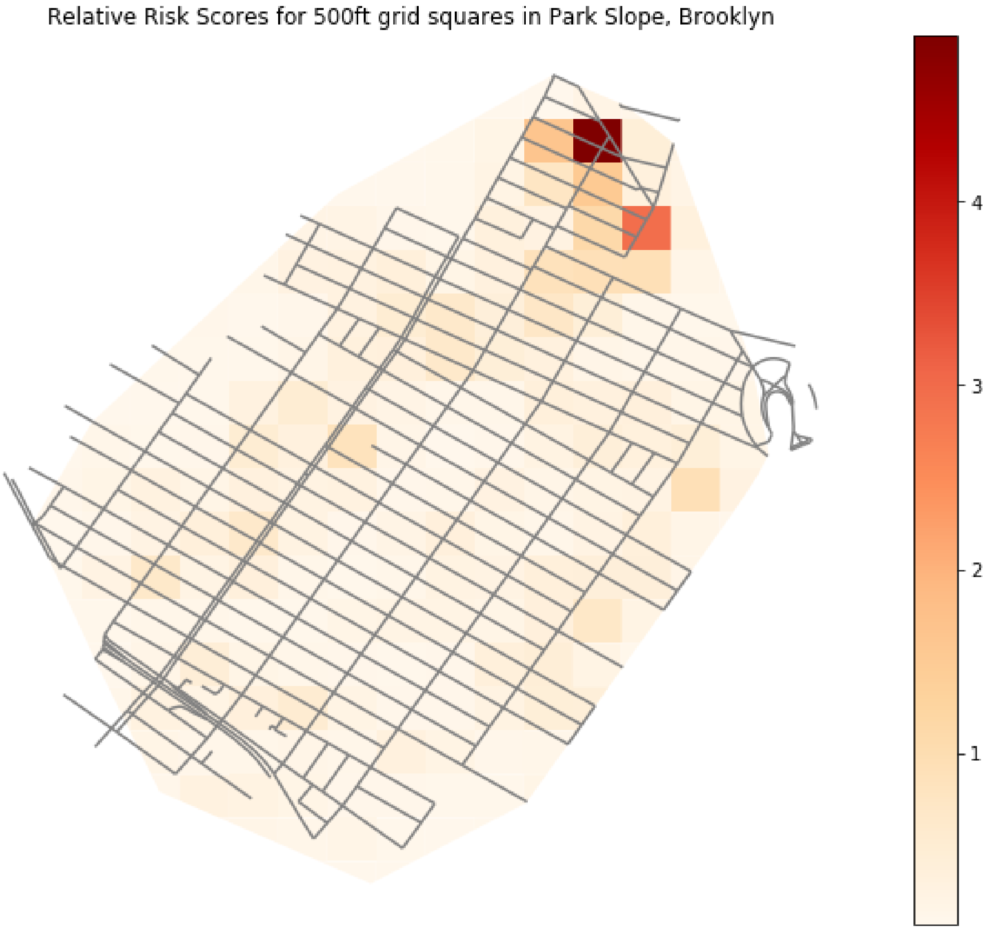


4.3 Short Term Forecasting

Short term forecasting using grid data was tested on a single neighborhood using weekly data. The neighborhood of Park Slope, Brooklyn was overlaid with a square grid and all data within the square was aggregated. The model included all vehicle collisions rather than only collisions causing pedestrian injuries because the data for pedestrian injuries is already quite sparse when looking at at the level of an individual neighborhood. The majority of grid squares will see no observed data for any given week. The grid model was tested with an out-of-sample period of 12 weeks while varying the historical training length between 3-12 preceding weeks. Distance was measured between the center of each grid square. The prior spatial expectation e_s was taken from the average weekly collisions in Park Slope during 2015 and scaled by the number of grid squares. Figure 4.6 shows a short-term model fitted to grid squares with an edge length of 500 feet.

The short-term model outperformed an AR(1) noticeably as measured by MSE, especially when using less time periods for fitting. This can be seen in 4.7. The difference in performance was especially relevant when forecasting futher into the out-of-sample period. The AR(1) performance converged to be comparable to the Gaussian Process model as the number of weeks used for fitting increased, while the PSIS-LOO score was 0.8.

Figure 4.6


Relative Risk Scores for 500ft grid squares in Park Slope, Brooklyn
4
3
2
1

4.3.1 Kernel Components

The kernel components were kept as close as possible to the neighborhood model, but the spatial Matern32 kernel proved to be too unstable, even after adding a jitter; another RBF kernel was used instead. The short-term model had the lowest share of explained variance between the three models, almost entirely attributable to the spatial kernel.

Figure 4.7 *Average Mean Squared Error across grid squares during the out-of-sample period for the LGCP model and a AR(1) comparison*

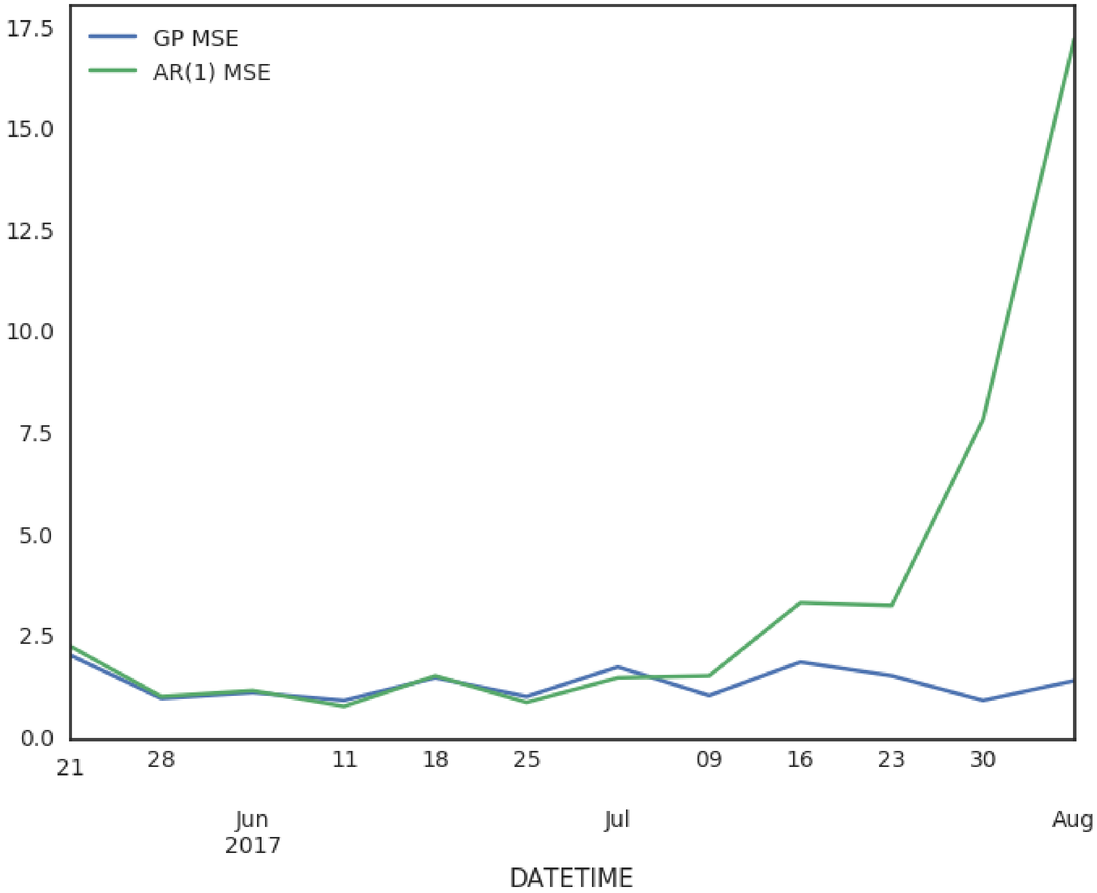


Table 4.5 *Grid Square Model Reduction-in-Variance*

	Reduction-in-Variance
Time (RBF)	-3.5%
Spatial (RBF)	35.6%
Periodic	0.0%
Product	26.1%
Combined	32.0%

Chapter 5

Discussion

The goal of this research was to evaluate the value of spatiotemporal Gaussian Process models for applications in urban policy evaluation and forecasting, and some of the results are quite promising. Gaussian processes are capable of convincingly modelling the incidence of traffic collisions in various resolutions of urban environment. The citywide temporal model offers a decomposition of the inevitably noisy observed data into components which are directly useful to policymakers and far better predictors than a simple average, all while relying only on previously observed count data as an input. The LGCP model of traffic collisions decomposes a longer term ’secular’ trend from short-term variability and yearly seasonality. The long term model however did have some statistical shortcomings, chiefly the ill-fitted credible intervals and poor k score. This should caution against relying on the model for overly precise predictions over the long term.

Modelling the spatial heterogeneity between neighborhoods as well as across time also has many potential uses. The interpretation of the latent function f_i at each location as a measure of relative risk is straightforward and easily applied for policy purposes. Again the decomposition of the kernel components allows for easy interpretation of the degree to which space and time contribute to the overall risk in each neighborhood, as well as being directly comparable to any other neighborhood.

The most ambitious modelling initiative was to attempt to forecast individual events at very high spatial detail, in some cases down to the space of one or two city blocks. This method offered better performance than normal autoregression when the lookback period was short while performing comparably to an AR(1) as the period was lengthened. The Reduction-in-Variance analysis suggest that the spatial distance component is valuable especially when timeseries data are otherwise sparse. It is also possible that a pure spatial model would perform better in this situation, based on how much of the Reduction-in-Variance is due to the spatial component.

5.1 Methodological Issues

5.1.1 Zero Inflation

The LGCP model as specified suffers when dealing with sparse observations, which happens naturally when events are recorded at increasing levels of spatial and temporal granularity. When using a very small grid edge sizes of 500-1000ft, The model would often converge to a trivial constant relative risk prediction of $f(s,t) = 1$, which is equivalent to a prediction of e_s at all points. Sometimes the approximation would fail to converge at all. It is possible the Poisson likelihood suffers from the 'zero-inflation' problem. With most grid and time points having no observations a majority of the time, a Poisson distribution does not have enough probability mass near zero to provide a good fit. An alternative likelihood like the Zero-Inflated Binomial may be better suited for very rare events.

5.1.2 Unstable Inference

The high PSIS-LOO scores of all the models is discouraging. This suggests potentially poor optimization results when fitting with Variational Inference. Using full posterior inference with Markov Chain Monte Carlo methods would assess whether the model fits from VI are problematic. GPflow does offer a Hamiltonian Monte Carlo sampling algorithm, but it is still experimental and also prone to instability. Using another probablistic programming language with reliable MCMC methods such as Stan's No-U-Turn Sampler (NUTS) would be more reliable and also likely more efficient than Hamiltonian Monte Carlo.

5.1.3 Scalability vs. Reliable and Full Inference

Another benefit of MCMC is that it will return estimates of full distributions for all model parameters. This would offer much more complete information for uncertainty-based decision making when using GP models for prediction and forecasting. For example, full distributions for each latent risk function $f(s,t)$ could be used to compare relative risk outcomes and allocate resources for e.g., traffic calming measures like additional enforcement based on an acceptable risk tolerance.

Gaussian Processes are slow to fit with MCMC, since the Cholesky Decomposition has to be done at every iteration of the algorithm. This makes doing full inference slow in some cases and prohibitively complex as the number of dimensions grows. There are some workarounds to this, such as exploiting Kronecker Algebra for fast computations (Flaxman, Wilson, Neill, Nickisch, & Smola, 2015), but the dimensionality is still limited to the thousands. While this is more than enough for most applications it does prevent scaling local spatial analysis to cover an area the size of a city while making estimates for relatively small grid squares. One promising alternative is making use of Bochner's theorem, which guarantees any valid stationary kernel can be represented using a Fourier transform (Rasmussen & Williams, 2005):

$$k(\tau) = \int_{\mathbb{R}^{\mathbb{D}}} e^{2\pi i s \tau} d\mu(\mathbf{s})$$

Hensman 2018 exploits this for a method called Variational Fourier Features that is capable of estimating high-dimensional models with millions of datapoints in minutes without specialized computing hardware (Hensman, Durrande, & Solin, 2016). Flaxman 2018 exploits Fourier features as well to apply a Poisson likelihood grid model for forecasting crime for the entire city of Portland, Oregon (Flaxman et al., 2018). Since the spatiotemporal LGCP models considered in this project assumed a stationary kernel by design they would be feasible candidates for using Variational Fourier Features while scaling to larger geographies. Avoiding relying on specialized GPU hardware would be another benefit.

5.2 Other Potential Data Sources

Other urban datasets may be just as- or even better suited to prediction using Guassian Processes. Many public issues where data are collected do not suffer from the sparsity issues of traffic collisions when making predictions at high resolution. Ubiqitious city annoyances like noise and pests are observed far more frequently than traffic collisions and would also benefit from reliable methods for prediction and comparative risk. Outside of the public sphere, taxi and ride-share vehicle demand are good candidates for heavily localized events which would benefit from granular forecasting.

5.3 Potential for Bias

The data used for modelling in this project were deliberately limited to the count data of interest, without relying on supplementary predictive or controlling variables. One obvious side effect of this choice that neverless bears stating directly is any bias in the data will be diligently reproduced in the predictions. Here 'bias' is meant in the sense of predictions skewed toward existing inaccuracies in the data provided rather than the technical definition of statistical bias.

This is especially relevant considering ongoing research into the performance of predictive algorithms when trained on data which had reporting bias. An illustrative example is a 2016 study of predictive policing systems in Oakland, California (Lum & Isaac, 2016). Lum and Isaac found that the proprietary algorithm provided by a private contractor replicated , rather than controlled for, reporting bias in the training data stemming from historical overpolicing of black

Americans for drug offenses. The models used here have also been used for experiments using policing data (Flaxman, 2014) (Flaxman et al., 2018). The risks of biased predictions based on biased data are not as visible outside of policing, but they should not be ignored when there is potential to practically apply these spatiotemporal models.

5.4 Measuring Distance

This project, like the previous research it was based on, relied on simple Euclidean distance in the spatial kernel. Euclidean distance is an excellent ’one-size-fits-all’ distance measure for general use, but other measures may be even more suited for predicting events taking place on a street grid. Manhattan distance would be a first alternative, and there is also the potential to use the actual street grid for measuring distance, see 5.6 for a representative example. This would also necessitate getting rid of the square grid structure, which would make the LGCP intractable (Teng et al., 2017). However, alternatives already exist and may be suitable for use in this case (Simpson et al., 2016). Finally, there is potential to use a distance measure that does not rely on geography at all; something like a social network could also be used as the spatial component.

5.5 Future Work

There are many possibilities for improvement and refinement. The most valuable would be to move to full Bayesian inference. It is possible to write one of the tested models in an MCMC friendly language such as Stan, which would be capable of generating posterior samples for all variables and parameters of interest rather than relying only on approximated point estimates, opening the door for explicit relative risk comparisons based on the modelled uncertainty. MCMC estimates would also provide a 'sanity check' on the variational approximations and potentially explain why model performance is sometimes poor or unstable. Practically speaking, moving to MCMC would also negate the need for a specialized computing environment with a GPU. if the MCMC results are promising, a further next step would be to experiment with scale by transferring the stationary models to Variational Fourier Features. With VFF it would become trivial to create grid models with high resolution that cover entire cities instead of just one neighborhood and fit them quickly on standard computer hardware.

A good portion of the necessary data manipulations could also be automated and abstracted. Automating the creation of the spatial grid, aggregating counts, re-centering variables, and other mundane tasks would make it much easier to experiment on different datasets quickly and efficiently.

5.6 Conclusion

The goal of this research was to evaluate the feasibility of spatiotemporal prediction and forecasting using a model framework that would be relatively easy to use ”out of the box” with only space and time data for each observation. Gaussian Process models are both theoretically and practically suitable in this case. The latent function structure of the Log Gaussian Cox Process provides direct estimates and predictions as well as a comparative statistic to gauge the relative prevalence of a phenomenona between places and times.

The experimental results show that the LGCP performs as well or better at prediction than a standard autoregression as measured by Mean Squared Error in long-, medium-, and short-term cases. The Reduction-in-Variance estimates show that the fitted models explain variations in spatial timeseries far better than relying on simple averages (an admittedly naive comparison), especially in the long-term model. The decomposed variance components also clearly delineate how time, space, seasonality, and long-term trends influence the total prediction.

While Gaussian Process models meet all the practical criteria for interpretable estimates and ease-of-use, the experiments also highlight some instability in fitting and unreliability in results. The worrisome PSIS-LOO scores as well as too-narrow credible intervals should not be ignored, especially when relying on variational inference methods that may not always suitably approximate the posterior distribution.

Still, Gaussian Process models are general and flexible enough to work with data at the city, neighborhood, and even hyperlocal level while also allowing flexible time inputs. There is potential to apply the same models to other uses of practical interest to urban planners, public safety officials, businesses, and anyone else who has some stake in an urban environment. Hopefully this project is a small step towards showing that ’cities are already smart.’

Appendices

.1 Additional Figures

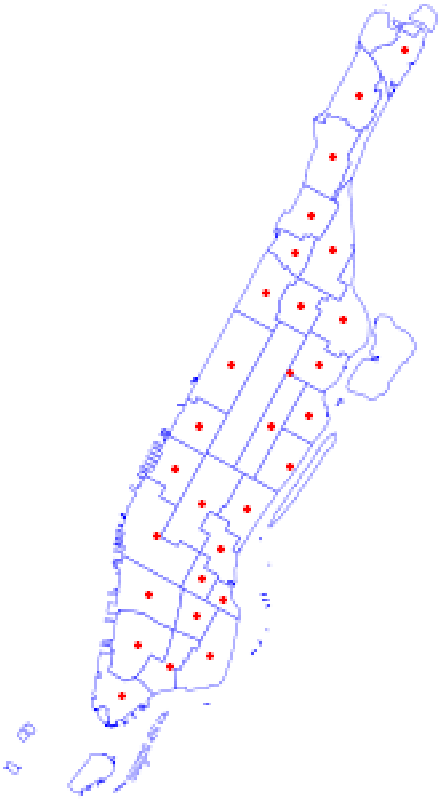


Figure 1 *Manhattan's 29 neighborhoods, with red dots indicating the centroid used for calculating distance.*

Figure 2 *Aggregated prediction of all Manhattan neighbhorhoods.*

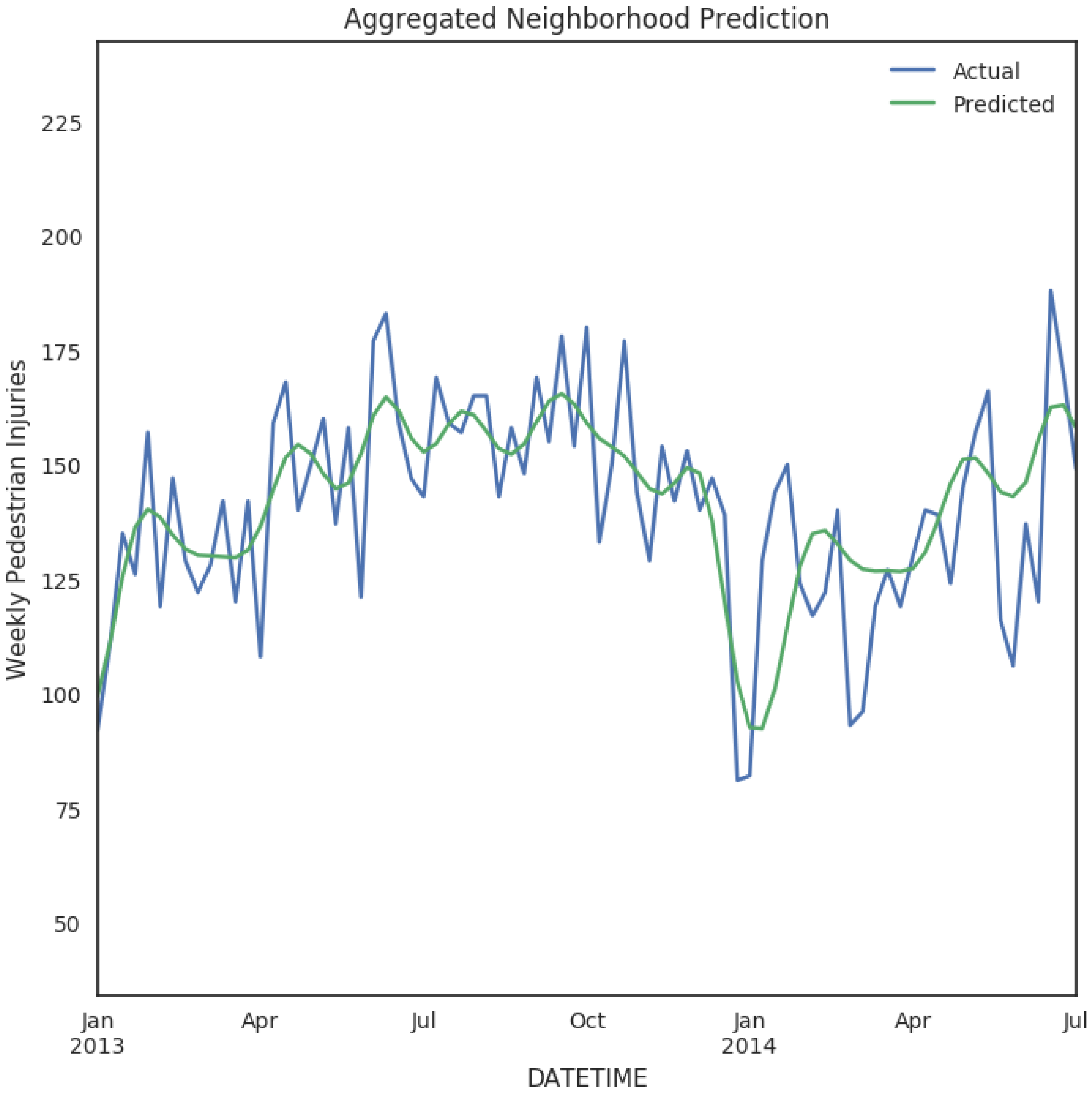
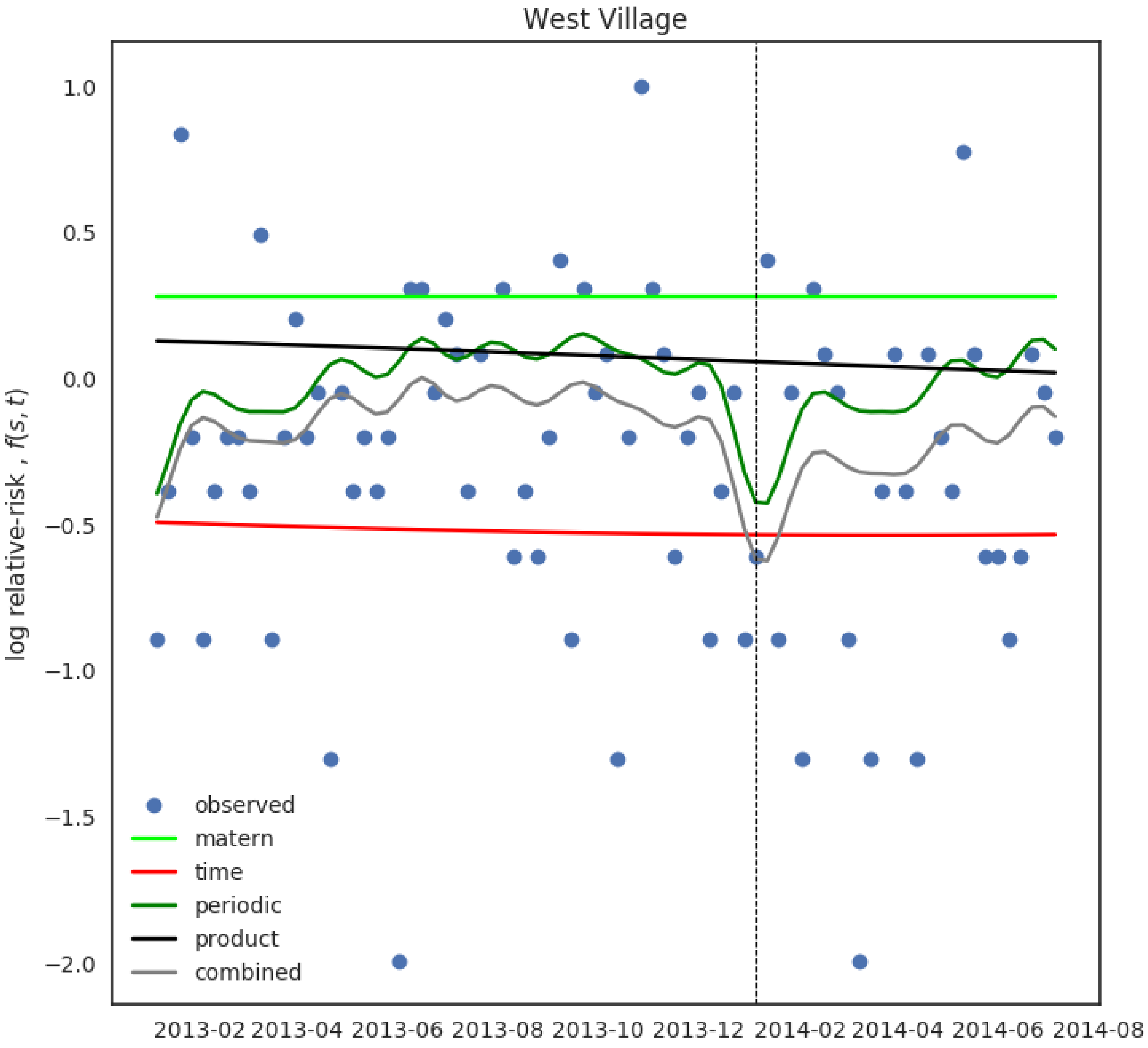


Figure 3 *Kernel component summary for the West Village neighborhood of Manhattan.*



.2 Variance Decomposition

Making new predictions $f^{\star}$ with a variational Gaussian approximation is done by integrating:

$$q\left(f^{\star}|\mathbf{y}\right) = \int p\left(f^{\star}|\mathbf{f}\right) q(\mathrm{f})\mathrm{d}\mathbf{f}$$

$$q\left(f^{\star}|\mathbf{y}\right) = \int \mathscr{N}\left(f^{\star}|\mathbf{K}_{\star f}\mathbf{K}^{-1}\mathbf{f}, \mathbf{K}_{\star\star} - \mathbf{K}_{\star f}\mathbf{K}^{-1}\mathbf{K}_{f\star}\right) \mathscr{N}(\mathbf{f}|\mathbf{K}\alpha, \blacksquare)\mathrm{d}\mathbf{f}$$

$$q\left(f^{\star}|\mathbf{y}\right) = \mathscr{N}\left(f^{\star}|\mathbf{K}_{\star\mathrm{f}}\alpha, \mathbf{K}_{\star\star} - \mathbf{K}_{\star f}(\mathbf{K}^{-1} - \mathbf{K}^{-1}\blacksquare\mathbf{K}^{-1})\mathbf{K}_{f\star}\right)$$

Here $\mathbf{K}_{\star\mathrm{f}}$ is the covariance between points used for prediction and the held-out points. $\mathbf{K}_{\star\star}$ is the covariance of the held out points and $\mathbf{K}$ is the covariance of the prediction points. The full derivation is available online (*Derivations of Variational Gaussian Process*, n.d.).

To make predictions using only some subset kernel/s of the full covariance it is only necessary to replace $\mathbf{K}_{\star\mathrm{f}}$ with that subset covariance K_s.

References

(n.d.).

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* Retrieved from `https://www.tensorflow.org/` (Software available from tensorflow.org)

Auerbach, J. (2017, jul). *De blasio wants to dramatically reduce nycs rat population. dont hold your breath.* Retrieved from `http://www.slate.com/articles/health_and_science/science/2017/07/de_blasio_s_team_consistently_misreads_the_stats_when_broadcasting_victories.html`

Blangiardo, M., & Cameletti, M. (2015). *Spatial and spatio-temporal bayesian models with r - inla.* Wiley. Retrieved from `https://books.google.com/books?id=QQ_QBwAAQBAJ`

CARROLL, C., & JOHNSON, D. S. (2008). The importance of being spatial (and reserved): Assessing northern spotted owl habitat relationships with hierarchical bayesian models. *Conservation Biology*, *22*(4), 1026–1036. Retrieved from `http://dx.doi.org/10.1111/j.1523-1739.2008.00931.x` doi: 10.1111/j.1523-1739.2008.00931.x

Cheng, T., Haworth, J., & Wang, J. (2012, Oct 01). Spatio-temporal autocorrelation of road network data. *Journal of Geographical Systems*, *14*(4), 389–413. Retrieved from `https://doi.org/10.1007/s10109-011-0149-5` doi: 10.1007/s10109-011-0149-5

CHUN, Y., KIM, Y., & CAMPBELL, H. (n.d.).

Derivations of variational gaussian process. (n.d.). Retrieved from `http://gpflow.readthedocs.io/en/latest/notebooks/vgp_notes.html`

Flaxman, S. (2014). A general approach to prediction and forecasting crime rates with gaussian processes..

Flaxman, S., Chirico, M., Pereira, P., & Loeffler, C. (2018, January). Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: a winning solution to the NIJ "Real-Time Crime Forecasting Challenge". *ArXiv e-prints.*

Flaxman, S., Gelman, A., Neill, D., Smola, A. J., & Vehtari, A. (2015). Fast hierarchical gaussian processes..

Flaxman, S., Wilson, A. G., Neill, D. B., Nickisch, H., & Smola, A. J. (2015). Fast kronecker inference in gaussian processes with non-gaussian likelihoods. In *Proceedings of the 32nd international conference on international conference on machine learning - volume 37* (pp. 607–616). JMLR.org. Retrieved from `http://dl.acm.org/citation.cfm?id=3045118.3045184`

Foreman, K. J., Li, G., Best, N., & Ezzati, M. (2017). Small area forecasts of cause-specific mortality: application of a bayesian hierarchical model to us vital registration data. *Journal of the Royal Statistical Society Series C*, *66*(1), 121-139. Retrieved from `https://EconPapers.repec.org/RePEc:bla:jorssc:v:66:y:2017:i:1:p:121-139`

Gelman, A. (2006, 09). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Anal.*, *1*(3), 515–534. Retrieved from `https://doi.org/10.1214/06-BA117A` doi: 10.1214/06-BA117A

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis, third edition.* CRC Press. Retrieved from `https://books.google.com/books?id=eSHSBQAAQBAJ`

Hensman, J., Durrande, N., & Solin, A. (2016, November). Variational Fourier features for Gaussian processes. *ArXiv e-prints.*

Illian, J. B., Sørbye, S. H., & Rue, H. (2013, January). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *ArXiv e-prints.*

Kitchin, R. (2014, Feb 01). The real-time city? big data and smart urbanism. *GeoJournal*, *79*(1), 1–14. Retrieved from `https://doi.org/10.1007/s10708-013-9516-8` doi: 10.1007/s10708-013-9516-8

Li, G., Haining, R., Richardson, S., & Best, N. (2014). Spacetime variability in burglary risk: A bayesian spatio-temporal modelling approach. *Spatial Statistics*, *9*, 180 - 191. Retrieved from `http://www.sciencedirect.com/science/article/pii/S2211675314000190` (Revealing Intricacies in Spatial and Spatio-Temporal Data: Papers from the Spatial Statistics 2013 Conference) doi: https://doi.org/10.1016/j.spasta.2014.03.006

Liesenfeld, R., Richard, J.-F., & Vogler, J. (2017). Likelihood-based inference and prediction in spatio-temporal panel count models for urban crimes. *Journal of Applied Econometrics*, *32*(3), 600–620. Retrieved from `http://dx.doi.org/10.1002/jae.2534` (jae.2534) doi: 10.1002/jae.2534

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, *63*(19).

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19.

MacKay, D. J. (1998). Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, *168*, 133–166.

Matthews, A. G. d. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., ... Hensman, J. (2017, apr). GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, *18*(40), 1-6. Retrieved from `http://jmlr.org/papers/v18/16-537.html`

Murray, I., Ghahramani, Z., & MacKay, D. (2012, June). MCMC for doubly-intractable distributions. *ArXiv e-prints*.

"new york city open data: Motor vehicle collisions". (n.d.). Retrieved from `https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95`

of Health, N. Y. S. D. (n.d.). *Traffic injury statistics.* Retrieved from `https://www.health.ny.gov/statistics/prevention/injury_prevention/traffic/county_of_residence.html`

of Transportation, N. Y. C. D. (n.d.). *New york city vision zero.*

of Transportation, N. Y. C. D. (2018, apr). *Vision zero map.* Retrieved from `https://www.nycvzv.info`

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian processes for machine learning (adaptive computation and machine learning).* The MIT Press.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, *71*(2), 319–392.

Schrdle, B., & Held, L. (2011). Spatio-temporal disease mapping using inla. *Environmetrics*, *22*(6), 725–734. Retrieved from http://dx.doi.org/10.1002/env.1065 doi: 10.1002/env.1065

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H., & Rue, H. (2016). Going off grid: Computationally efficient inference for log-gaussian cox processes. *Biometrika*, *103*(1), 49–70.

Stan Development Team. (2018). *RStan: the R interface to Stan.* Retrieved from http://mc-stan.org/ (R package version 2.17.3)

Tascikaraoglu, A. (2018). Evaluation of spatio-temporal forecasting methods in various smart city applications. *Renewable and Sustainable Energy Reviews*, *82*, 424 - 435. Retrieved from http://www.sciencedirect.com/science/article/pii/S1364032117313308 doi: https://doi.org/10.1016/j.rser.2017.09.078

Teng, M., Nathoo, F. S., & Johnson, T. D. (2017, January). Bayesian Computation for Log-Gaussian Cox Processes–A Comparative Analysis of Methods. *ArXiv e-prints*.

Tran, D., Ranganath, R., & Blei, D. M. (2015, November). The Variational Gaussian Process. *ArXiv e-prints*.

Vehtari, A., Gelman, A., & Gabry, J. (2015, July). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *ArXiv e-prints*.

Wakefield, J. (2006). Disease mapping and spatial regression with count data. *Biostatistics*, *8*(2), 158–183.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018, February). Yes, but Did It Work?: Evaluating Variational Inference. *ArXiv e-prints*.