PREDICTING AMBULANCE DEMAND

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Zhengyi Zhou

August 2015

PREDICTING AMBULANCE DEMAND

Zhengyi Zhou, Ph.D.

Cornell University 2015

Predicting ambulance demand accurately on a fine resolution in time (e.g., every hour) and space (e.g., every 1 km^2) is critical for staff, fleet management and dynamic deployment. There are several challenges: although the dataset is typically large-scale, the number of observations per time period and locality is almost always zero. The demand arises from complex urban geography and exhibits complex spatio-temporal patterns, both of which we need to capture and exploit. We propose three new methods to address these challenges, and provide spatio-temporal predictions for Toronto, Canada and Melbourne, Australia.

First, we introduce a Bayesian time-varying Gaussian mixture model. We fix the mixture component distributions across time, while representing the spatio-temporal dynamics through time-varying mixture weights. We constrain the weights to capture weekly seasonality, and apply autoregressive priors on them to model location-specific patterns.

Second, we propose a spatio-temporal kernel density estimator. We weight the spatial kernel of each historical observation by its informativeness to the current predictive task. We construct spatio-temporal weight functions to incorporate various temporal and spatial patterns in ambulance demand.

Third, we propose a kernel warping method to incorporate complex spatial features. For each prediction we build a kernel density estimator on a sparse set of most similar data (labeled data), and warp these kernels to a larger set of past

data regardless of labels (point cloud). The point cloud represents boundaries, neighborhoods, and road networks. Kernel warping can be interpreted as a regularization and a Bayesian prior imposed for spatial features.

We show that these methods give much higher statistical predictive accuracy, and reduce error in predicting EMS operational performance by as much as two-thirds compared to the industry practice.

BIOGRAPHICAL SKETCH

Zhengyi Zhou was born in Chengdu China, despite abnormal Down syndrome markers and passing of her father.

She grew up in an academic and loving family with plenty of high expectations for music, painting, and mathematics. After a chance application and 3 days of IQ tests, she became one of the five scholarship winners to attend the best middle and high schools in Singapore.

After spending half of her time in Singapore studying math and "further math", she decided to go to a Liberal Arts college, Occidental College. She majored in Economics. And Mathematics, because it was too good to pass up so many transfer credits. Her bachelor thesis empirically studied human risk preferences and irrationality, and concluded that in those matters, one is born, not made.

She chose Cornell Center for Applied Mathematics for her PhD studies. She met her husband Philipp on her first day in Ithaca. She very much enjoyed teaching, the waterfalls, complaining about the cold, and researching the problems presented in this thesis.

To my mother, for supplying immeasurable love

and to Philipp, for providing vital buoyancy

ACKNOWLEDGEMENTS

This thesis would not have taken shape, and certainly would not have made me proud, if it were not for my advisor David Matteson. He gave me more guidance, help and encouragement that I could ask for. My other co-authors Dawn Woodard, Shane Henderson and Sakis Micheas have extended themselves to help me in essential ways. John Guckenheimer sparked my interest in computation and numerical analysis, and offered much advice amid uncertainty. Before I came to Cornell, Sita Slavov and Gregory Tollison at Occidental College planted the intellectual seed that started me off on this path. My family has given me so much unwavering love and support. CAM friends and my students have given me a great time. I am very grateful for this incredible journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

1.1.1 Motivation

A primary goal of emergency medical services (EMS) is often to minimize response times to emergencies while managing operational costs. Sophisticated operations research methods have been developed to optimize many management decisions, such as locations of bases, fleet size, staffing, and dynamic deployment strategies [79, 33, 38]. However, these methods require ambulance demand estimates as inputs, and their performances rely critically on the accuracy of these demand estimates. Demand predictions that are too high lead to over-staffing, unnecessary vehicles and high cost, while estimates that are too low result in slow response times to potentially life-threatening emergencies.

In practice, two types of demand estimates are of interest: aggregate temporal demand, i.e., total expected demand volume, and spatio-temporal demand, or the spatial distribution of demand over time. Temporal aggregate demand estimates inform effective staffing and fleet planning; spatio-temporal estimates are critical for choosing base locations and dynamic deployment strategies. These estimates are ideally needed at high temporal resolution (e.g., four-hour work shifts). Similarly, spatio-temporal estimates at fine spatial granularities are required for accurate dynamic deployment. The industry typically predicts for every hour and every 1 km^2 region.

1.1.2 Challenges

There are several typical challenges to predicting ambulance demand.

- Ambulance demand is often exceedingly sparse at the temporal and spatial resolution required for prediction. For example,Toronto receives only 23 calls per hour on average; 96% of the 1-km^2 spatial regions have zero calls in any hour. Similarly, Melbourne only receives 37 calls per hour; 99.6% of the 1-km^2 regions receive no calls in any hour.
- This demand arises from complex urban geography. The city boundary is often highly irregular. Ambulance demand can be very high (coastal and downtown) or very low (suburbs) along the boundary. Within this boundary, demand follows closely the city's infrastructure and terrain; there might be high demand along central highways and zero demand within an internal lake. High resolutions covariates of these features are often not readily available.
- There are notable spatial and temporal patterns in ambulance demand. Usually, weekly seasonality is prominent [21, 50]; the industry relies heavily on this seasonality to make predictions. In the case of Toronto and Melbourne, we have noted stronger weekly, daily seasonalities and short-term serial dependence at densely-populated regions than other locations [98, 95, 97].
- Ambulance demand data for large cities is often large-scale. Toronto receives about 200,000 calls per year, and Melbourne receives more than 330,000. This presents computational challenges, especially since predictions are needed very frequently.

It is particularly difficult to simultaneously resolve these challenges. Overcoming sparsity requires considerable smoothing, while capturing complex spatio-temporal patterns requires fine-resolution modeling. At high granularities, data sparsity makes it difficult to detect spatio-temporal characteristics accurately. At low granularities, differences across regions and times are not sufficiently captured for optimal ambulance planning.

1.1.3 Industry methods

The current industry practice for predicting ambulance demand often uses a simple averaging formula. Demand in a 1 km^2 spatial region over an hour is typically predicted by averaging a small number of historical counts, from the same spatial region and over the corresponding hours from previous weeks or years [33]. In current practice, Toronto EMS averages four historical counts in the same hour of the year over the past four years, while the EMS of Charlotte-Mecklenburg, North Carolina averages twenty historical counts in the same hour of the preceding four weeks for the past five years (MEDIC method) [70]. Averaging so few historical counts, which are mostly zeros, produces highly noisy and flickering predictions, resulting in haphazard and inefficient deployment. Such methods are also sensitive to how the spatial domains are partitioned [33].

1.1.4 Overview

The remainder of this dissertation is organized as follows. We survey relevant literature in Section 1.2 and introduce the data in Section 1.3. In Chapter 2, we present a new time-varying Gaussian mixture model for predicting spatio-temporal ambulance demand. We describe the model (2.1), computation of the model (2.2), additional model refinements (2.3), results on Toronto ambulance demand data (2.4), and model performance (2.5). In Chapter 3, we propose a spatio-temporal weighted kernel density estimation method for fast and accurate spatio-temporal ambulance demand prediction. We introduce the model (3.1), its computation (3.2), and results on Toronto ambulance demand data (3.3). In Chapter 4, we propose a kernel warping method to predict spatio-temporal ambulance demand on complex spatial domains. We construct the kernel warping model in 4.1, and show the results on Melbourne ambulance demand data in 4.2. Chapter 5 concludes and discusses these methods, and look forward to future research in ambulance demand prediction and in general, data mining in

healthcare, operations and business.

1.2 Literature

This section surveys the relevant literature. We review some current approaches in predicting ambulance demand in 1.2.1. Then we introduce literature on spatio-temporal Poisson point process (1.2.2), Bayesian mixture modeling (1.2.3), and kernel density estimation (1.2.4). We review literature on manifold learning and kernel warping (1.2.5).

1.2.1 Ambulance demand prediction

Several studies have modeled aggregate ambulance demand as a temporal process. Channouf et al model daily total ambulance demand in Calgary, Canada using Gaussian autoregressive moving-average models with seasonality and special day effects (day-of-week, month-of-year, fixed day-month interactions) [21]. Hourly demand is then obtained by using hour-of-day effects and assigning a multinomial distribution, conditional on the daily total. Matteson et al directly model hourly call arrival rates in Toronto, Canada by combining a dynamic latent factor structure with integer time series models [50]. Covariate information is captured as constraints on the factor loadings, and smoothing splines are applied to the factor levels and loadings. Other aggregate demand studies for ambulance demand or the closely related problem of call center demand have also considered dynamic harmonic regression [86], spectral analysis [88], fixed-effects, mixed-effects and bivariate models [2, 39], Bayesian multiplicative models [90], Singular Value Decomposition [71], and Cox processes [72].

While these temporal estimates inform staffing and fleet size, spatio-temporal demand estimates are critical for selection of base locations and for dynamic deployment planning, but have received far less attention. Setzler,

Saydam and Park use artificial neural networks (ANN) on discretized spatial and temporal domains, and compare it to industry practice [70]. ANN is superior at low spatial granularity, but both methods produce noisy results at high spatial resolutions.

1.2.2 Spatio-temporal Poisson point process

Spatial point processes have frequently been modeled using non-homogeneous Poisson processes (NHPP) [23, 55, 40]. In particular, Bayesian semi-parametric mixture modeling has been proposed to account for heterogeneity in the spatial intensity function. Examples include Dirichlet processes with beta or Gaussian densities [47, 42], and finite Gaussian mixture models with a fixed number of components [20]. However, EMS data is sparse at the desired temporal granularity for estimation in this industry; the average number of observations in each two-hour period is only 45. This makes it difficult to estimate an accurate spatial structure at each time period.

Recently, dependent Dirichlet processes have been proposed to model correlated spatial densities across discrete time [80, 81, 24, 82]. These methods allow the stick-breaking weights of the Dirichlet process to evolve in an autoregressive manner, but necessitates a simple first-order dependence structure common to all components. For EMS applications, it is essential to capture a much more complex set of temporal dynamics, including short-term serial dependence as well as daily and weekly seasonalities. Moreover, some of these dynamics vary from location to location. To consider only the first-order dependence, and enforce it across the entire spatial domain may be very limiting. On the other hand, extending the dependent Dirichlet processes to include higher-order serial dependence and multiple seasonalities is not straightforward. It is also not easy to make these dynamics location-specific. Discretizing the spatial domain into sub-regions and imposing a different autoregressive parameter on each region would add substantial computational complexity, and is sensitive to spatial partitioning. Furthermore, given the large amount of data and the large number of

time periods considered, using an infinite-dimensional Dirichlet process can be computationally challenging.

All the methods we propose adopt the NHPP framework and aim at capturing complex spatial and temporal patterns in flexible and efficient ways.

1.2.3 Bayesian mixture modeling

Mixture models provide a convenient framework to model data that originates from different groups or cannot be modeled well by a single parametric distribution. Using a fixed number of components, we can carry out Bayesian estimation of mixture models via Gibbs sampling and data augmentation [83]. One major computational challenge is lack of good convergence and mixing [49]. Components with small number of observations and small variances produce very concentrated modes, and there is very low probability of escaping from these modes.

Using a variable number of components mitigates this difficulty to some extent, and offers even more flexibility in modeling. One popular approach is Richardson and Green's reversible jump Markov chain Monte Carlo (RJMCMC) [37, 64]. In RJMCMC, one periodically proposes a move to a different model and rejects that proposal with the appropriate probability to ensure stationarity of the process. Three types of moves are possible: birth-and-death of components, split-and-combine of components, and fixed-component update.

Stephens proposes an alternative scheme by constructing a continuous time birth-and-death Markov chain Monte Carlo (BDMCMC). Each iteration of Stephens' BDMCMC is a two-stage process. In the first stage, new components are "born" or existing ones "die" in a continuous time framework. New components are born in a constant rate; parameters of new-born components are sampled from their respective priors. Components die at a rate so as to maintain sampling stationarity; they die according to their relative implausibility as computed from the likelihood of observations and priors. After each

birth or death, the mixture weights are scaled proportionally to maintain sum-to-one invariance within each time period. After a fixed duration of births and deaths, in the second stage, the number of components are fixed and distributional parameters and mixture weights of the components are updated using data augmentation, Gibbs sampling or Metropolis-Hastings.

Both RJMCMC and BDMCMC facilitate better mixing and convergence of parameters. Allowing the dimension of the mixture model to vary helps mixing of parameters within each configuration. RJMCMC and BDMCMC are similar frameworks; one can construct a sequence of RJMCMC samplers that converges to BDMCMC, and vice versa [19]. Although BDMCMC is more computationally expensive in each iteration than RJMCMC, it is shown to have a greater ability to move to very unlikely spaces, since births are always accepted regardless of data. This leads to slightly improved mixing of all parameters [78, 19]. BDMCMC has been used for spatial and spatio-temporal pattern recognition [53, 52].

These studies have also suggested choices for priors for mixture model Bayesian estimation. Richardson and Green define a set of independent, weakly informative and hierarchical priors conjugate to univariate Gaussian mixture models [64], which Stephens extends to the multivariate case [78].

The method we propose in Chapter 2 uses Bayesian mixture modeling and BDMCMC.

1.2.4 Spatio-temporal kernel density estimation

Kernel density estimation (KDE) is a powerful tool for non-parametric density estimation in spatio-temporal data. It has been widely applied to visualize or forecast spatio-temporal crime incidence [16, 56], disease spread [93, 91], product demand [41], and data streams [1, 59]. It allows for rapid visualization and identification of hotspots and their evolutions in time and space. In most cases, the time dimension is treated differently from the space dimension(s). The most traditional approach is to build a separate spatial KDE for each discrete time pe-

riod. However, this approach may result in uneven subset size and sparse subsets with too little data for accurate density estimates. Recent studies assume a multiplicative orthogonal relationship between the time and space dimensions. For example, Aggarwal multiplies a spatial kernel with a linear function in time [1]. Studies such as [16, 56, 93] multiply a spatial kernel and a temporal kernel with different bandwidths and kernel functions for the two kernels. The method we proposed in Chapter 3 is an extension of these multiplicative spatio-temporal KDE methods [95].

We can select the bandwidth of KDE via the plug-in method [89] or smoothed cross-validation [44, 26]. We typically use the Gaussian kernel, or for additional computational savings, the Epanechnikov kernel with bounded support. There are many fast computational methods for KDE, including KD-trees [10], ball trees [58], dual trees [36] and statistical regular pavings [68]. When data show large variation in density, using one fixed bandwidth may not be optimal [75, 17, 69]. A bandwidth too large wipes out local features where we have sufficient data; a bandwidth too small leads to spurious peaks where data is sparse. Ambulance demand varies substantially in space (downtown vs. suburbs) and time (midnight vs. rush hours); we may be motivated to consider a spatial- and/or time-varying bandwidth.

The methods we propose in Chapter 3 and 4 are extensions on this type of spatio-temporal KDEs.

1.2.5 Kernel warping

Few studies have focused on modeling spatial or spatio-temporal point processes on complex spatial structures. Most studies assume a boundary defined *a priori* (either polygon or pixelated). If not, *ad hoc* methods based on the convex hull of all observed points are typically used [65, 5]. This invariably results in a convex boundary that may be inaccurate where data is sparse. Even with a boundary optimally defined, few methods are equipped to handle complex

boundary features. Ramsey proposes a finite window smoother with known boundary conditions computed using an expensive finite element approach [60]. Building on that, Wood, Bravington and Hedley model the boundary condition as a loop of wire and the point process as a soap film suspended from the boundary wire [92]. They represent this smoother as a penalized basis, compute it via multi-grid, and select smoothness via generalized cross-validation. They acknowledge the lack of an elegant solution when the boundary conditions are unknown. Apart from boundary, other geographical characteristics are rarely incorporated in modeling. In Chapter 4, we propose a method that can efficiently capture and exploit a wide range of spatial characteristics. We draw from theory and methods developed in manifold learning.

Manifold learning, a branch of machine learning, is concerned with learning and exploiting the underlying structures of data. The assumption is that data in a high-dimensional space resides on or near a lower-dimensional sub-manifold. In practice, we do not have access to this sub-manifold, but we can approximate it from a point cloud, i.e., a mass of historical data. The most common method is to construct an adjacency graph of this point cloud and make use of the properties and structures of this graph. This idea has led to many popular learning methods, including isomap [84], local linear embedding [67], Hessian eigenmaps [25], and Laplacian eigenmaps [7] (see [87] for some review). These methods have been initially designed for data representation or visualization, but have been adapted for semi-supervised classification [8, 99, 94], and clustering [57, 73].

In particular, a variant of Laplacian eigenmaps, kernel warping, has been proposed for semi-supervised classification [77, 8, 76, 9]. Using a small number of labeled data and a larger number of point cloud data (labeled and unlabeled), the method classifies new examples by constructing kernels on the labeled data that warp to the geometry of the point cloud. This geometry is represented by the adjacency graph of the point cloud. Smoothing orthogonal to this geometry is penalized heavily, whereas smoothing along this geometry is not. This method is designed for high-dimensional classification, and has good perfor-

mance on text and image data.

1.3 Data

We use ambulance demand data from Toronto, Canada and Melbourne, Australia. The Toronto data is from Toronto Emergency Medical Services, for years 2007 and 2008. The data consist of $391,296$ priority emergency events received by Toronto EMS for which an ambulance was dispatched. Each record contains the time and the location to which the ambulance was dispatched. This includes some calls not requiring lights-and-sirens response, but does not include scheduled patient transfers. We include only the first event in our analysis when multiple responses are received for the same event; explanatory analysis did not reveal any spatial or temporal pattern for these cases, and we treat them as a single ambulance dispatch. We have removed all redundant calls and calls with no locations. There was no call received for more than two hours on March 10, 2007 due to a recording system malfunction, and we have also removed all calls from that day. These removals totaled less than 2% of the data.

Figure 1.1 (a) shows all $15,393$ observations from February 2007. Figure 1.1 (b) shows all 31 observations from 8-10 am February 1 2007. Data is considerably sparse at the resolution of 2-hour periods; the average number of observations per 2-hour period is only 45.

Figure 1.2 explores some temporal characteristics at various locations in Toronto. For Figure 1.2 (a), we compute the proportions of observations that arise from the downtown region (outlined by the rectangle in Figure 1.1 (a)) out of all observations for each 2-hour period from February 2007. We analyze the autocorrelation of this time series of proportions. At downtown, we observe evidence for weekly (84 time periods) and daily (12 time periods) seasonality as well as low-order autocorrelation (dashed lines represent approximate 95% point-wise confidence intervals). For Figure 1.2 (b), we repeat this procedure at another locality in Toronto, but we only found evidence of weekly seasonality.

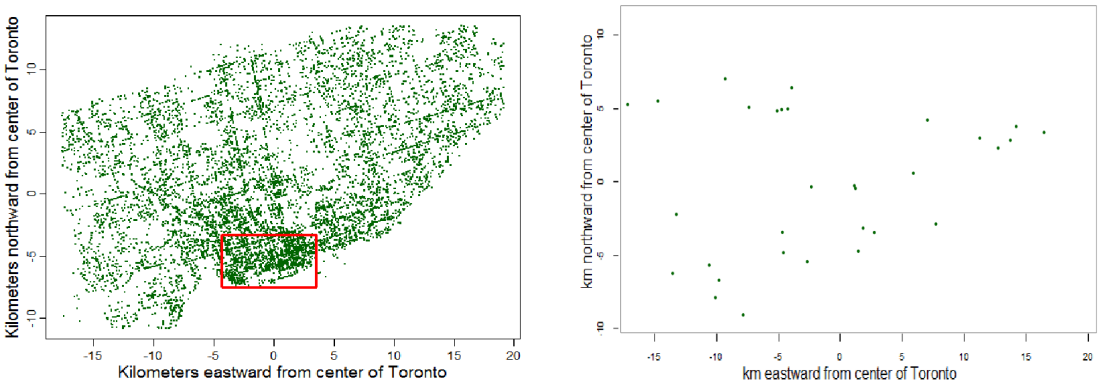

Figure 1.1: Left: all 15,393 observations in the training data (February 2007), with downtown subregion outlined by a rectangle. Right: all 31 observations from 8-10 am February 1 2007.

In general, we consistently found weekly seasonality, but daily seasonality and low-order autocorrelation tend to be stronger at locations such as downtown or dense residential regions, and weaker at others such as dispersed residential areas or large parks.

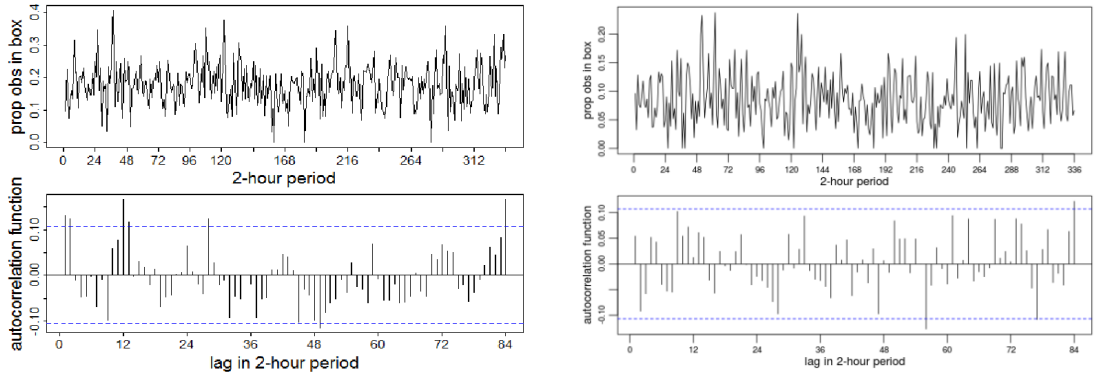

Figure 1.2: Left: Time series across 2-hour periods (top) and autocorrelation function (bottom) of the proportions of observations arising from the downtown Toronto (enclosed in rectangle in Figure 1.1 (a)). Weekly, daily seasonality, and low-order autocorrelation are observed. Right: Time series across 2-hour periods (top) and autocorrelation function (bottom) of the proportions of observations arising from another region from Toronto. Only weekly seasonality is observed.

We also use ambulance demand data from Melbourne Emergency Medical Services, for years 2011 and 2012. Again, each observation contains the time and the location to which the ambulance was dispatched. There are altogether 696, 975 observations. In Figure 1.3 (a), we show in gray the locations of 696,975 demand incidents for these two years and in black those of 38 demand incidents for a typical 1-hour period. Comparing with the map of Melbourne in Figure 1.3 (b) (map data: Google [35]), we observe a highly complex spatial boundary as Melbourne encloses a large bay to its southwest. Demand is high near the bay, but low on the outskirt suburban areas. Demand is visibly higher at small satellite suburban neighborhoods and along major highways radiating out from the city center to the suburbs. There is lack of demand due to several reservoirs and a national park to the west and northwest. Consistent with typical patterns, the demand exhibits strong weekly seasonality.

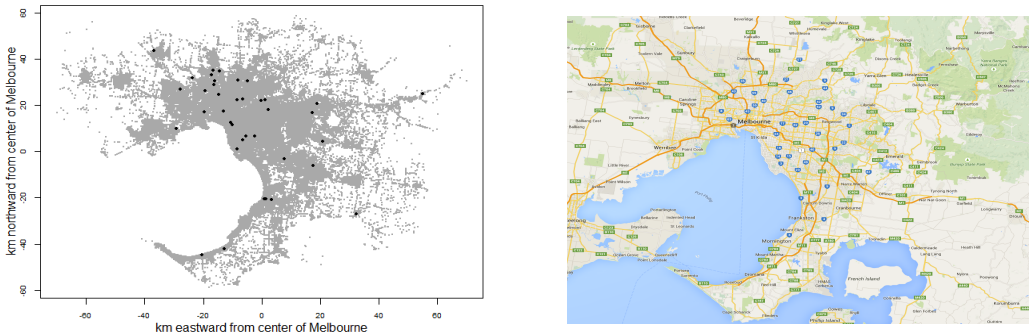

Figure 1.3: Left: spatial locations of all 696, 975 Melbourne ambulance demand incidents from years 2011 - 2012 (in gray), and 38 demand incidents for a typical 1-hour period (in black). We observe complex boundary and geographical features (e.g., highways, roads, satellite suburbs). Right: map of Melbourne [35].

CHAPTER 2
TIME-VARYING GAUSSIAN MIXTURE MODELS

We propose in this chapter a novel specification of a time-varying finite mixture model. We assume a common set of mixture components across time periods, to promote effective learning of the spatial structure across time, and to overcome sparsity within each period. We allow the mixture weights to vary over time, capturing temporal patterns and dynamics in the spatial density by imposing seasonal constraints and applying autoregressive priors on the mixture weights. The number of mixture components may be fixed or estimated via birth-and-death Markov chain Monte Carlo [78]. We compare the proposed method with a current industry practice, as well as a proposed extension of this practice. The proposed method is shown to have highest statistical predictive accuracy, as well as the least error in measuring operational performance.

Material from this chapter was accepted in August 2014 by the Journal of the American Statistical Association in an article titled "A Spatio-Temporal Point Process Model for Ambulance Demand", authored by Zhengyi Zhou, David S. Matteson, Dawn B. Woodard, Shane G. Henderson, and Athanasios C. Micheas [98]. This paper was the winner of ASA health policy statistics section student paper competition (2014) and a finalist of INFORMS data mining section student paper award (2013). Material from this chapter was also accepted in May 2014 as a chapter titled "Temporal and spatio-temporal models for ambulance demand" authored by Zhengyi Zhou and David S. Matteson. This is a chapter of the edited volume "Healthcare Data Analytics, Wiley Series in Operations Research and Management Science" edited by H. Yang and E.K. Lee [97].

We define the general setting and propose a spatio-temporal mixture model using a fixed number of components in Section 2.1. We discuss the computation in Section 2.2. We further extend the mixture model in Section 2.3 to assume an unknown number of components (Section 2.3.1), include covariates (Section 2.3.2), and incorporate Toronto's spatial boundary (Section 2.3.3). We show the results of estimating ambulance demand in Toronto in Section 2.4, and assess the

performance and validity of the proposed approach in Section 2.5. We conclude in Section 2.6.

2.1 Model

We investigate Toronto's ambulance demand on a continuous spatial domain $\mathcal{S} \subseteq \mathbb{R}^2$ and a discretized temporal domain $\mathcal{T} = \{1, 2, \ldots, T\}$ of two-hour intervals ($T = 336$ for 28 days in February 2007). In Section 2.3.3 we define $\mathcal{S}$ to lie within the boundary of Toronto. The proposed method trivially extends to other spatial domains.

Let $\boldsymbol{s}_{t,i}$ denote the spatial location of the ith ambulance demand event occurring in the tth time period, for $i \in \{1, \ldots, n_t\}$. We assume that the set of spatial locations in each time period independently follows a non-homogeneous Poisson point process over $\mathcal{S}$ with positive integrable intensity function λ_t. The intensity function for each period t can be decomposed as

$$\lambda_t(\boldsymbol{s}) = \delta_t f_t(\boldsymbol{s}), \tag{2.1}$$

for $\boldsymbol{s} \in \mathcal{S}$, in which $\delta_t = \int_{\mathcal{S}} \lambda_t(\boldsymbol{s})\, d\boldsymbol{s}$ is the aggregate demand intensity, or total call volume, for period t, and $f_t(\boldsymbol{s})$ is the spatial density of demand in period t, i.e., $f_t(\boldsymbol{s}) > 0$ for $\boldsymbol{s} \in \mathcal{S}$ and $\int_{\mathcal{S}} f_t(\boldsymbol{s})\, d\boldsymbol{s} = 1$. Then we have $n_t|\lambda_t \sim \text{Poisson}(\delta_t)$ and $\boldsymbol{s}_{t,i}|\lambda_t, n_t \overset{iid}{\sim} f_t(\boldsymbol{s})$, for $i \in \{1, \ldots, n_t\}$. Many prior studies propose sophisticated methods for estimating $\{\delta_t\}$. Here, we focus on estimating $\{f_t(\boldsymbol{s})\}$, which has received little consideration in the literature.

The proposed model is constructed in three steps. We first introduce in Section 2.1.1 the general framework of mixture models with common component distributions across time. We add constraints on the mixture weights in Section 2.1.2 to describe weekly seasonality. We also place autoregressive priors on the mixture weights to capture location-specific dependencies in Section 2.1.3. For now, we fix the number of mixture components K; estimation of K is incorpo-

rated in Section 2.3.1.

2.1.1 Gaussian mixture model

We consider a bivariate Gaussian mixture model in which the component distributions are common through time, while mixture weights change over time. Fixing the component distributions allows for information sharing across time to build an accurate spatial structure, because each time period typically has few observations. It is also natural in this application, which has established hotspots such as downtown, residential areas and central traffic routes. Letting the mixture weights vary across time enables us to capture dynamics in population movements and actions at different locations and times. The proposed methods can be trivially extended to other distributional choices such as a mixture of bivariate Student's t distributions. For any $t \in \mathcal{T}$, we model the spatial distribution by a K-component Gaussian mixture

$$f_t(\boldsymbol{s}; \{p_{t,j}\}, \{\boldsymbol{\mu}_j\}, \{\boldsymbol{\Sigma}_j\}) = \sum_{j=1}^{K} p_{t,j}\, \phi(\boldsymbol{s}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \quad \forall\ \boldsymbol{s} \in \mathcal{S}, \tag{2.2}$$

in which ϕ is the bivariate Gaussian density, with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$, for $j \in \{1, \ldots, K\}$. The mixture weights are $\{p_{t,j}\}$, satisfying $p_{t,j} \geq 0$ and $\sum_{j=1}^{K} p_{t,j} = 1$ for all t and j. In this specification, the number of components K is assumed fixed across time. However, components can be period-specific when their weights are 0 in all other periods. The component means and covariances are also the same in all time periods; only the mixture weights are time-dependent.

2.1.2 Constraints for seasonality

We observe weekly seasonality in ambulance demand across the spatial domain (Section 1.3); previous research has also confirmed that EMS total call counts

vary greatly with time of the day and day of the week, but change little from week to week [21, 50]. We represent this weekly seasonality by constraining all time periods with the same position within a week (e.g., all periods corresponding to Monday 8 - 10 am) to have common mixture weights.

Let $B \in \mathbb{N}$ ($B \ll T$) denote a time block, corresponding to the desired cycle length. In this application, $B = 84$, the number of 2-hour periods in a week. Each $t \in \mathcal{T}$ is matched to the value of $b \in \{1, \ldots, B\}$ such that $b \bmod B = t \bmod B$. We modify Equation (2.2) to

$$f_t(\boldsymbol{s}; \{p_{t,j}\}, \{\boldsymbol{\mu}_j\}, \{\boldsymbol{\Sigma}_j\}) = f_b(\boldsymbol{s}; \{p_{b,j}\}, \{\boldsymbol{\mu}_j\}, \{\boldsymbol{\Sigma}_j\}) = \sum_{j=1}^{K} p_{b,j}\, \phi(\boldsymbol{s}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \tag{2.3}$$

so that all periods with the same position within the cycle have the same set of mixture weights.

The usefulness of such constraints on mixture weights is not limited to representing seasonality. We can also exempt special times, such as holidays, from seasonality constraints, or combine consecutive time periods with similar characteristics, such as rush hours or midnight hours.

2.1.3 Autoregressive priors

We also observe that EMS demand exhibits low-order serial dependence and daily seasonality whose strengths vary with locations (Section 1.3). We may capture this in the proposed mixture model by placing a separate conditionally autoregressive (CAR) prior on each series of mixture weights, i.e., $\{p_{b,j}\}_{b=1}^{B}$ for each j in Equation (2.3). CAR priors are widely used in spatial analysis to encourage similar parameter estimates at neighboring locations [13, 6], and in temporal analysis to smooth parameter estimates at adjacent times [12, 46, 45].

With such priors, we can represent a rich set of dependence structures, including complex seasonality and high-order dependence structures, which may

be especially helpful for analyzing temporal patterns across fine time scales. We can also use unique specification and parameters for each mixture weight, allowing us to detect location-specific temporal patterns. These autoregressive priors also create smoothing, or shrinkage, of the estimated spatial density across discrete time periods, which is desirable since the spatial density is typically believed to vary smoothly across time.

The mixture weights, $p_{b,j}$, are subject to non-negativity and sum-to-unity constraints; placing autoregressive priors and manipulating them would require special attention. Instead, we transform them into an unconstrained parametrization via the multinomial logit transformation (also used in [54])

$$\pi_{b,r} = \log\left[\frac{p_{b,r}}{1 - \sum_{j=1}^{K-1} p_{b,j}}\right], \quad r \in \{1, \ldots, K-1\},\ b \in \{1, \ldots, B\}, \tag{2.4}$$

with the following inverse transformation

$$p_{b,j} = \frac{\exp\{\pi_{b,j}\}}{1 + \sum_{r=1}^{K-1} \exp\{\pi^{b,r}\}}, \quad j = 1, \ldots, K-1, \quad \text{and} \quad p_{b,K} = \frac{1}{1 + \sum_{r=1}^{K-1} \exp\{\pi^{b,r}\}}.$$

We then specify autoregressive priors on the transformed weights $\{\pi_{b,r}\}$. For this application, we apply the CAR priors to capture first-order autocorrelation and daily seasonality.

We assume that the de-meaned transformed weights from any time period depend most closely on those from four other time periods: immediately before and after (to represent short-term serial dependence), and exactly one day before and after (to capture daily seasonality). We impose the following priors

$$\begin{aligned}
\pi_{b,r}|\boldsymbol{\pi}_{-b,r} &\sim \mathrm{N}\left(c_r + \rho_r\left[(\pi_{b-1,r} - c_r) + (\pi_{b+1,r} - c_r) + (\pi_{b-d,r} - c_r) + (\pi_{b+d,r} - c_r)\right], \nu_r^2\right), \\
c_r &\sim N(0, 10^4) \\
\rho_r &\sim \mathrm{U}(0, 0.25) \\
\nu_r^2 &\sim U(0, 10^4),
\end{aligned} \tag{2.5}$$

for $r \in \{1, \ldots, K-1\}$ and $b \in \{1, \ldots, B\}$, in which $\boldsymbol{\pi}_{-b,r} = (\pi_{1,r}, \ldots, \pi_{b-1,r}, \pi_{b+1,r}, \ldots, \pi_{B,r})'$, and d is the number of time periods in a day ($d = 12$ in this case). Since every week has the same sequence of spatial densities, we define priors of $\{\pi_{b,r}\}$ circularly in time, such that the last time period is joined with the first time period. In the prior specification of $\{\pi_{b,r}\}$, the CAR parameters ρ_r determine the persistence in the transformed mixture weights over time, while the intercepts c_r determine their mean levels, and the variances v_r^2 determine the conditional variability. These three parameters are component-specific, and therefore location-specific. For any $\rho_r \in (-0.25, 0.25)$, the joint prior distribution of $[\pi_{1,r}, \ldots, \pi_{B,r}]$ is a proper multivariate normal distribution [14]; we take the priors of ρ_r to be $\mathrm{U}(0, 0.25)$ because exploratory data analysis only detected evidence of nonnegative serial dependence. The priors on c_r and v_r are diffuse, reflecting the fact that we have little prior information regarding their values.

Alternative to this circular definition of mixture weights with symmetric dependence on past and future, one can also specify the marginal distribution of $\pi_{1,r}$ and let each $\pi_{b,r}$ depend only on its past. In either setting, we can represent a wide range of complex temporal patterns.

2.2 Computation

We apply Bayesian estimation, largely following [64] and [78] in our choices of prior distributions and hyperparameters. Richardson and Green define a set of independent, weakly informative and hierarchical priors conjugate to univariate Gaussian mixture models [64], which Stephens extends to the multivariate case [78]. We extend to incorporate time-varying mixture weights, and instead of imposing independent Dirichlet priors on $\{p_{b,j}\}$, we impose CAR priors on $\pi_{b,r}$ as in Equation (2.5). For all other parameters, we have for $j \in \{1, \ldots, K\}$ and

$t \in \{1, \ldots, T\}$,

$$\begin{aligned} \boldsymbol{\mu}_j &\sim \text{Normal}\left(\boldsymbol{\xi}, \boldsymbol{\kappa}^{-1}\right) \\ \boldsymbol{\Sigma}_j^{-1} | \boldsymbol{\beta} &\sim \text{Wishart}\left(2\alpha, (2\boldsymbol{\beta})^{-1}\right) \\ \boldsymbol{\beta} &\sim \text{Wishart}\left(2g, (2\boldsymbol{h})^{-1}\right), \end{aligned} \tag{2.6}$$

in which we set $\alpha = 3$, $g = 1$ and

$$\boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}, \qquad \boldsymbol{\kappa} = \begin{bmatrix} \frac{1}{R_1^2} & 0 \\ 0 & \frac{1}{R_2^2} \end{bmatrix}, \qquad \boldsymbol{h} = \begin{bmatrix} \frac{10}{R_1^2} & 0 \\ 0 & \frac{10}{R_2^2} \end{bmatrix},$$

in which ξ_1 and ξ_2 are the medians of all observations in the first and second spatial dimensions, respectively, and R_1 and R_2 are the lengths of the ranges of observations in the first and second spatial dimensions, respectively. The prior on each $\boldsymbol{\mu}_j$ is diffuse, with prior standard deviation in each spatial dimension equal to the length of the range of the observations in that dimension. The inverse covariance matrices $\boldsymbol{\Sigma}_j^{-1}$ are allowed to vary across j, while centering around the common value $E(\boldsymbol{\Sigma}_j^{-1} | \boldsymbol{\beta}) = \alpha \boldsymbol{\beta}^{-1}$. The constant α controls the spread of the priors on $\boldsymbol{\Sigma}_j^{-1}$; this is taken to be 3 as in [78], yielding a diffuse prior for $\boldsymbol{\Sigma}_j^{-1}$. The centering matrix $\boldsymbol{\beta}^{-1}$ is given an even more diffuse prior, since g is taken to be a smaller positive constant. Our choice of $\boldsymbol{h}$ is the same as [78].

We perform estimation via Markov chain Monte Carlo (MCMC) [66, 85]. We augment each observation $\boldsymbol{s}_{t,i}$ with its latent component label $z_{t,i}$ [83], simulating a Markov chain with limiting distribution equal to the joint posterior distribution of $\{z_{t,i}\}$, $\{\boldsymbol{\mu}_j\}$, $\boldsymbol{\beta}$, $\{\boldsymbol{\Sigma}_j\}, \{\pi_{b,r}\}, \{c_r\}, \{\rho_r\}$, and $\{\nu_r\}$. After initializing all parameters by drawing from their respective priors, we update $\{z_{t,i}\}, \{\boldsymbol{\mu}_j\}, \boldsymbol{\beta}$ and $\{\boldsymbol{\Sigma}_j\}$ by their closed-form full conditional distributions (Equation 2.7), and update $\{\pi_{b,r}\}, \{c_r\}, \{\rho_r\}$ and $\{\nu_r\}$ via random-walk Metropolis-Hastings.

$$
\begin{aligned}
P(z_{t,i} = j|\cdot) &\propto p_{b(t),j}\phi(\boldsymbol{s}_{t,i}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \\
[\boldsymbol{\beta}|\cdot] &\sim \text{Wishart}\left(2g + 2K\alpha, \left[2\boldsymbol{h} + 2\sum_j \boldsymbol{\Sigma}_j^{-1}\right]^{-1}\right), \\
[\boldsymbol{\mu}_j|\cdot] &\sim \text{Normal}\left([m_j\boldsymbol{\Sigma}_j^{-1} + \boldsymbol{\kappa}]^{-1}[m_j\boldsymbol{\Sigma}_j^{-1}\bar{\boldsymbol{s}}_j + \boldsymbol{\kappa}\boldsymbol{\xi}], [m_j\boldsymbol{\Sigma}_j^{-1} + \boldsymbol{\kappa}]^{-1}\right), \\
[\boldsymbol{\Sigma}_j^{-1}|\cdot] &\sim \text{Wishart}\left(2\alpha + m_j, \left[2\boldsymbol{\beta} + \sum_{(t,i):z_{t,i}=j} (\boldsymbol{s}_{t,i} - \boldsymbol{\mu}_j)(\boldsymbol{s}_{t,i} - \boldsymbol{\mu}_j)'\right]^{-1}\right),
\end{aligned}
\tag{2.7}
$$

for $t \in \{1, \ldots, T\}$, $i \in \{1, \ldots, n_t\}$ and $j \in \{1, \ldots, K\}$, in which $m_j = \sum_t \sum_{i=1}^{n_t} \mathbb{1}_{\{z_{t,i}=j\}}$ is the number of observations in all periods assigned to component j, and $\bar{\boldsymbol{s}}_j = \frac{1}{m_j}\sum_{\{(t,i):z_{t,i}=j\}} \boldsymbol{s}_{t,i}$ is the mean of these observations.

We could also perform the above estimation in two stages. In the first stage, we would estimate the distributions of all mixture components using the locations of all observations in consideration (disregarding their times). In the second stage, we would estimate the evolution of mixture weights over time, conditional on fixed component distributions. This two-stage estimation approach can be applied to all mixture models we propose. The two-stage method would perhaps provide some computation savings, while the simultaneous estimation of all parameters would probably result in estimates with lower variance; there is likely a trade-off between computational simplicity and statistical efficiency. For simplicity, we only demonstrate simultaneous estimation of both component distributions and mixture weights.

After estimation, we numerically normalize $f_t(\cdot)$ for each t with respect to Toronto's boundary to obtain final density estimates. As a result, we predict outside of Toronto's boundary with probability zero, and the density within the boundary is elevated proportionally for each t. Here we do not impose this boundary during estimation; we consider doing so in Section 2.3.3.

2.3 Additional model refinements

2.3.1 Number of components

We assumed a fixed number of mixture components in our proposed model in Section 2.1; in this section we estimate a variable number of components. As mentioned in Section 1.2.3, allowing the number of components to vary typically improves the mixing (efficiency) of the MCMC computational method, by allowing the Markov chain to escape local modes more quickly. Stephen's birth-and-death MCMC (BDMCMC) is especially effective at this [78] (reviewed in Section 1.2.3).

We adapt BDMCMC to a spatio-temporal setting. We can generalize Stephens' BDMCMC to incorporate time-varying mixture weights in a straightforward way, by maintaining the same number of components across different time periods within each iteration. Since the birth and death process applies in the same way to all time periods, it is easy to show that stationarity holds for this generalized sampling method. Following [78], we assume a truncated Poisson prior on the number of components K, i.e., $P(K) \propto \tau^K/K!,\ K \in \{1, \ldots, K_{max}\}$ for some fixed τ and K_{max}. All other priors and hyperparameters are as specified in (2.5) and (2.6), and the spatial density function at each time is as in Equation (2.3).

BDMCMC algorithm for time-binned data
Start with an initial configuration of K components,

$$y = \left\{ \left(\{p_{b,1}\}_{b=1}^{B}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1 \right), \ldots, \left(\{p_{b,K}\}_{b=1}^{B}, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K \right) \right\},$$

and iterate the following:

1. For $j \in \{1, \ldots, K\}$, calculate death rate for the j^{th} component, $d_j(y)$, accord-

ing to

$$d_j(y) = \lambda_b \frac{L\left(y \backslash (\{p_{b,j}\}_{b=1}^B, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right)}{L(y)} \frac{P(K-1)}{K\,P(K)} \frac{P\left(\{p_{b,1}, \ldots, p_{b,j-1}, p_{b,j+1}, \ldots, p_{b,K}\}_{b=1}^B\right)}{P\left(\{p_{b,1}, \ldots, p_{b,K}\}_{b=1}^B\right)},$$

in which $L(\cdot)$ is joint likelihood function and $P(\cdot)$ represents the priors. When calculating the contribution of an observation to the likelihood function, we use the mixture weights corresponding to that observation's time period.

2. Calculate the total death rate, $d(y) = \sum_{j=1}^K d_j(y)$.
3. Draw time to next jump from an exponential distribution with mean $\lambda_b + d(y)$.
4. Simulate whether the next jump is a birth or a death with Pr(birth) = $\lambda_b/(\lambda_b + d(y))$ and Pr(death) = $d(y)/(\lambda_b + d(y))$.
5. If it is a birth, independently sample the new component's weight from Beta$(1, K)$ and parameters from their respective priors. We add this component to all time periods and adjust all other mixture weights proportionally so that the sum of mixture weights is 1 for each period. If it is a death, simulate which component dies, where the j^{th} component of all time periods dies with probability $d_j(y)/d(y)$. After death, we also adjust all other mixture weights accordingly.
6. Repeat Steps 1 to 5 for a fixed amount of time T_0 (increasing T_0 and increasing the birth rate, λ_b, have the same effect; without loss of generality, we can let $T_0 = 1$).
7. Fix the number of components, and update all other parameters via full-conditionals or Metropolis-Hastings as illustrated in Section 2.2.

Given the large amount of data and the complexity of spatial-temporal methods, imposing a vague prior on the number of components would result in an unfeasibly large number of mixture components, and leads to overfitting. We therefore use strong priors on K to regularize the number of components.

2.3.2 Covariates

There are many spatial and/or temporal covariates that we may want to incorporate to improve modeling accuracy and explanatory power. We can incorporate these covariates in the proposed time-varying Gaussian mixture model (Section 2.1) in the following ways.

If temporal covariates $\boldsymbol{x}_t$ are available (e.g., temperature and precipitation), they may be incorporated into the hierarchical CAR priors for the (transformed) mixture weights. Previously, in Equation (2.5), we assume *a priori* that the centered weights $\pi_{b,r} - c_r$ follows a conditional autoregressive structure componentwise. This may be modified such that covariate-adjusted weights $\pi_{t,r} - c_{b(t),r} - \boldsymbol{a}_r' \boldsymbol{x}_t$ follow the same conditional autoregressive structure, in which $\boldsymbol{x}_t$ is a vector of temporal covariate values for time t, and $\boldsymbol{a}_r$ is a vector of location-specific coefficients for the rth component. One advantage of this specification is that it can differentiate the impacts of covariates to ambulance demand at different component locations in space. In the presence of no covariates, the $\{c_{b(t),r}\}$ are simply the mean values of $\pi_{t,r}$ across the weekly temporal blocks $b(t)$, utilized to capture the intra-week pattern. Equation (2.5) then becomes

$$\begin{aligned}
\pi_{t,r} | \boldsymbol{\pi}_{-t,r} \sim \mathrm{N}\big(& c_{b(t),r} + \boldsymbol{a}_r' \boldsymbol{x}_t + \rho_r [(\pi_{t-1,r} - c_{b(t-1),r} - \boldsymbol{a}_r' \boldsymbol{x}_{t-1}) + (\pi_{t+1,r} - c_{b(t+1),r} - \boldsymbol{a}_r' \boldsymbol{x}_{t+1}) \\
& + (\pi_{t-d,r} - c_{b(t-d),r} - \boldsymbol{a}_r' \boldsymbol{x}_{t-d}) + (\pi_{t+d,r} - c_{b(t+d),r} - \boldsymbol{a}_r' \boldsymbol{x}_{t+d})], \ \nu_r^2\big), \\
c_{b,r} \sim\ & N(0, 10^4), \\
\rho_r \sim\ & U(0, 0.25), \\
\nu_r^2 \sim\ & U(0, 10^4).
\end{aligned}$$

To include a spatial or spatio-temporal covariate (e.g., population density), we may proceed by adding it to the current model as an additional mixture "component." Consider a modified Equation (2.2)

$$f_t(\boldsymbol{s}; \{p_{t,j}\}, \{\boldsymbol{\mu}_j\}, \{\boldsymbol{\Sigma}_j\}) = p_{t,0} g_t(\boldsymbol{s}) + (1 - p_{t,0}) \sum_{j=1}^{K} p_{t,j} \, \phi(\boldsymbol{s}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),$$

in which $g_t(\boldsymbol{s})$ is a (possibly time-varying) spatial density, and $p_{t,0} \in [0,1]$ is a time-varying probability that ambulance demand arises directly from the covariate density. As such, demand arises from component j of the Gaussian mixture model with the time-varying probability of $(1 - p_{t,0})p_{t,j}$, for $j \in \{1, \ldots, K\}$. In this framework, we can still incorporate weekly seasonality (Equation (2.3)) and CAR priors (Equation (2.5)) on the mixture weights. The Bayesian estimation extends naturally, by way of an additional data augmentation to include the component label for $g_t(\boldsymbol{s})$. If the covariate density is the primary factor in realized demand, we expect $p_{t,0}$ to be large. In general, we can include multiple such spatial covariates, as long as we make sure all mixture weights are well-defined.

We consider two temporal covariates: mean daily temperature and total daily precipitation in Toronto. We found a statistically significant negative relationship between precipitation level (snow only; there was no rain in Feb 2007) and ambulance demand at three downtown components and one at a central traffic route. This is perhaps not surprising because snow was likely removed from these locations more promptly; as a result there was a relatively lower demand for ambulances (responding to traffic related events) relative to other locations within the city. Temperature was not found to be statistically significant at the 5% level. Including these two temporal covariates do not drastically improve the predictive accuracy. We therefore did not include these covariates ultimately.

Unfortunately, there are very limited spatial or spatio-temporal covariate data available. We were only able to find scalar population density estimates for the entire city for the entire census year. What we would like is a spatial map of population density, or more ideally, a spatio-temporal population density that varies in fine time scales (e.g., hourly as population shifts). This type of data is typically not collected, especially with privacy concerns. In fact, we may have provided a possible proxy measure of it. Indeed, it would be very interesting to see what other factors affect ambulance demand, as well as when and where they play a role. Perhaps some early regression models on EMS demand can

shed some light. Studies have found that demand is higher at areas with lower socioeconomic status and higher concentration of elderly people [3, 74, 48, 18]. However, such demographic data is also not available at a fine time scale as population shifts. This lack of data is a widely acknowledged in the literature of EMS demand prediction.

2.3.3 Boundary

In Section 2.1 we model Toronto's ambulance demand on a continuous spatial domain within $\mathbb{R}^2$, and we post-process the the density estimates by normalizing with respect to Toronto's boundary after the estimation. In this section, we consider incorporating the boundary of Toronto in the density estimation. Let this boundary be $\overline{\mathcal{S}}$ and define it as the convex hull of ambulance demand from years 2007 and 2008.

In the context of Toronto EMS, assuming a boundary is natural, given Lake Ontario to the south of the city and the suburban areas around Toronto, for which data is not available. In our application, we observe a relatively high density of observations along the lake edge, which is difficult for a small number of mixture components to capture well. If we incorporate Toronto's boundary in estimation, the truncation of spatial densities at the boundary would encourage mixture components that are close to the lake to move towards/beyond the lake or take on higher weights in order to better describe the high density of observations.

We can normalize the density estimates in Equation (2.3) according to

$$\overline{f_b}(\boldsymbol{s};\boldsymbol{\theta}) = \frac{f_b(\boldsymbol{s};\boldsymbol{\theta})\mathbb{1}_{\{\boldsymbol{s}\in\overline{\mathcal{S}}\}}}{\int_{\overline{\mathcal{S}}} f_b(\boldsymbol{s};\boldsymbol{\theta})\,d\boldsymbol{s}}, \tag{2.8}$$

in which $\boldsymbol{\theta}$ is a vectorization of all the parameters. We numerically approximate the denominator of (2.8) as follows. Grid the spatial domain $\overline{\mathcal{S}}$ into E fine, equal-sized spatial cells each with area A_E, and label centers of these E cells as

$\{\boldsymbol{u}_e, e = 1, \ldots, E\}$. Using $\{\boldsymbol{u}_e\}$ as evaluation points, we approximate the normalizing integral using its Riemann sum $\sum_{e=1}^{E} A_E \left[f_b(\boldsymbol{u}_e; \boldsymbol{\theta})\right]$.

Computationally, we lose the closed-form full-conditionals for $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ and a numerical integration is needed to compute likelihood in every Metropolis-Hastings proposal, for every parameter. We have investigated this extension fully, but we found the computation to be prohibitively expensive on samples this large, and only a minor out-of-sample improvement in some test cases. As a result, we did not impose this boundary in estimation in our proposed model. we propose an alternative method in Chapter 4 that incorporates boundary and other spatial features elegantly.

2.4 Predicting Toronto ambulance demand

We fit the full Gaussian mixture model with seasonality constraints and autoregressive priors (Section 2.1) on the Toronto EMS data from February 2007. First, we use a fixed number of 15 components. We found 15 components to be large enough to capture a wide range of residential, business and transportation regions in Toronto, yet small enough for computational ease given the large number of observations. We then fit the model again with a variable number of components (Section 2.3.1). Given the large amount of data and the complexity of spatial-temporal methods, imposing a vague prior on the number of components would result in an unfeasibly large number of mixture components, and leads to overfitting. We therefore set the *a priori* maximum number of components $K_{max} = 50$ and chose two small values for the prior mean of the number of components τ. These prior choices lead to posterior average numbers of components of 19 or 24 (with posterior standard deviations of 3.1 and 4.6, respectively).

Each MCMC algorithm is run for 50,000 iterations, with the first 25,000 iterations discarded as burn-in. We compute the effective sample sizes and Gelman-Rubin diagnostics [31] of the minimum and maximum of component means and variances along each spatial dimension. In a typical simulation, the mean pa-

rameters have effective sample size averaged 2,606 and Gelman-Rubin below 1.05, and for covariance parameters, 6,065 and 1.09, respectively. This suggests burn-in and mixing may be sufficient. We focus on the minimum and maximum of these parameters instead of relying on component labels because any mixture models may encounter the label switching problem, in which the labeling of component parameters can permute while yielding the same posterior distribution. However, this label switching problem does not affect estimation of ambulance demand in time and space, because we are interested in the entire posterior distribution, instead of inferences on individual mixture parameters. In Section 2.5.2 we also report the estimated MCMC standard errors of the performance measures [29, 15]; they are small enough to provide accuracy to 3 significant digits, further suggesting the run length may be satisfactory.

Using a personal computer, the computation times for the proposed model with 15 components is about 4 seconds per iteration, compared to 7 and 8 seconds using variable numbers of components averaged 19 and 24, respectively. In practice, estimation using the proposed model only needs to be performed infrequently (at most once a month in this application); density prediction of any future time period can then be immediately calculated as the corresponding density using the most recent parameter estimation results.

Figures 2.1 and 2.2 present results from fitting using 15 components. Figure 2.1 shows all 15 Gaussian component ellipses at the 90% level, using the parameter values from the $50,000$th iteration of the Markov chain. Each component ellipse is shaded by the posterior mean of ρ_r for that component, except for the 15th component because $r \in \{1, \ldots, 14\}$. Components at the denser greater downtown and coastal regions of Toronto have the highest estimates of ρ_r; these regions exhibit the strongest low-order serial dependence and daily seasonality. This is in line with exploratory study in Section 1.3. The proposed model is able to easily differentiate temporal patterns and dynamics at different locations.

Figure 2.2 shows the log predictive densities at Wednesday around midday and midnight, as computed by the proposed mixture model with 15 compo-

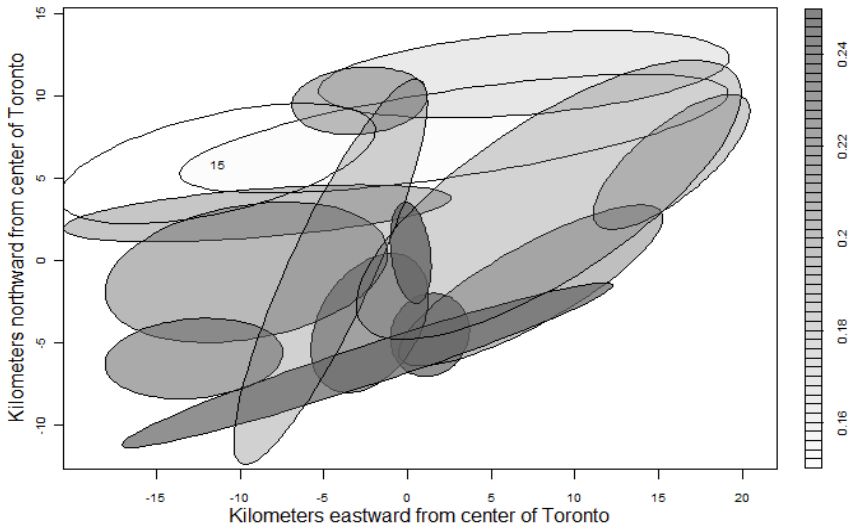

Figure 2.1: Using 15 components: Gaussian component ellipses at the 90% level. Each component (except the 15th) is shaded with the posterior mean of ρ_r for that component. The greater downtown and coastal regions exhibit stronger low-order serial dependence and daily seasonality.

nents and averaged across the last 25,000 Monte Carlo samples. Note that the demand is concentrated at the heart of downtown during working hours in the day, but is more dispersed throughout Toronto during the night.

Figure 2.3 shows the log predictive densities using variable numbers of components around Wednesday midnight; these spatial densities are similar to that using 15 components (shown in Figure 2.2 (b)).

2.5 Model performance and validation

We evaluate the performance and validity of the proposed models in several ways. For performance, we attempt to predict ambulance demand densities on two sets of test data (March 2007 and February 2008). To do this using mixture models, we train the models on data from February 2007, and use the resulting

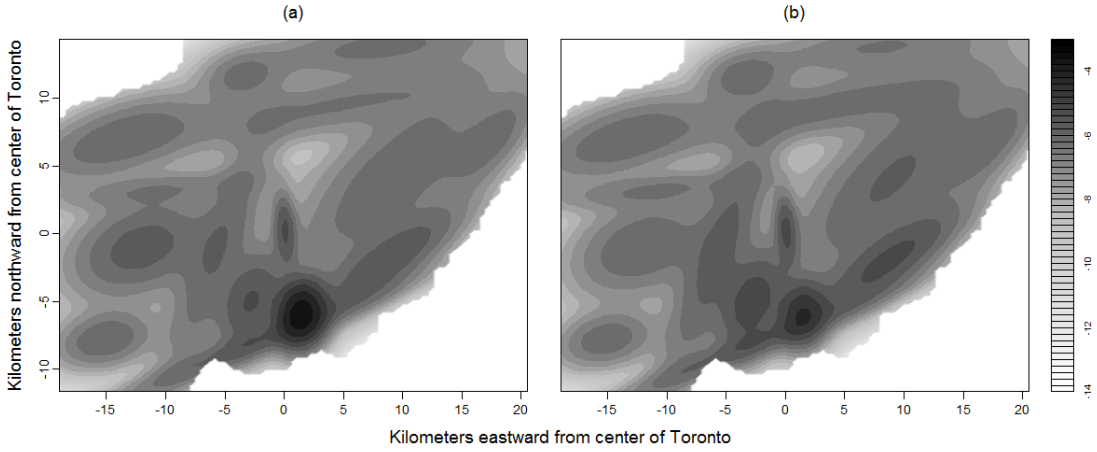

Figure 2.2: Using 15 components: (a) posterior log spatial density for Wednesday 2-4pm (demand concentrated at downtown during the day); (b) posterior log spatial density for Wednesday 2-4am (demand more spread out during the night).

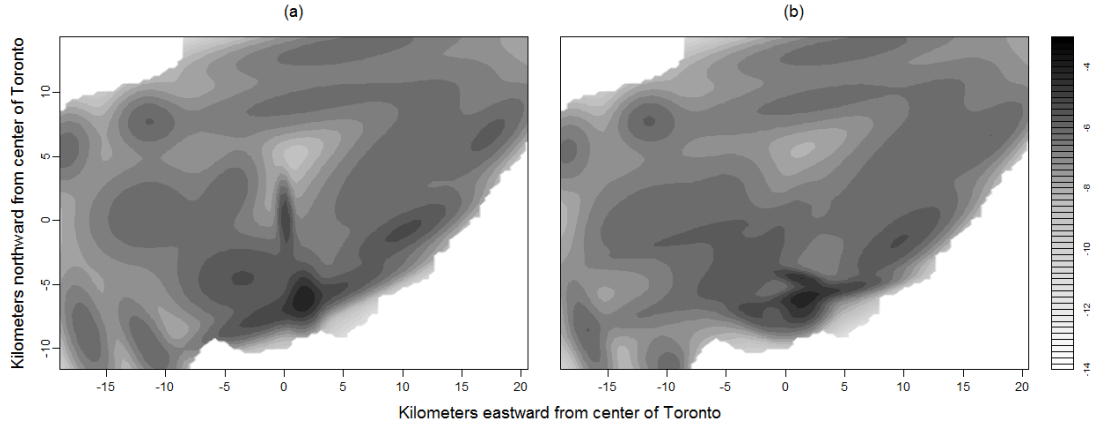

Figure 2.3: Using variable number of components: (a) posterior log spatial density for Wednesday 2-4am (night) using an average of 19 components; (b) that using an average of 24 components.

density estimates to predict for both sets of test data. We introduce in Section 2.5.1 two methods for comparisons. We compare the statistical predictive accuracies for all methods in Section 2.5.2. In Section 2.5.3 we then put these predictive accuracies in the context of EMS operations. We verify the validity of the method in Section 2.5.4.

2.5.1 Comparison methods

We compare the proposed mixture models to a current industry practice, and to a proposed extension of the industry practice that uses kernel density estimation (KDE). As mentioned in Section 1.1.3, Toronto EMS employed an averaging method based on a discretized spatial and temporal domains. The demand forecast at a spatial cell in a particular time period is the average of four corresponding realized demand counts for the past four years (from the same location, week of the year, day of the week, and hour of the day). Each spatial cell is 1 km by 1 km. A similar practice described in [70], the MEDIC method, uses the average of up to twenty corresponding historical demands in the preceding four weeks, for the past five years. These industry practices capture, to various extents, yearly and weekly seasonalities present in EMS demand.

We implement the MEDIC method as far as we have historic data available. Since we focus on predicting the demand density, we normalize demand volumes at any place by the total demand for the time period. For any 2-hour period in March 2007, we average the corresponding demand densities in the preceding four weeks. For any 2-hour period in February 2008, we average the corresponding demand densities in the preceding four weeks in 2008 and those same four weeks in 2007. For example, to forecast the demand density for 8 - 10 am on the second Monday of February 2008, we average the demand densities at 8 - 10 am in the first Monday of February 2008, the last three Mondays of January 2008, the first Monday of February 2007 and the last three Mondays of January 2007. Note that this means the MEDIC method is trained on at least as much data, which is at least as recent as that used in the mixture models. We adopt the same 1 km by 1 km spatial discretization used by Toronto EMS.

Since the proposed method is continuous in space, we also propose to extend the MEDIC method to predict continuous demand densities as a second comparison method. The demand density for each 2-hour period is taken to be the KDE for all observations from that period. Here we use a bivariate normal kernel function, and bandwidths chosen by cross-validation using the predictive

accuracy measure in Section 2.5.2. We predict demand densities for March 2007 and February 2008 by averaging past demand densities using the MEDIC rule described above. To ensure fair comparisons, we also numerically normalize the predictive densities produced by the two comparison methods with respect to Toronto's boundary.

Figure 2.4 shows the log predictive density using these two competing methods for February 6, 2008 (Wednesday) 2 - 4 am. These two densities are comparable with Figures 2.2 (b) and 2.3, which are the log predictive densities for the same time period estimated from the proposed mixture models. Compared to the proposed model, both the MEDIC and MEDIC-KDE produce less smooth predictions compared to the proposed mixture models.

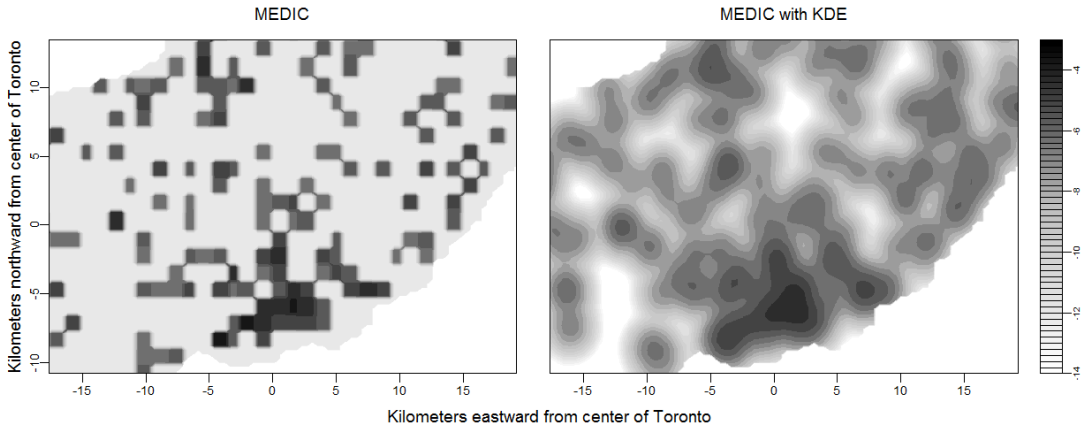

Figure 2.4: Log predictive densities using two current industry estimation methods for 2-4am (night) on February 6, 2008 (Wednesday). Figure 2.2(b) and Figure 2.3 show the log predictive densities for the same period using mixture models. Compared to mixture models, estimates from the MEDIC and MEDIC-KDE are less smooth.

2.5.2 Statistical predictive accuracy

To measure the predictive accuracy of density estimates obtained from the proposed mixture models, MEDIC, and the proposed MEDIC-KDE, we use the av-

erage logarithmic score (ALS). First proposed by [34], this performance measure is advocated for being a strictly proper scoring rule and its connections with Bayes factor and Bayes information criterion [32, 22, 11]. We define

$$\mathrm{ALS}\left(\{\tilde{\boldsymbol{s}}_{t,i}\}\right) = \frac{1}{\sum_{t=1}^{T} n_t} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \log \hat{f}_t(\tilde{\boldsymbol{s}}_{t,i}), \tag{2.9}$$

in which $\hat{f}_t(\cdot)$ is the density estimate for time t obtained using various methods, and $\tilde{\boldsymbol{s}}_{t,i}$ denotes observations from the test data (March 2007 or February 2008). For the proposed mixture models, we use the Monte Carlo estimate of Equation (2.9)

$$\mathrm{ALS}_{mix}\left(\{\tilde{\boldsymbol{s}}_{t,i}\}\right) = \frac{1}{M} \sum_{m=1}^{M} \left[\frac{1}{\sum_{t=1}^{T} n_t} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \log \hat{f}_t(\tilde{\boldsymbol{s}}_{t,i} | \boldsymbol{\theta}^{(m)}) \right],$$

in which $\boldsymbol{\theta}^{(m)}$ represents the mth-iteration posterior parameter estimates generated from the training data, for $m \in \{1, \ldots, M\}$ and some large M.

The predictive accuracies of various methods for two test data sets (March 2007 and February 2008) are shown in Table 2.1. The predictive accuracies for Gaussian mixture models are presented with their 95% consistent, non-overlapping batch means confidence intervals [see 43], which reflect the accuracy of the MCMC estimates. Here, a less negative predictive accuracy indicates better performance. The proposed mixture models outperform the two current industry methods. Allowing for a variable number of components improves the predictive accuracy slightly, but the rate of improvement diminishes as the average number of components grows. Given that the computational expense almost doubles to obtain these modest improvements, we conclude that using a fixed number of 15 components is largely sufficient in this application.

2.5.3 Operational predictive accuracy

In this section, we quantify the advantage of the proposed model over the industry practice. We show that the proposed model gives much more accurate

Estimation method		ALS for Mar 07	ALS for Feb 08
Gaussian Mixture	15 components (2.1)	-6.1378 ± 0.0004	-6.1491 ± 0.0005
	Variable # of comp (2.3.1):		
	average 19 comp	-6.080 ± 0.002	-6.128 ± 0.002
	average 24 comp	-6.072 ± 0.003	-6.122 ± 0.004
Competing Methods	MEDIC	-8.31	-7.62
	MEDIC-KDE	-6.87	-6.56

Table 2.1: Predictive accuracies of proposed Gaussian mixture models and competing methods on test data of March 2007 and February 2008. The predictive accuracies for mixture models are presented with their 95% batch means confidence intervals.

forecasts of the industry's operational performance measure. The standard EMS operational performance measure is the fraction of events with response times below various thresholds (e.g., 60% responded within 4 minutes). Obtaining an accurate forecast of this performance is of paramount importance because many aspects of the industry's strategic management aim to optimize this performance. Accuracy in estimating this performance depends crucially on the accuracy of spatio-temporal demand density estimates.

For each of the three methods of interest, we have a set of 2-hour demand density estimates for March 2007 and February 2008. Using density estimates generated by method $\mathscr{M}$ for time period t, we predict the operational performance by computing the proportions of demand, $\mathscr{P}_{\mathscr{M},t}(r)$, reachable within response time threshold r from any of the 44 ambulance bases in Toronto (see Figure 2.5 (a)). To do so, we first discretize Toronto into a fine spatial grid and outline the regions that can be covered within any response time threshold. We then numerically integrate within these regions the demand density estimates from $\mathscr{M}$, for each t and r, to obtain $\mathscr{P}_{\mathscr{M},t}(r)$. We also compute the realized performances using the test data, $\mathscr{P}_{test,t}(r)$. For simplicity, we assume ambulances always travel at the median speed of Toronto EMS trips, 46.44 km / hour. We also use the L_1 (Manhattan) distance between any base and any location. We

consider response time thresholds ranging from 60 seconds to 300 seconds at 10-second intervals.

We compute the average absolute error in predicting operational performance made by each method under various response time thresholds, as compared to the truth. We define

$$\text{Error}(\mathscr{M}, r) = \frac{1}{T}\sum_{t=1}^{T} |\mathscr{P}_{\mathscr{M},t}(r) - \mathscr{P}_{test,t}(r)|.$$

In Figure 2.5 (b) and (c), we show Error($\mathscr{M}, r$) against r for each method $\mathscr{M}$ (mixture model with 15 components, MEDIC, and MEDIC-KDE), using test data from March 2007 and February 2008, respectively. The 95% point-wise confidence bands for Error($\mathscr{M}, r$) are shown in gray; these bands indicate interval estimates for the average absolute errors for each $\mathscr{M}$ and r given a series of errors. We find that the proposed method predicts the operational performance much more accurately, given the same set of operational assumptions about base locations, speed and distance. It reduces error by as much as two-thirds compared to the MEDIC method, despite sometimes using less recent training data. We expect similar orderings of the three methods under different sets of operational strategies.

2.5.4 Model validation

We assess the goodness-of-fit of the proposed models and the validity of the NHPP assumption. We use the model checking approach in [82], where each marginal of the point event data is transformed into quantities that are assumed to be uniformly distributed, and compared to the true uniform distribution graphically. In particular, we have assumed that the point process follows a NHPP with time-varying intensity $\lambda_t(\boldsymbol{s}) = \delta_t f_t(\boldsymbol{s})$. We have posterior estimates of $\{f_t(\boldsymbol{s})\}$ from the proposed mixture models. We estimate δ_t in two ways. First, we assume that $\delta_t = n_t$, in which n_t is the actual, realized demand counts in

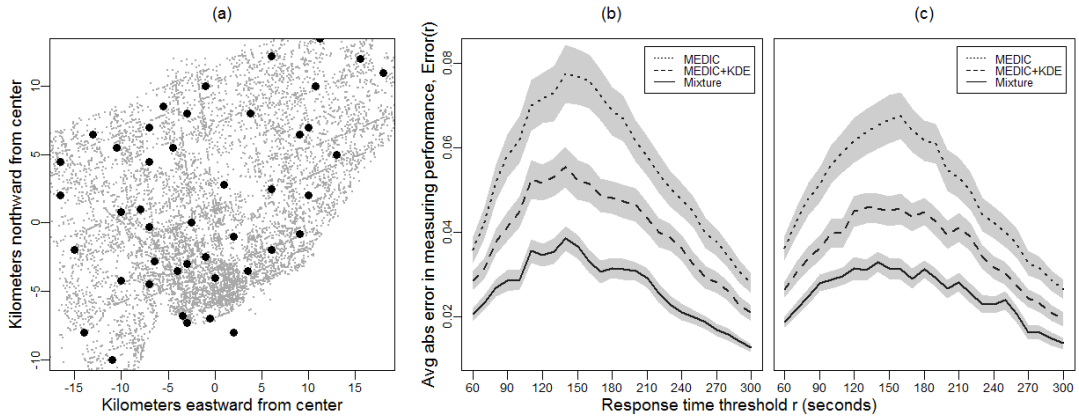

Figure 2.5: (a) all 44 ambulance bases in Toronto; (b) and (c) average absolute error in measuring operational performance made by the proposed mixture model (15 components), MEDIC, and MEDIC-KDE, using test data from March 2007 and February 2008, respectively (with 95% point-wise confidence intervals in gray). The proposed mixture model outperforms the competing methods.

the t-th period. In this case, we attempt to validate the proposed model only with respect to the spatial densities with no uncertainty from δ_t. Second, we use method proposed in Matteson et al to estimate [50].

Point locations along the first and the second spatial dimension thus follow one-dimensional NHPP with marginalized intensities of $\lambda_t(\cdot)$, denoted as $\lambda_{1,t}(\cdot)$ and $\lambda_{2,t}(\cdot)$, respectively. We compute the corresponding cumulative intensities $\Lambda_{1,t}(\cdot)$ and $\Lambda_{2,t}(\cdot)$ and sort the observations for each time period into ordered marginals $\{\bar{s}_{j,1}, \ldots, \bar{s}_{j,n_t}\}$ for each dimension $j \in \{1,2\}$. If the assumptions are valid and the models have perfect goodness-of-fit, then $\{\Lambda_{j,t}(\bar{s}_{j,i}) : i = 1, \ldots, n_t\}$ for each t and $j \in \{1,2\}$ follows a homogeneous Poisson process with unit rate, and $u_{i,j,t} = 1 - \exp\{-(\Lambda_{j,t}(\bar{s}_{j,i}) - \Lambda_{j,t}(\bar{s}_{j,i-1})\}$ for $i \in \{1, \ldots n_t\}$, $j \in \{1,2\}$ and $t \in \{1, \ldots, T\}$ are i.i.d uniform random variables on (0,1). We compare the $u_{i,j,t}$ samples obtained from the models with the uniform distribution via quantile-quantile (Q-Q) plots. We have a set of $\{u_{i,j,t}\}$ for each set of posterior parameter estimates. In Figures 2.6 and 2.7 we show the mean Q-Q line using actual, realized values for δ_t and estimated δ_t from Matteson et al [50], respectively. We include the 95%

point-wise intervals reflecting the uncertainty in MCMC sampling. All plots indicate high goodness-of-fit, whether we are using a fixed or a variable number of components and whether we are using actual or estimated δ_t.

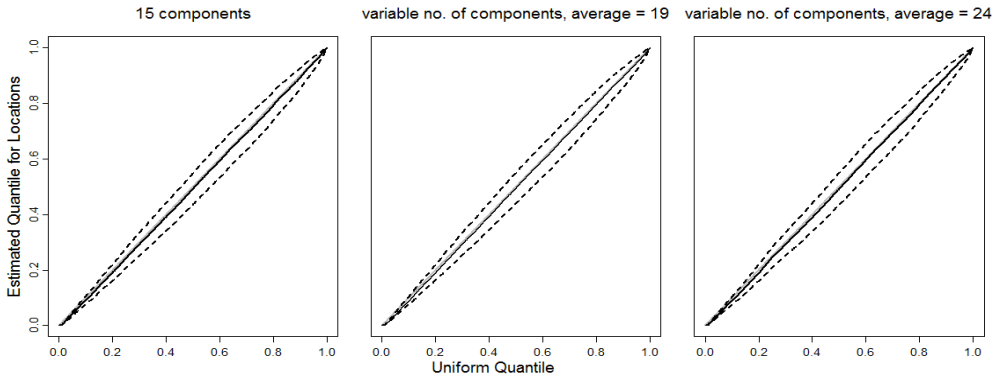

Figure 2.6: Posterior Q - Q plots (solid) and 95% posterior intervals (dash) for the proposed mixture models using actual, realized demand aggregate counts for δ_t. All three plots show that the models are adequate and appropriate.

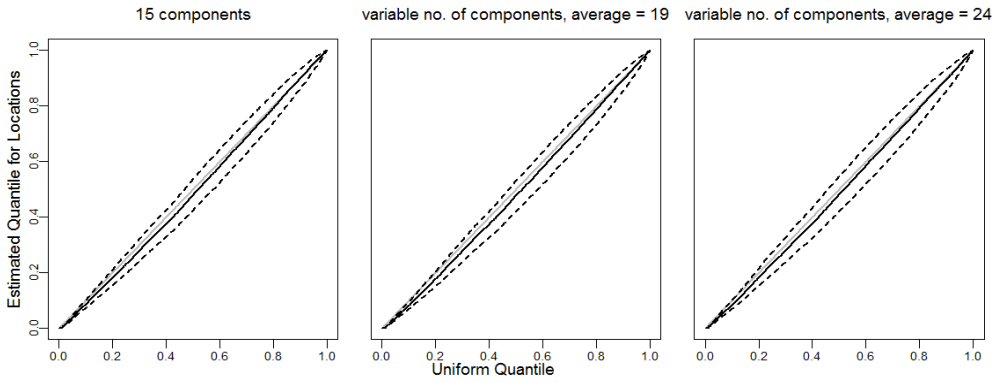

Figure 2.7: Posterior Q - Q plots (solid) and 95% posterior intervals (dash) for the proposed mixture models using methods of Matteson et al [50] to estimate δ_t. All three plots indicate that our models fit the data well.

We repeat the model checking procedure on data at downtown and data along the boundary near the lake and the proposed mixture model with a fixed number of 15 components. We show the result for downtown in Figure (2.8),

and that the lake edge in Figure (2.9). We observe that the proposed mixture model has reasonably high goodness-of-fit in both regions, although not as perfect as goodness-of-fit for Toronto as a whole. Similar results are obtained using variable number of components. This shows that it may be advantageous to impose the boundary of Toronto during estimation. As we discussed in Section 2.3.3, we did not pursue this due to high computational expense.

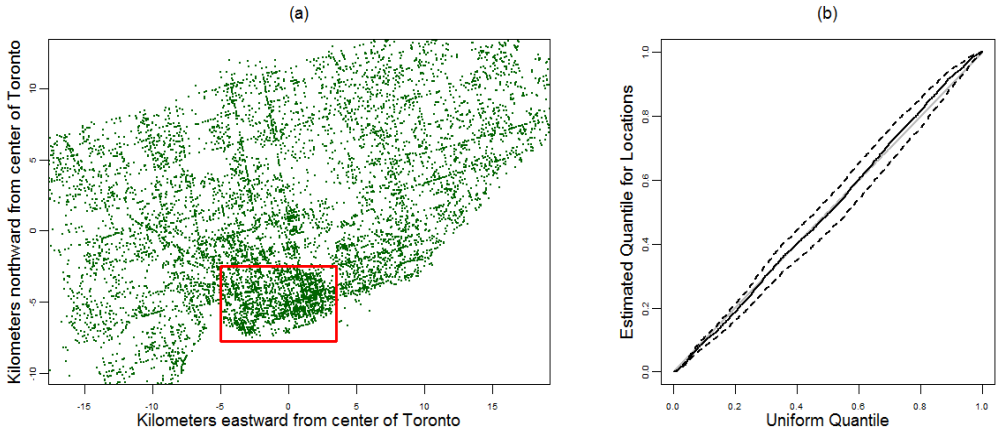

Figure 2.8: Model Checking at downtown: (a) all training data, with downtown outlined in a rectangle; (b) Q-Q plot (solid) and 95% posterior interval (dash) using data from the downtown rectangle and the proposed 15-component mixture model.

2.6 Discussion

We predict spatio-temporal ambulance demand in fine resolutions in time and space by extending Gaussian mixture models. We jointly estimate mixture component distributions over time to promote efficient learning of spatial structures even though data is sparse within each time period. We express a diverse set of location-specific seasonalities and serial dependence typical in the spatial densities of ambulance demand by re-parameterizing the mixture weights and applying conditionally autoregressive priors. We additionally illustrate how to estimate the number of components, incorporate covariate information and im-

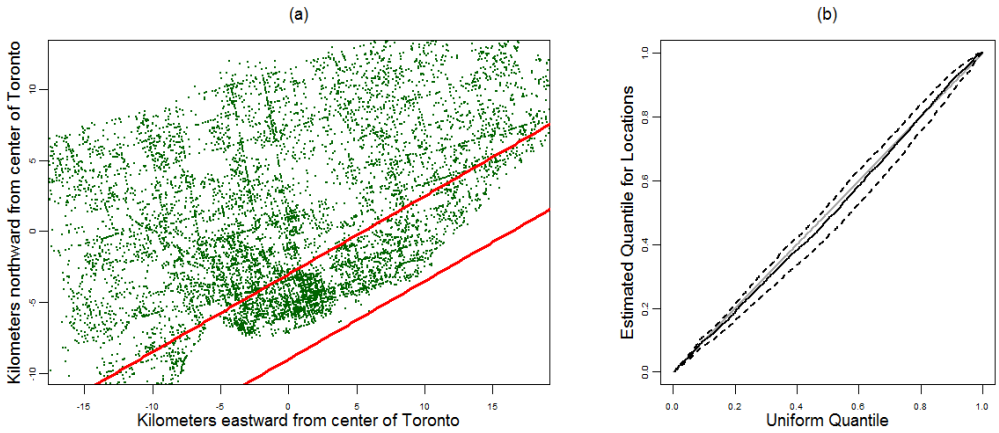

Figure 2.9: Model Checking along the boundary with lake: (a) all training data, with the boundary region enclosed by two lines; (b) Q-Q plot (solid) and 95% posterior interval (dash) using data between the two lines and the proposed 15-component mixture model.

pose spatial boundary.

Practically, our method outperforms the current EMS industry practice, both statistically and operationally. Methodologically, we have developed a set of easily generalizable tools to analyze a wide range of spatio-temporal point process applications. Our method is parsimonious, straightforward to implement, and computationally feasible for large-scale datasets.

CHAPTER 3

SPATIO-TEMPORAL WEIGHTED KERNEL DENSITY ESTIMATION

We propose a fast and accurate method for predicting spatio-temporal ambulance demand that is practical and scalable in industrial settings. We follow Chapter 2 in predicting in discrete time and continuous space. We propose a novel specification of spatio-temporal kernel density estimation (stKDE). First, we learn parametrically the temporal and spatial characteristics of the demand. Each historical observation is annotated with a weight based on what we have learned. This spatio-temporal weight function scales how helpful different historical observations are to a given predictive task. Then we construct a spatial kernel density estimator weighted by the informativeness weight function, and use the resulting kernel density estimates as predictions. In this way, we efficiently emphasize the historical data most important to prediction and, as far as possible, exploit the spatial and temporal characteristics in the data.

The proposed stKDE method have three main advantages

1. accessibility: stKDE is fully automated and robust. It is easy to interpret and use by non-specialized personnel, while other approaches such as artificial neural network (ANN [70]), or time-varying Gaussian mixture models (GMM, Chapter 2) may need special expertise (e.g., tuning, MCMC diagnostics).
2. efficiency: stKDE has low computational complexity. It is faster than GMM; inferring latent component label in GMM is costly. It can afford more frequent parameter estimation updates and online predictions.
3. accuracy: stKDE gives more accurate predictions than current industry practice with similar computational expense. It also outperforms naive KDE methods (and ANN via [70]); it is at least as accurate as GMM.

Material from this chapter is accepted in May 2015 by ACM SIGKDD International Conference on Knowledge Discovery and Data Mining in an article

titled "Predicting Ambulance Demand: A Spatio-Temporal Kernel Approach", authored by Zhengyi Zhou and David S. Matteson [95].

We propose the stKDE model in Section 3.1 and discuss computational methods in Section 3.2. We show the empirical results on Toronto ambulance demand in Section 3.3, and conclude in Section 3.4.

3.1 Model

We model Toronto's ambulance demand on a continuous spatial domain $\mathcal{S} \subseteq \mathbb{R}^2$ and a discretized temporal domain of one-hour intervals $\mathcal{T} = \{1, 2, \ldots\}$. Let $\mathbf{s}_{t,i}$ be the location of the i-th ambulance demand arising from the t-th time period, for $i \in \{1, \ldots, n_t\}$, where n_t is the total number of ambulances demanded in the t-th period. Similar to Chapter 2, we model $\{\mathbf{s}_{t,i} : i = 1, \ldots n_t\}$ for each time period t independently follow an NHPP over $\mathcal{S}$, with positive intensity function λ_t. As in Equation 2.1, the intensity function $\lambda_t(\mathbf{s})$ can be decomposed as $\delta_t f_t(\mathbf{s})$, for $\mathbf{s} \in \mathcal{S}$, where δ_t is the aggregate demand intensity over the spatial domain, and $f_t(\cdot)$ is the continuous spatial *density* of the demand at time t. Like Chapter 2, we also focus on predicting the spatio-temporal demand density $\{f_t\}$, which has received far less attention in the literature.

3.1.1 Spatio-temporal kernel density estimation

Suppose we observe and utilize historical ambulance demand from a set of past time periods $\mathcal{T}_{obs}$, and we would like to forecast demand for a future time period u. We propose to predict f_u using a spatio-temporal weighted kernel density estimator. We place a bivariate spatial kernel K at the location of each past observation in $\mathcal{T}_{obs}$, and weight each kernel by the helpfulness of the corresponding

observation in predicting for the uth time period. Specifically, we have for $\mathbf{s} \in \mathcal{S}$

$$f_u(\mathbf{s}) = \frac{1}{\sum_{t \in \mathcal{T}_{obs}} w(\mathbf{s}_{t,i}, u)} \sum_{t \in \mathcal{T}_{obs}} w(\mathbf{s}_{t,i}, u)\, K_{\boldsymbol{H}}(\mathbf{s} - \mathbf{s}_{t,i}). \tag{3.1}$$

Here, $w(\mathbf{s}_{t,i}, u)$ is the helpfulness weight function multiplied with the spatial kernel of the past observation $\mathbf{s}_{t,i}$. This weight function is defined in detail in Section 3.1.2. $K_{\boldsymbol{H}}$ is the chosen bivariate spatial kernel with bandwidth $\boldsymbol{H}$, i.e.,

$$K_{\boldsymbol{H}}(\mathbf{s} - \mathbf{s}_{t,i}) = \frac{1}{|\boldsymbol{H}|} K\left(\boldsymbol{H}^{-1/2}(\mathbf{s} - \mathbf{s}_{t,i})\right)$$

3.1.2 Weight function

The weight function w aims to capture the utility of a past observation in predicting demand at a future period. Specifically, we would like to incorporate in w the spatial and temporal dependencies in the demand. Domain knowledge on EMS demand densities focuses our attention on weekly and daily seasonalities and short-term serial dependence of a few hours, which have varying strengths in different neighborhoods.

We can therefore discretize the spatial domain into C large spatial cells, representing a rough division into neighborhoods. We assume that temporal dependencies within each cell remain constant in space. Let w_c denote the weight function local to each discretized cell $c \in \{1, \ldots, C\}$. Within this cell, we further assume that the informativeness of a past observation from time t in predicting for future time u only depends on how far back t is from u. We use the weight function to measure how positively correlated two demand densities $(u - t)$ periods apart are in each cell. We model the weight function as

$$w_c(u - t) = \rho_{1,c}^{u-t} + \rho_{2,c}^{u-t} \rho_{3,c}^{\sin^2\left(\frac{\pi(u-t)}{T_1}\right)} \rho_{4,c}^{\sin^2\left(\frac{\pi(u-t)}{T_2}\right)}, \tag{3.2}$$

for $c \in \{1, \ldots, C\}$. We combine all w_c to have

$$w(\mathbf{s}_{t,i}, u) = \sum_{c=1}^{C} w_c(u-t)\, \mathbb{1}_{\{\mathbf{s}_{t,i} \in \text{cell}\, c\}}. \tag{3.3}$$

Here, $\{\rho_{1,c}\}, \{\rho_{2,c}\}, \{\rho_{3,c}\}$ and $\{\rho_{4,c}\}$ for $c \in \{1, \ldots, C\}$ are all constrained to take values in $[0, 1]$. We use a separate ρ parameter to capture each typical EMS patterns for easy interpretation and comparisons across locations (e.g., downtown vs suburbs) and times (e.g., winter vs summer). The term $\rho_{1,c}^{u-t}$ describes any potential short-term serial dependence. Its parametric form is the same as a stationary first-order autoregressive model, AR(1), and is also equivalent to the squared exponential function often used in Gaussian processes [61]. The term with $\rho_{3,c}$ describes any potential daily seasonality with $T_1 = 24$, whereas the term with $\rho_{4,c}$ describes any potential weekly seasonality with $T_2 = 24 \times 7 = 168$. The parametric form of these two terms corresponds to the periodic function used in Gaussian processes [61]. These two seasonality terms are multiplied, and further discounted by a squared exponential function, $\rho_{2,c}^{u-t}$. Finally, we sum the short-term dependency effect and the seasonality effects. The different ρ terms are combined in similar to the typical approach to combining covariance functions in Gaussian processes. . There may be other weight functions that work similarly; we draw inspirations from Gaussian processes because these functions are well-studied and have some nice properties (e.g., infinite differentiability). This parametrization of the weight function is easy to interpret and visualize, and flexible to experiment with, even for non-experts.

The weight function is bounded between 0 and 2. We avoid negative weights to avoid negative kernels in the kernel density estimator, which complicates bandwidth selection, results in negative density estimates that need to be floored at zero, and produces discontinuities in the derivatives of the estimates [69]. The magnitudes of the weights are nominal, as long as they are comparable across all C regions, since they are normalized in Equation (3.1).

Equations (3.1), (3.2) and (3.3) together form the model.

3.2 Computation

3.2.1 Parameter estimation

We must select or estimate the kernel function K and bivariate bandwidth $\boldsymbol{H}$ in Equation (3.1), as well as the spatial discretization C and $4C$ number of ρ parameters in the weight function (3.2). Since the nature of ambulance demand does not change drastically over time, these estimations may be performed infrequently in practice (at most several times a year).

For K, we can use the typical Gaussian kernel, or for additional computational savings, the Epanechnikov kernel with bounded support. We can select the bandwidth $\boldsymbol{H}$ via the plug-in method [89] or smoothed cross-validation [26]. We can also adopt one of many fast computational methods for KDE, including kd-trees [10], ball trees [58], dual trees [36] and statistical regular pavings [68].

For the weight function (3.2), we can choose the discretization mesh or C *a priori* or via cross-validation. A larger value of C allows personalized temporal patterns on a finer grid, but if C is too large, data may become too sparse for accurate estimation of temporal dependencies. In our application, C is best chosen to be close to 20, yielding discrete regions that are roughly 5 km by 5 km each. The $4C$ number of ρ parameters in the weight function could be chosen in a number of standard ways; for instance we could use stochastic gradient ascent to maximize the joint likelihood of training data. For accurate estimation, we would need to use training data with tens of thousands of observations, and incur non-trivial computational cost. Here we introduce a much faster alternative method to estimate these parameters.

In Equation (3.2), w_c measures how positively correlated two demand densities $(u - t)$ periods apart are at cell c. We can directly estimate this correlation as follows. For each cell c, we can approximate its demand density for any period by the proportion of observations arising from this cell out of all observations from that period. We can then obtain a time series of proportions and compute

its (discrete) autocorrelation function $A_c(\ell)$ for lag $\ell \in \{1, \ldots, L\}$, where L is the maximum lag considered. Typically L can be taken to be around several weeks (hundreds of one-hour periods). The non-negative part of this autocorrelation, $A_c^+(\ell)$, or a smoothed version of it, is precisely what w_c aims to capture. For example, Figure 2 (a) shows an example of the autocorrelation function $A_c(\ell)$, and the grey lines in Figure 2 (b) shows the corresponding $A_c^+(\ell)$, for $\ell \in \{1, \ldots, 672\}$ (up to 4 weeks of one-hour periods).

The goal is to find appropriate ρ parameters such that w_c best fits the shape of A_c^+. To do this, we would like to minimize the sum of squared errors between $\rho_{0,c} w_c(\ell)$ and $A_c^+(\ell)$ at all time lags ℓ from 1 to L. We can find the optimal $\rho_{0,c}$ to $\rho_{4,c}$ for this minimization using gradient descent or grid search. The extra parameter $\rho_{0,c}$ is needed to scale w_c to curve-match the magnitude of A_c^+, and is of no real significance. Of greater importance is curvature or shape of A_c^+, which is captured in $\rho_{1,c}$ to $\rho_{4,c}$. To make w_c comparable across all C cells, we need to normalize w_c such that the area under w_c up to L is the same across different cells.

In summary, we estimate the ρ parameters in C minimization problems: for each $c \in \{1, \ldots, C\}$,

$$\begin{aligned} &\min_{\rho_{j,c}, \forall\, j \in \{0,\ldots,4\}} \sum_{\ell=1}^{L} \left(A_c^+(\ell) - \rho_{0,c} w_c(\ell)\right)^2 \\ &\text{s.t.} \sum_{\ell=1}^{L} w_c(\ell) = 1. \end{aligned} \tag{3.4}$$

This computation is much more efficient than the joint estimation of $4C$ parameters by maximizing likelihood. Here, we do not need to involve the kernel density estimator, nor loop through tens of thousands of ambulance demand observations. We can easily compute the C minimization problems in parallel. For each cell, we have a low-dimensional (5 parameters) problem with a small number of observations L (around hundreds of hours of time lags). A wide array of standard algorithms for solving optimization problems can be applied.

For example, we can Lagrangian relax the constraint into the objective and use the genetic algorithm or particle swarm.

3.2.2 Prediction

Once the infrequently performed parameter estimation is done, predictions for any future time period can be calculated instantaneously using short sliding windows of length L. We can additionally refine or customize the prediction procedure in the following two ways.

First, to boost predictive accuracy, we can bilinearly interpolate the weight values smoothly over the spatial domain instead of taking only C sets of values on a discretized grid. This is appropriate since we believe that the temporal patterns vary smoothly across the spatial domain. It also mitigates the sensitivity to predictions induced by choices of C.

Secondly, we can impose an omission threshold value, O, for the weights. If the weight of a past observation $\boldsymbol{s}_{t,i}$ is below this threshold, i.e., if $w(\boldsymbol{s}_{t,i}, u) < O$, we can omit this observation in the calculation of weighted KDE by overriding $w(\boldsymbol{s}_{t,i}, u) = 0$. The threshold can be chosen to balance the tradeoff between computational expense and predictive accuracy.

3.3 Predicting Toronto ambulance demand

The computation has two stages. In the first stage, we estimate or choose all parameters, including the kernel K, bandwidth $\boldsymbol{H}$, discretization C and $4C$ number of ρ parameters. This estimation only needs to be performed infrequently. For this parameter estimation, we use Toronto ambulance data from January to July 2008. Figure 3.1 (a) shows the spatial locations of all observations from this 7-month period. In the second stage, we predict future ambulance demand on a sliding window of length $L = 672$ (4 weeks, around $15,000$ observations) for

each one-hour period from August to December 2008.

In estimation, we choose the Gaussian kernel for K, select the bandwidth $\boldsymbol{H}$ via the plug-in method [89] and discretize Toronto into $C = 21$ equally-sized regions. We estimate the ρ parameters in the weight function using the method detailed in Section 3.2. As an example, we outline the cell c covering downtown Toronto in Figure 3.1 (a). We show in the top panel of Figure 3.1 (b) the autocorrelation function A_c for the proportions of observations arising from this downtown cell out of all observations across hourly time periods. This autocorrelation function indicates weekly, daily seasonalities and low-order serial dependence. The bottom panel of Figure 3.1 (b) shows in gray A_c^+ and in black the fitted weight function $\rho_{0,c} w_c$ for downtown, with $\rho_{1,c} = 0.95$ (short-term serial dependence), $\rho_{3,c} = 0.001$ (daily seasonality), $\rho_{4,c} = 0.145$ (weekly seasonality) and $\rho_{2,c} = 0.9995$ (discounting for seasonalities). The fitted weight function provides interpretable basis to understand exactly which historical observations are the most important for prediction. For example, from Figure 3.1 (b), an EMS manager can recognize that at downtown, ambulance demand in the past day or two and corresponding hour of the past few weeks are the most important. We checked the cross-correlation among the 21 weight functions estimated at different regions in Toronto. Neighboring weight functions showed some association, but those far apart are not correlated.

Once parameter estimation is done, we predict forward using a sliding window of 4 weeks for each one-hour period from August to December 2008. Figure 3.2 shows the predictive densities on August 6, 2008 (Wednesday) at two different time periods. The ambulance demand is, not surprisingly, concentrated at the heart of downtown during day time on Wednesday (Figure 3.2 (b)), and more spread out throughout the city during night time on Wednesday (Figure 3.2 (a)). This illustrates that the proposed model can differentiate temporal patterns at different time periods and locations.

We compare stKDE to the following competing methods

(a) The MEDIC method, which is an industry practice implemented in

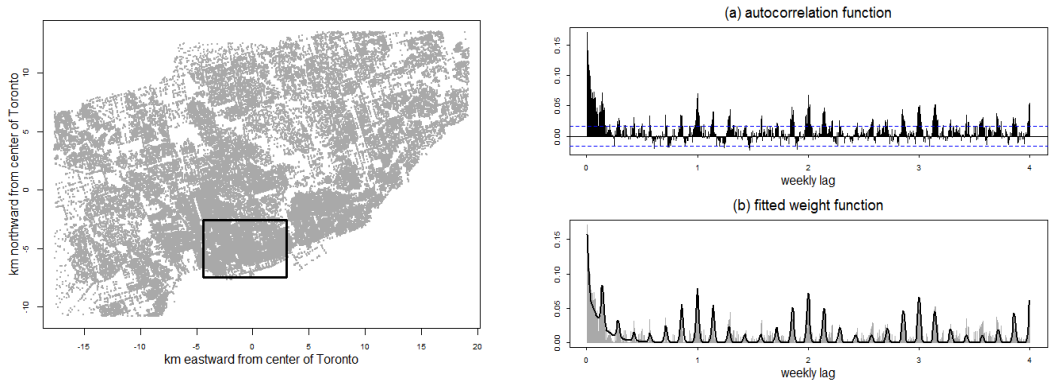

Figure 3.1: Left: spatial locations of all Toronto ambulance demand data from January to July 2008. To evaluate location-specific weight functions, we discretize the spatial domain into 21 cells, and here we outline the cell containing downtown Toronto. Right: (top) the autocorrelation function of the proportions of observations arising from the rectangle in Figure 1 over all observations across one-hour periods; (bottom) the fitted weight function (black) against the nonnegative part of the autocorrelation function (gray).

Charlotte-Mecklenburg, NC (§1). We implement this method as far as we have data. The cell count in a 1-km^2 region and a 1-hour period is predicted by the average of corresponding cell counts in the preceding 4 weeks in the past two years. The log predictive density produced by MEDIC for August 6, 2008 (Wednesday) at 2 - 3 am is shown in Figure 3.3. Compared to Figure 3.2 (a), the MEDIC prediction appears much noisier.

(b) Two naive KDEs, (i) using data from the most recent hour to predict the next hour, and (ii) using all data from the past four weeks with equal weights (this produces a spatial only model, with almost no temporal variation).

(c) A time-varying Gaussian mixture model. We quote results from Table 2.1 implemented on Toronto data with different training / testing months and various modeling specifications (e.g., number of components). The computational expense is considerable.

To compare the statistical predictive accuracies of our model and the indus-

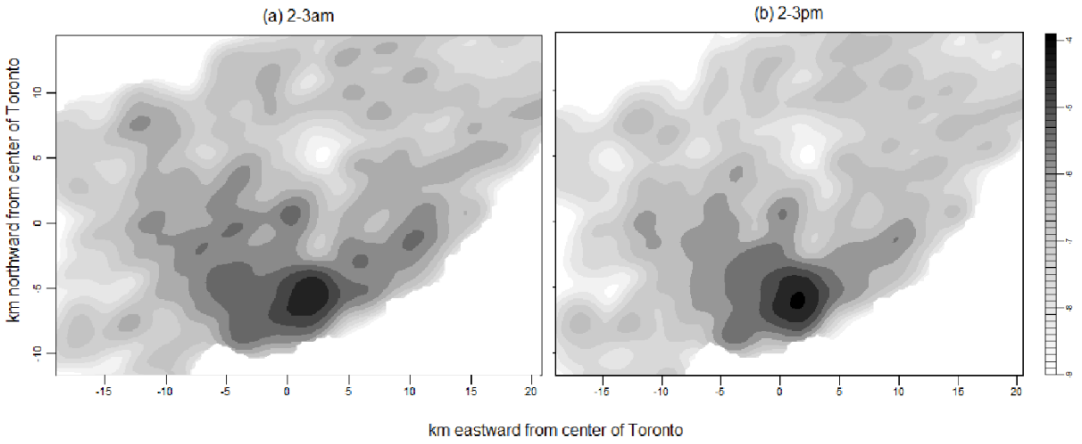

Figure 3.2: Log predictive density using stKDE for Aug 6, 2008 (Wednesday) at (a) 2 - 3 am (demand more spread out at night) and (b) 2 - 3 pm (demand concentrated at downtown during the day).

try method, we use the metric of average log score (ALS, defined in Equation (2.9) in Chapter 2). It is the average log likelihood of test data. A less negative average log score indicates better performance.

We show in Table 3.1 the predictive accuracies produced by our method. We present three variations of prediction: (i) using the estimated discretized weight functions w_c as they are, (ii) spatially interpolating the estimated weight values, and (iii) imposing an omission threshold on the estimated weight values such that each prediction uses a comparable amount of data as the industry method (about 200 observations).

The stKDE method significantly outperforms the MEDIC method (industry practice). It also outperforms the naive KDE methods, demonstrating the utility of incorporating spatio-temporal patterns via the weight functions. Our performance is comparable to time-varying GMM as it is implemented on Toronto data with different training / testing months and modeling specifications. Among the three variations of stKDE, allowing for bilinear interpolation of weight values improves the predictive accuracy slightly. In the third variation, including the omission threshold leads to a small loss of accuracy but re-

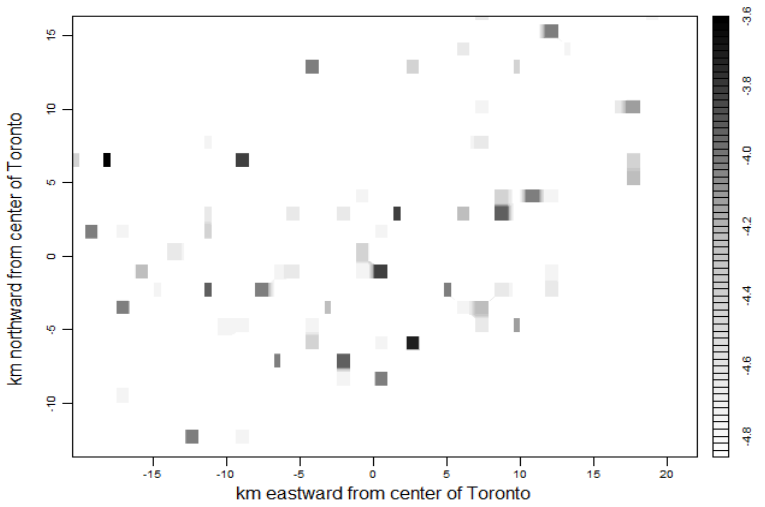

Figure 3.3: Log predictive density using industry method for Aug 6, 2008 (Wednesday) at 2 - 3 am. Figure 3.2 (a) shows the prediction by stKDE for the same period, which is less noisy.

duces computational cost significantly to be comparable to the industry method.

Prediction method		**Accuracy**
stKDE		−6.106
	+ interpolation	−6.102
	+ threshold (less data)	−6.635
MEDIC		−8.642
naiveKDE	most recent hour	−6.921
	all equal weights	−6.254
GMM		−6.072 to −6.149

Table 3.1: Predictive accuracies of stKDE and competing methods. Results of GMM are quoted from Table 2.1 implemented on Toronto data with various training / testing months and model specifications.

The infrequent estimation of weight functions and bandwidth takes several hours on a personal computer. This offline training is significantly shorter than that of GMM (inferring latent component labels in GMM is costly). It does not

take much longer than estimating bandwidths for naive KDE methods. Once estimation is done, making each new prediction is instantaneous (a few seconds). We could further reduce the computational expense of stKDE by parallelizing weight estimation, using a tree-based algorithm for fast KDE computation, using a bounded kernel function, or creating a look-up table of densities (none of these was done).

3.4 Discussion

We propose a spatio-temporal weighted kernel density estimator to predict spatio-temporal ambulance demand in Toronto with higher accuracy than and comparable computational cost as a typical industry practice.

We propose a spatio-temporal weighted kernel density estimator. The spatial kernel of each historical observation is multiplied with a weight value to indicate the informativeness of this historical observation to the current predictive task. The spatio-temporal weight functions are inferred from dependencies in data, are unique to each neighborhood and can be updated regularly. This is an improvement from the ad hoc heuristic that only accounts for the weekly and yearly seasonality across the entire city. The weight functions are also flexible to represent various spatial and temporal characteristics. They are easy to experiment with, visualize and interpret by non-experts. Moreover, stKDE easily handles missing data by placing zero weight and scaling up weights on other data proportionally. It can also easily predict many hours or days into the future.

The proposed method provides efficient estimation of the weight function, and offers customizable prediction to balance the tradeoffs between accuracy and computational cost. It is straightforward to implement by non-specialized users and scalable to large-scale datasets, and can be easily generalized to a wide range of other applications with spatial-temporal point process data.

CHAPTER 4
SPATIO-TEMPORAL KERNEL WARPING

We propose a novel method for modeling spatio-temporal ambulance demand against a complex set of spatial structures and geographical features. We follow Chapters 2 and 3 in predicting in discrete time and continuous space. To predict ambulance demand for a future time period, we only have a sparse set of historical data that is very relevant for this prediction (labeled data). We fit a KDE on them, but warp the kernels to a larger set of historical data regardless of their direct relevance to this particular predictive task (point cloud). This point cloud describes our belief about the spatial structure on which the labeled data lies. It captures exterior and interior boundaries without needing to explicitly define boundaries and boundary conditions. It also incorporates a wide range of complex spatial similarities and discontinuities, such as roads, city blocks, and neighborhoods of varying shapes and densities. Intuitively, this warping can be thought of as a regularization that penalizes radical departure from and encourages flow of information along our intuition of the geography. In a Bayesian sense, it can also be thought of as imposing a prior based on how similar or different the point process is across different locations. Such a regularization or prior is especially beneficial when the labeled data is sparse. We select the kernel bandwidth and the degree of warping efficiently via cross-validation. Both of these parameters can be made time- and/or location-specific.

We implement this method on ambulance demand data from Melbourne in years 2011 and 2012 introduced in Section 1.3. The proposed kernel warping model gives significantly more accurate predictions than previous approaches, including the MEDIC method as an industry practice, unwarped KDE, and GMM.

Material from this chapter is to be submitted in an article titled "Predicting ambulance demand using kernel warping", authored by Zhengyi Zhou and David S. Matteson [96].

We develop the kernel warping model in Section 4.1. We construct an un-

warped KDE in Section 4.1.1, warp the kernels to the point cloud in Section 4.1.2, and allow for time and location-specific warping in Section 4.1.3 for the Melbourne data. Some details on computation are included in Section 4.1.4. We show the empirical results for Melbourne ambulance demand in Section 4.2, and conclude in Section 4.3.

4.1 Model

We model Melbourne's ambulance demand on a continuous spatial domain $\mathcal{S} \subseteq \mathbb{R}^2$ and a discretized temporal domain of one-hour intervals $\mathcal{T} = \{1, 2, \ldots\}$. Let $\mathbf{s}_{t,i}$ be the location of the i-th ambulance demand arising from the t-th time period, for $i \in \{1, \ldots, n_t\}$, where n_t is the total number of ambulances demanded in the t-th period. Similar to Chapters 2 and 3, we model $\{\mathbf{s}_{t,i} : i = 1, \ldots n_t\}$ as an NHPP over $\mathcal{S}$ for each t, with intensity λ_t. We decompose the intensity function as $\lambda_t(\mathbf{s}) = \delta_t f_t(\mathbf{s})$ (as in Equation 2.1), for $\mathbf{s} \in \mathcal{S}$, where $\delta_t = \int_{\mathcal{S}} \lambda_t(\mathbf{s})\, d\mathbf{s}$ is the aggregate demand intensity over the spatial domain, and $f_t(\cdot)$ is the continuous spatial *density* of the demand at time t. We focus on predicting the spatio-temporal demand density $\{f_t\}$ as in previous chapters.

4.1.1 Spatio-temporal KDE

Suppose we want to predict Melbourne's ambulance demand for a future 1-hour period u. Given the prominent weekly seasonality, the most relevant observations are from the corresponding hour of the week for the past M weeks. They constitute the labeled data for this predictive task. This approach is aligned with the industry practice, and is shown to work well in [98]. We choose the sliding window width M *a priori*. With a larger M, the training data is less sparse, but each training becomes more expensive and less adaptive to recent changes in demand patterns (e.g., summer vs. winter). The industry and recent studies have considered M between 4 and 8. We set $M = 8$, resulting in an average labeled data size of about 300 points (ranging from 100 to 450 for different peri-

ods). Let $\mathcal{T}_u = \{u - 168m : m \in \{1, \ldots, M\}\}$ denote the set of labeled time periods, in which 168 is the number of 1-hour periods in a week.

Starting with a simple KDE on the labeled data, we predict for any $\mathbf{x} \in \mathcal{S}$,

$$f_u(\mathbf{x}) = \frac{1}{\sum_{t \in \mathcal{T}_u} n_t} \sum_{t \in \mathcal{T}_u} \sum_{i=1}^{n_t} k(\mathbf{x}, \mathbf{s}_{t,i} \,|\, \boldsymbol{H}) \,. \tag{4.1}$$

Here, k is the chosen bivariate spatial kernel with bandwidth matrix $\boldsymbol{H}$. As before, we use the Gaussian kernel, and choose bandwidth $\boldsymbol{H}$ via the plug-in method [89] or smoothed cross-validation [44, 26]. As mentioned in Section 1.2.4, we may be motivated to consider a spatial- and/or time-varying bandwidth $\boldsymbol{H}$ since data density varies substantially in space (downtown vs. neighborhoods) and time (midnight vs. rush hours).

4.1.2 Kernel warping

We would like to warp each kernel k in Equation (4.1) to a larger set of point cloud data that describes the spatial boundary and characteristics of Melbourne. We choose the point cloud data, construct an adjacency graph on the point cloud, define the graph Laplacian matrix, and warp the kernel to this Laplacian matrix. We discuss in detail each step.

Step 1 [Choosing the point cloud]: Typically in Laplacian eigenmap and kernel warping applications, all labeled and unlabeled data is used as the point cloud. In the context of spatial statistics and our application, there are several points of consideration:

(a) *Which points?* We consider all observations in the near past, irregardless of the time period. If we use the same sliding window width of $M = 8$ previous weeks, we are choosing from about 50,000 points.

(b) *How many points?* There is a trade-off: the more points we use for the point cloud, the better the quality of our approximation of the geography,

but the slower the computation. Since we are in a low-dimensional space of $\mathbb{R}^2$, we may not need a very large number of points to depict the most salient boundary and spatial structures. In our application, we find 1000 spatial points to represent Melbourne's geography reasonably well.

(c) *Points or mesh?* Alternative to using past observations, we can also use past data to define a pixelated spatial domain and use the centers of included pixels as the point cloud. Doing so we lose some resolution and information on data density, but may gain computationally if it can reduce the number of point cloud data significantly. A regularly spaced point cloud also induces a sparse, band-diagonal graph Laplacian matrix (to be discussed later), leading to further savings.

(d) *Global or local?* We can have one global point cloud for the entire spatial domain. We can also discretize the spatial domain into several regions with separate local point clouds. Local point clouds can provide computational advantages if they are smaller. They may also offer accuracy advantages if they depict finer-grain characteristics or allow for customized degree of warping at each locale. We discuss this further in Section 4.1.3.

In our application, we randomly sample 1000 historical observations as the point cloud for each "component" (to be explained in Section 4.1.3). We denote the set of point cloud data as $\{\boldsymbol{z}_i\}$ for $i \in \{1, \ldots, Z\}$. See Figure 4.1 (a) for an example cloud of 1000 points over the entire city of Melbourne. For our application, we find that predictive accuracy is not sensitive to the random sampling of the point cloud data. If it were, a larger point cloud might be needed, or predictions might be repeated and averaged over several point cloud samples.

Step 2 [Constructing the adjacency graph]: We construct a graph with nodes at each point in the point cloud and edges connecting points that are close. We represent this graph using a symmetric, positive semidefinite adjacency matrix A.

(a) *Which nodes to connect?* Knowledge about the spatial domain (e.g., inside

a building vs outside) or regularity of the point cloud (e.g., regular mesh) may inform a natural way to define how nodes should be connected. Without such knowledge, we can connect nodes z_i and z_j if z_i is among the n nearest neighbors of z_j *or* z_j is among the n nearest neighbors of z_i (symmetric relation). This requires us to choose n. In our experience, n should be big enough to ensure that the point cloud is sufficiently connected instead of being very fragmented, but small enough to emphasize local relationships. A second way is to connect nodes if the (Euclidean) distance between them is smaller than a threshold. This requires us to choose the threshold.

(b) *Weighted edges?* In the simplest case, we can set $A_{ij} = 1$ if nodes z_i and z_j are connected and 0 otherwise. Another idea suggested in [7] is to define weighted edges depending on the distance between points, i.e., $A_{ij} = \exp\{-\|z_i - z_j\|^2/r\}$ if z_i and z_j are connected and 0 otherwise. The authors note that they do not have a principled way of choosing r; we find it reasonable to choose r empirically by fitting an exponential distribution on all distances between connected nodes. They also note that in practice a binary adjacency graph works well, and we agree.

In our application, we use $n = 5$ nearest neighbors and binary weights to construct A. Figure 4.1 (a) shows the adjacency graph of a sample point cloud of size 1000. Again, we find our predictive accuracy to be insensitive towards any reasonable variations in these choices.

Step 3 [Constructing the Laplacian matrix]: The graph Laplacian matrix L is defined to be $L = D - A$, in which D is the diagonal degree matrix, with its diagonal entries being the column (or equivalently, row) sum of A, i.e., $D_{ii} = \sum_j A_{ij}$. L is a symmetric, positive semidefinite matrix. If the graph has multiple connected components, L can be rearranged into a block diagonal matrix, where each block is the respective Laplacian matrix for each connected component.

Here is the intuition of the Laplacian matrix. The (discrete) point cloud adja-

cency graph is an empirical approximation to our target (continuous) manifold of Melbourne geography. The (discrete) graph Laplacian matrix L is then an approximation to the (continuous) Laplace-Beltrami operator on this manifold. The Laplace-Beltrami operator is a manifold generalization of the Laplace operator, which is a linear second order differential operator on functions (in our case, kernels). This L induces a semi-norm on kernels which penalizes changes between adjacent nodes. There is a close analogy to heat flow; the heat (partial differential) equation has a Laplace operator in space. Intuitively, L guides how information (heat) spreads on the spatial structure (manifold approximated by graph) from any initial KDE (initial heat distribution).

Step 4 [Warping the kernels]: We warp each kernel k from Equation (4.1) to the point cloud to generate a new warped kernel $\tilde{k}$. For any $\mathbf{x} \in \mathcal{S}$ and any $\mathbf{s}$ in the set of labeled data,

$$\tilde{k}(\mathbf{x}, \mathbf{s} \,|\, \boldsymbol{H}) = k(\mathbf{x}, \mathbf{s} \,|\, \boldsymbol{H}) - \boldsymbol{k}_{\mathbf{x}}^T (\boldsymbol{I} + \lambda L \boldsymbol{K})^{-1} \lambda L \boldsymbol{k}_{\mathbf{s}}, \tag{4.2}$$

in which $\boldsymbol{k}_{\mathbf{x}} = [k(\mathbf{x}, \mathbf{z}_1 \,|\, \boldsymbol{H}), \ldots, k(\mathbf{x}, \mathbf{z}_Z \,|\, \boldsymbol{H})]$ and $\boldsymbol{k}_{\mathbf{s}} = [k(\mathbf{s}, \mathbf{z}_1 \,|\, \boldsymbol{H}), \ldots, k(\mathbf{s}, \mathbf{z}_Z \,|\, \boldsymbol{H})]$ are vectors of kernels evaluated at $\mathbf{x}$ or $\mathbf{s}$ and the point cloud data $\{\boldsymbol{z}_i\}$. Matrix $\boldsymbol{K} = \left[k(\mathbf{z}_i, \mathbf{z}_j \,|\, \boldsymbol{H})\right]_{i,j \in \{1,\ldots,Z\}}$ is a symmetric matrix of kernels evaluated at all pairs of point cloud data, and $\boldsymbol{I}$ is a Z by Z identity matrix. The parameter $\lambda > 0$ represents the degree of deformation. When $\lambda = 0$, we have $\tilde{k} = k$. When $\lambda \to \infty$, $\tilde{k}$ approaches a positive constant on the point cloud (steady state heat distribution).

Equation (4.2) is obtained by warping the Reproducing Kernel Hilbert Space (RKHS) associated with the chosen kernel. We modify the RKHS with a point-cloud semi-norm λL. This deforms the kernel k along a finite-dimensional subspace given by the point cloud data. The modified RKHS is shown to be another RKHS, i.e., $\tilde{k}$ is a properly defined kernel. See [76] and [9] for more details (they use the point cloud semi-norm of λL^p; we consider the simplified case where $p = 1$).

There are three interpretations of this type of kernel warping. The first is that of heat flow as mentioned before. We allow information (heat) to spread along the graph of the point cloud (approximately the manifold of geography). The second interpretation is a graph regularizer. Variations between adjacent nodes in the graph are penalized, and thus violation of the spatial structure implied by the point cloud are penalized. Lastly, in the Bayesian framework, kernel warping can informally be thought of as imposing a data-dependent informative prior to describe our belief of the data geometry.

We replace the regular Gaussian kernel k in Equation (4.1) with the new warped kernel $\tilde{k}$ defined in (4.2) to predict the density of ambulance demand f_u at a future time period u. We set *a priori* the sliding window width M, the point cloud data type / size, the number of nearest neighbors n, and the weights used to construct the Laplacian matrix. We estimate the Gaussian kernel bandwidth $\boldsymbol{H}$ and the degree of deformation λ.

We show in Figure 4.1 (b) and (c) examples of warping kernels. Three kernels of bandwidth $\boldsymbol{H} = \text{diag}(2, 2)$ are placed on three observations drawn and circled in red in Figure 4.1 (a). They are warped towards the point cloud in (a) with degree of deformation $\lambda = 0.5$ (b) and 2 (c). With a larger λ, the kernels conform to the spatial boundary and characteristics to a greater extent.

4.1.3 Spatio-temporal kernel warping

Melbourne's ambulance demand shows substantial density variations with patterns in time (midnight vs rush hour) and in space (downtown vs neighborhoods). It may be beneficial to allow bandwidth $\boldsymbol{H}$ and degree of deformation λ to vary with time and space. Ideally, we would like to find, in time and space, pockets of the point process with similar characteristics, and apply similar smoothing and deformation.

We discretize time according to our modeling aims, i.e., into 1-hour time periods. For each hour, we further discretize the spatial domain into a small

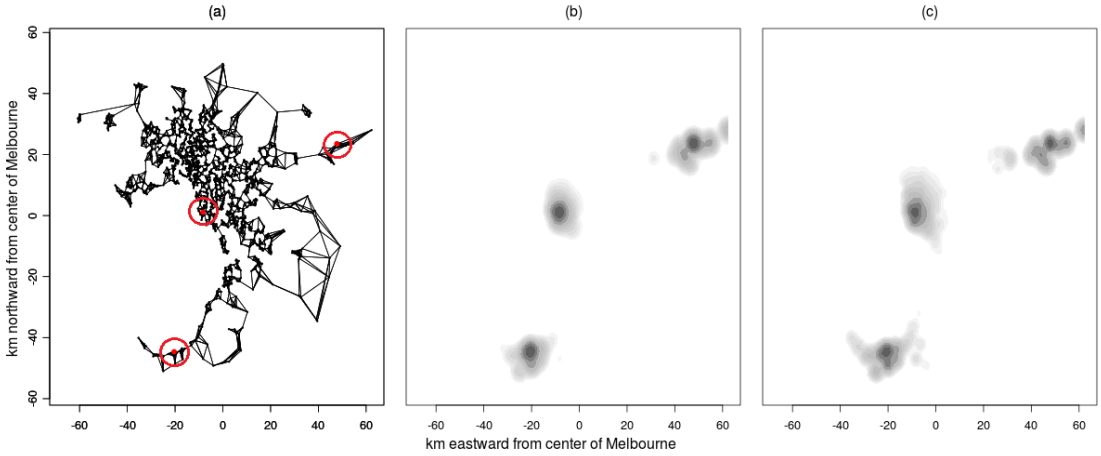

Figure 4.1: Examples of kernel warping: (a) the adjacency graph of a sample point cloud of size 1000; three observations are highlighted in red; (b) and (c), warped kernels centered at the these three observations with degrees of deformation $\lambda = 0.5$ and 2, respectively.

number of regions, as motivated by the behavior of labeled data for that time period. We call each subregion of each hour a *component*, and perform estimations and predictions independently on each component. The spatial discretization splits a global point cloud into local ones, cuts all edges connecting across regions, and decomposes the Laplacian matrix into blocks. Labeled data are also matched into components. We estimate a separate set of $\boldsymbol{H}$ and λ for each component by cross-validation (details in Section 4.1.4).

We discretize spatially by clustering. For any given future time period, we cluster on its labeled data (about 300 points). We allow different numbers of clusters and clustering configurations for each time period. In our application, this gives more accurate predictions than imposing a universal clustering configuration across time. We also obtain better results by clustering on labeled data rather than clustering on the point cloud data (the point cloud is much more similar across time than the labeled data). In our case, spatial characteristics across time are different enough that the gain in personalized modeling exceeds the loss in stablization offered by a common arrangement.

We choose to cluster using K-means based on Euclidean distance. K-means is fast and automated, clusters all points, and gives even clusters. Even cluster sizes are desirable because a very small cluster does not provide enough labeled data to reliably estimate parameters via cross-validation. To avoid this, we set a threshold minimum number of points in any cluster. We set the threshold at 15 points, which in practice limits the number of clusters to be below 8. If we fail to clear this threshold, we lower the number of clusters. Density-based clustering algorithms such as DBSCAN [28] and shared nearest neighbors [27] do not classify all points, and do not allow easy specification of the number of clusters. Graph-based clustering algorithms such as affinity propagation [30] and spectral clustering [57] do not cluster on Euclidean distance, and may be less intuitive for spatial point patterns. In our case, hierarchical clustering gives very uneven cluster sizes.

For each time period, we binary search for the best number of clusters based on validation likelihood. Increasing the number of clusters leads to two opposing forces. On the one hand, we add in another 1000 points into the cloud and the flexibility to customize parameters locally. On the other hand, we sparsify the labeled data for each cluster and destabilize parameter estimation. It is an empirical question for each time period whether we have enough labeled data to afford this increase in complexity. In our case, we find the number of clusters to be largely proportional to the size of labeled data.

4.1.4 Computation

We estimate the kernel bandwidth $\boldsymbol{H}$ and the degree of deformation λ for each spatio-temporal component. To reduce the dimensionality, we parametrize $\boldsymbol{H}$ to be a scalar multiple of the plug-in bandwidth $\boldsymbol{H}_{pi}$ obtained if we fit an unwarped KDE for the same component. That is we define $\boldsymbol{H} = \alpha \boldsymbol{H}_{pi}$, and estimate α. Alternatively, we can define a radial bandwidth $\boldsymbol{H} = \text{diag}(\beta, \beta)$, reducing the Gaussian kernel to a radial basis function. We use $\boldsymbol{H} = \alpha \boldsymbol{H}_{pi}$ because this parametrization gives slightly better performance in our preliminary

analysis. To estimate a full $\boldsymbol{H}$ is more difficult because $\boldsymbol{H}$ needs to be positive semi-definite.

We choose $\boldsymbol{H}$ and λ for each component using 5-fold cross-validation to maximize average validation likelihood. We implement a surrogate, derivative-free optimization procedure called the stochastic radial basis function (RBF) method [62, 63]. It is a fast algorithm for global optimization of computationally expensive objective functions. Each iteration builds an RBF model to approximate the expensive function, selects subsequent candidate points, and evaluates them in parallel. We choose this approach because our objective function (likelihood) evaluation is not instantaneous. It takes between 0.5 and 4 seconds, depending on the sizes of the labeled data and point cloud (Python code on a personal computer). We also do not have simple derivative computations. In our experience, 100 such evaluations are sufficient to provide a good optimum, competitive to those found by grid search, pattern search, or evolutionary algorithms. However, a wide range of optimization tools can be applied here.

In our application, we find a typical optimal α to be between 0.05 and 0.3. We need a concentration of heat which is then spread or warped to the point cloud. A typical optimal λ is between 0 and 2. Most time periods choose between 1 and 3 spatial components. We warm start the binary search for the number of clusters based on the size of the labeled data. The best configuration is usually found within 3 searches.

Given the prominent weekly seasonality, we believe that the corresponding parameter values are also similar from week to week. In fact, we believe that the nature of deformation and smoothing does not vary significantly over several months, and thus only estimate the parameters for a one-week cycle once every few months. With the most recent weekly set of parameter values, we predict forward in an online fashion with a sliding window of $M = 8$ weeks, making use of the most recent 8 weeks of data available. Each prediction is instantaneous.

The most expensive part of the computation is evaluating kernels between all pairs of point cloud data and taking the inverse of a large matrix. Several

local point clouds of reasonable sizes (< 2000) is computationally more efficient than one massive global point cloud. There are ways to optimize this computation, including using right division instead of inversion, saving precomputed kernel evaluation matrices and vectors, exploiting sparse, banded-diagonal Laplacian matrix, using a tree-based algorithm for fast KDE computation [36], and using a look-up table for Gaussian densities (most of these optimizations are not used in our implementation). The computation is "embarrassingly parallelizable", across validation likelihood evaluations and across spatio-temporal components.

4.2 Predicting ambulance demand for Melbourne

We would like to predict ambulance demand in Melbourne for every 1-hour period in March 2011. There are two stages to this computation. In the first stage, we estimate all parameters for a weekly cycle. The parameters include the spatial clustering configuration for each 1-hour period, as well as the parameters λ (degree of warping) and α (in bandwidth $\boldsymbol{H} = \alpha \boldsymbol{H}_{pi}$) for each spatial component in each 1-hour period. This estimation only needs to be performed very infrequently, and in our case, once. For this estimation, we use Melbourne ambulance demand data from 8 weeks in January and February 2011. In the second stage, we use the estimated weekly set of parameter values to predict future ambulance demand on a sliding window of 8 weeks for each 1-hour period in March 2011.

Figure 4.2 shows the predictive density predicted by kernel warping for two time periods on March 2, 2011 (Wednesday). We have only about 150 labeled data to predict for 2 - 3 am (a), and cross validate to use only 1 spatial component. We have almost 400 labeled data for 2 - 3 pm (b) and cross-validate to choose 5 spatial components.

We consider two variations in estimation: (i) spatio-temporal kernel warping (S-T param), in which we separately estimate parameters for each 1-hour

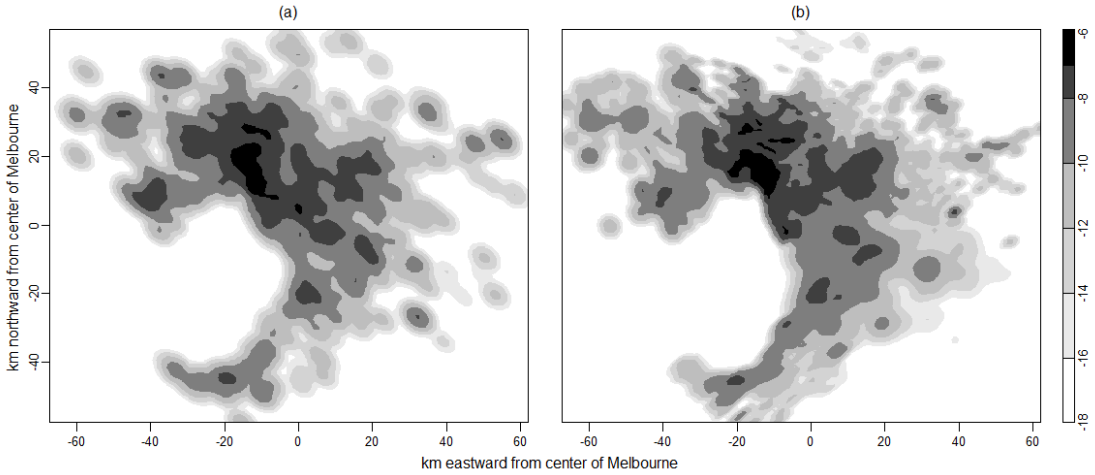

Figure 4.2: Log predictive densities using spatio-temporal kernel warping for March 2, 2011 (Wednesday) at (a) 2 - 3 am (night), and (b) 2 - 3 pm (day). For time period (a), we have sparse data and cross-validate to choose 1 spatial component. For time period (b), we have more data and choose 5 spatial components.

period and spatial region (via clustering, Section 4.1.3), and (ii) temporal kernel warping (T param), in which we separately estimate parameters for each 1-hour period (no spatial clustering). We show in Figure 4.3 the predictive densities produced by these two approaches for the same time period. The densities look similar, with slightly more details when we use spatio-temporal kernel warping (we cross-validate to select 3 spatial clusters).

We compare the proposed kernel warping models to the following competing methods

(a) The MEDIC method, which is an industry practice implemented in Charlotte-Mecklenburg, NC (Section 1.1.3). We implement this method as far as we have data. The cell count in a 1-km^2 region and a 1-hour period is predicted by the average of corresponding cell counts in the preceding 8 weeks.

(b) Unwarped KDE, corresponding to Equation (3.1). The bandwidth $\boldsymbol{H}$ is chosen via the plug-in method (PI) [89] and smoothed cross-validation

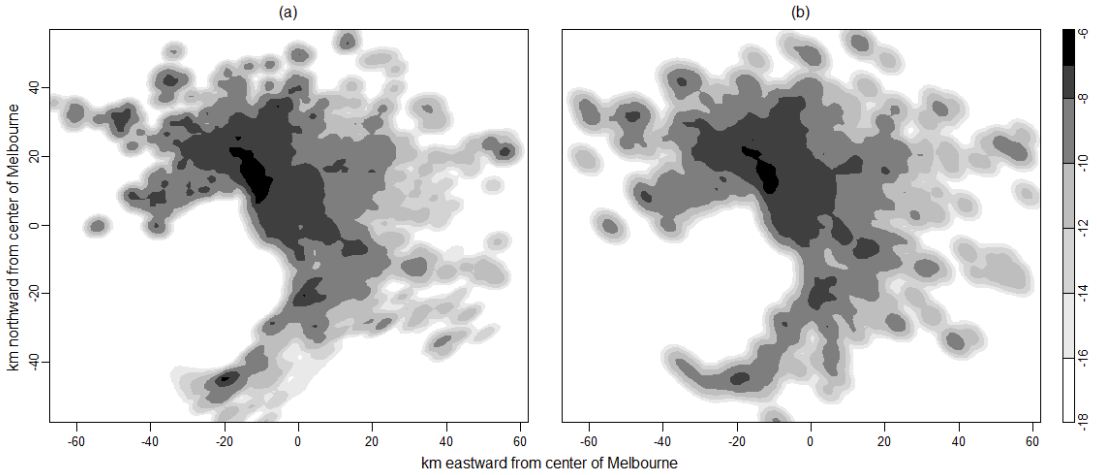

Figure 4.3: Log predictive densities for March 2, 2011 (Wednesday) at 10 - 11 am using (a) spatio-temporal kernel warping (3 spatial clusters), and (b) temporal kernel warping. The density in (a) shows slightly more details.

(SCV) [26]. This $\boldsymbol{H}$ is separately estimated for each time period, but does not vary in space.

(c) A simplified version of Gaussian mixture model (GMM) as detailed in [98]. The means and covariances of Gaussian components are fixed through time, and the mixture weights vary in time and constrained to be the same across weeks (but no autoregressive priors as detailed in Section 2.1.3. We use labeled data from the last 8 weeks, and consider fixed numbers of 15, 30 and 50 components.

Figure 4.4 show the log predictive density using the MEDIC method, unwarped KDE (PI), and GMM (30 components) for March 2, 2011 at 2 - 3 pm. These densities are comparable with Figure 4.2 (b), which shows the log predictive density for the same period predicted by the proposed kernel warping. Even with 400 labeled data, the MEDIC method gives exceedingly noisy predictions, while unwarped and KDE produce over-smoothed densities that do not adapt well to the spatial features of Melbourne.

We use several performance metrics to compare the statistical predictive ac-

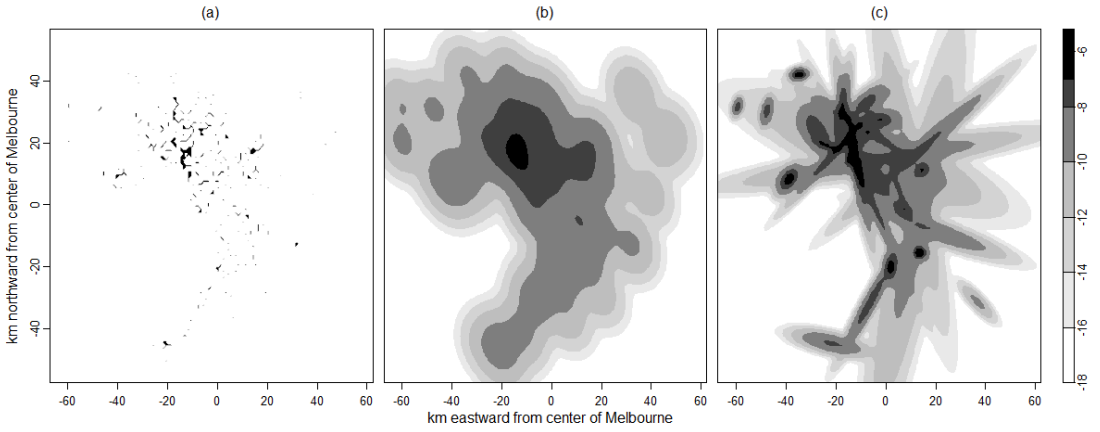

Figure 4.4: Log predictive densities using comparison methods for 2 - 3 pm on March 2, 2011 (Wednesday): (a) the MEDIC method (an industry practice); (b) unwarped KDE with bandwidth selected by the plug-in method (PI); (c) time-varying Gaussian mixture model with 30 components. These densities are to be compared to Figure 4.2 (b), which is the prediction using kernel warping for the same period.

curacies of different methods. First, we use average log score (ALS). Similar to Equation 2.9, we define

$$\text{ALS}\,(u) = \sum_{i=1}^{n_u} \log \hat{f}_u(\tilde{\mathbf{s}}_{u,i}),$$

for each test time period u in the set of all test time periods $\mathcal{T}_{test}$, in which $\{\tilde{\mathbf{s}}_{u,i}\}$ are the test data, and $\hat{f}_u(\cdot)$ is the predictive density for period u obtained by various methods. For the MEDIC method, we normalize cell counts to discrete density by dividing over the total count in each period.

Secondly, we compare accuracy in cell counts for every 1-km^2 region and 1-hour period. For the proposed kernel warping, unwarped KDE, and GMM, we discretize continuous predictions in space to each 1 km^2, and convert to counts by multiplying the total count for the period as predicted by the MEDIC method. We compute the root-mean-square error, both within the smallest rectangle enclosing all data (plotting window in Figures 4.2 and 4.4) (RMSE) and

within a pixelated data-driven boundary of Melbourne B (RMSE$_B$). For each test time period $u \in \mathcal{T}_{test}$,

$$\text{RMSE}\ (u) = \sqrt{\frac{1}{C}\sum_{c=1}^{C}(y_{u,c} - \hat{y}_{u,c})^2},$$

where C is the number of 1 km^2 cells in the rectangular observation window, $y_{u,c}$ and $\hat{y}_{u,c}$ are the actual and predicted count for period u and cell c respectively. For RMSE$_B$, we use cells c within the pixelated boundary B and C as the number of 1 km^2 cells within this boundary.

Additionally, since these cell counts (mostly 0s and 1s) are more appropriately modeled by a discrete distribution such as the Poisson distribution, we also compute the root-mean-square Anscombe residuals [4, 51], which specifically adjusts to measure predictive accuracy for Poisson data. Similarly, we consider within all of the rectangular window (ANSC) and within the boundary of Melbourne (ANSC$_B$). Using the same notations as above,

$$\text{ANSC}\ (u) = \sqrt{\frac{1}{C}\sum_{c=1}^{C}\left(\frac{(3/2)(y_{u,c}^{2/3} - \hat{y}_{u,c}^{2/3})}{\hat{y}_{u,c}^{1/6}}\right)^2},$$

and ANSC$_B$ similarly defined. We show in Table 4.1 the mean predictive accuracies of various methods, averaged across all test time periods $\mathcal{T}_{test}$ (all 1-hour periods in March 2011). A less negative ALS, and smaller RMSE, RMSE$_B$, ANSC, and ANSC$_B$ indicate better predictive accuracy. Both versions of kernel warping have a significant advantage over the comparison methods in all performance measures, especially in RMSE$_B$ and ANSC$_B$. Between the two versions of kernel warping, allowing parameters to be location-specific (in addition to being time-specific) provides additional benefits, even though a large number of time periods choose to use only 1 spatial component. We further show in Figure 4.5 the box-plots illustrating the variations of some of these metrics across time periods. Kernel warping has not only the best mean performance, but also the smallest variations across time periods.

Prediction method		ALG	RMSE	RMSE_B	ANSC	ANSC_B
Kernel warping	S-T param	−7.53	0.0500	0.0498	0.176	0.171
	T param	−7.56	0.0518	0.0514	0.178	0.172
(a) MEDIC		−10.11	0.0589	0.0996	0.479	0.810
(b) Unwarped KDE	PI	−8.14	0.0562	0.0950	0.199	0.334
	SCV	−8.15	0.0562	0.0950	0.194	0.325
(c) GMM	15 comp	−7.96	0.0562	0.0949	0.181	0.304
	30 comp	−7.87	0.0561	0.0948	0.191	0.323
	50 comp	−7.93	0.0561	0.0949	0.188	0.316

Table 4.1: Mean predictive accuracies across all 1-hour periods in March 2011 of the proposed kernel warping and competing methods. Kernel warping outperforms the competing methods.

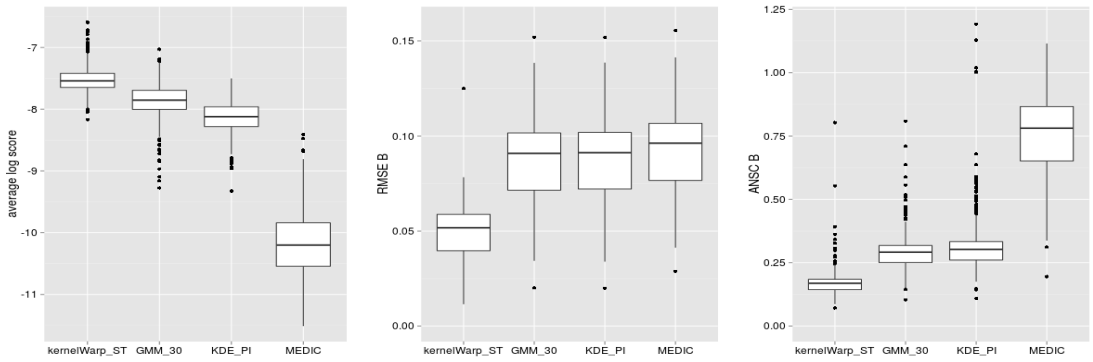

Figure 4.5: Box-plots of predictive accuracies of kernel warping (S-T parameters), GMM (30 comp), KDE (PI bandwidth), and the MEDIC method (an industry practice) over 672 test periods, as measured by average log score (left, less negative is better), RMSE_B (middle, smaller is better), and ANSC_B (right, smaller is better).

4.3 Discussion

We propose a kernel warping method that smooths intelligently towards geographical characteristics to overcome sparsity and model complex spatial fea-

tures at the same time. We demonstrate our proposed method predicts EMS demand in Melbourne more accurately than the state of the art in the practice and research of ambulance demand prediction.

To predict ambulance demand for any hour, we use a spatio-temporal kernel density estimator on the sparse set of most similar labeled data, but warp these kernels to a larger set of point cloud drawn from all historical observations regardless of labels. We construct an adjacency graph on this point cloud to approximate Melbourne's spatial boundaries, neighborhoods, and road networks in a data-driven manner. Kernels on labeled data are warped to encourage flow along and penalize flow orthogonal to this graph structure. This kernel warping can be interpreted (i) physically as heat flow on an empirical manifold, (ii) computationally as a regularization against large variations across adjacent nodes, and (iii) in the Bayesian framework as a prior of the spatial structure.

CHAPTER 5

DISCUSSION

5.1 Predicting spatio-temporal ambulance demand

Fine-resolution spatio-temporal ambulance demand predictions are crucial to optimal ambulance planning. The EMS industry practice and early studies are often simple and do not give accurate estimates. We provide three much-needed and highly accurate methods to predict spatio-temporal ambulance demand at fine scales.

First, in Chapter 2, we use finite mixture models to capture the complex temporal patterns and dynamics in this large-scale dataset. We demonstrate that the proposed method predicts the EMS operational performance much more accurately, reducing error by as much as two-thirds compared to an industry practice. Many management decisions seek to optimize this estimated operational performance; the proposed method predicts this optimization objective with more accuracy, leading to more confidence in optimization.

We have developed a set of easily generalizable tools suitable to analyze a wide range of spatio-temporal point process applications. We jointly estimate mixture component distributions over time to promote efficient learning of spatial structures, and describe spatial and temporal characteristics using mixture weights. This approach can be applied to various settings in which particular spatial aspects of the point process are time invariant, or data are too sparse at the desired temporal granularity to describe spatial structures accurately. The evolution of mixture weights provides a flexible and simple framework to explore complex temporal patterns, dynamics, and their interactions with space in a spatio-temporal point process. In this application, we capture seasonalities by constraining the mixture weights, and represent any location-specific dependence structure by imposing CAR priors on the mixture weights. We have also shown that prediction may be implemented with a variable number of components, inclusion of covariate information and incorporation of a spatial bound-

ary. The proposed method is parsimonious, flexible, straightforward to implement, and computationally-feasible for large-scale datasets.

The proposed model utilizes the same data as the current industry methods, and does not require any additional data collection. Future work can investigate the use of additional covariates, such as weather, special events, population and demographic variables, in addition to historical data. A further challenge is to collect and make use of data on population and demographic shifts across fine time scales, e.g., hourly. Additionally, a computationally-feasible way of incorporating the boundary of Toronto and accounting for the high concentration of observations near the boundary would be an important contribution.

Second, in Chapter 3, we use a spatio-temporal weighted kernel density estimator (stKDE) to predict spatio-temporal ambulance demand in Toronto with higher accuracy and comparable computational cost as a typical industry practice.

Methodologically, we multiply the spatial kernel of each historical observation with a weight value to indicate the informativeness of this historical observation to the current predictive task. The spatio-temporal weight functions are inferred from dependencies in data, are unique to each neighborhood and can be updated regularly. This is a step up from the ad hoc heuristic that only accounts for the same simply weekly and yearly seasonality across the entire city. The weight functions are also flexible to represent various spatial and temporal characteristic. They are easy to experiment, visualize, interpret and implement by non-specialized personnel from the EMS industry without requiring special statistical expertise. Moreover, stKDE easily handle missing data by placing zero weight and scaling up weights on other data proportionally. It can also easily predict many hours or days into the future.

We design efficient estimation of the weight function, and offer customizable prediction to balance the trade-offs between accuracy and computational cost. The tools we have developed can be easily generalized to a wide range of other applications with spatial-temporal point process data.

Third, in Chapter 4, we propose a novel predictive method using spatio-temporal kernel warping to overcome data sparsity and capturing complex spatial feature. To predict for each hour, we build a kernel density estimator on a sparse set of most similar data from relevant past time periods (labeled data), but warp these kernels to a larger set of past data irregardless of time periods (point cloud). The point cloud represents the spatial structure and geographical characteristics of Melbourne, including complex boundaries, road networks, and neighborhoods. Borrowing from manifold learning, kernel warping is done through graph Laplacian of the point cloud and can be interpreted as a regularization towards and imposing a prior of features of spatial domain. Kernel bandwidth and degree of warping can be efficiently estimated via cross-validation, and can be made time- and/or location-specific. Our proposed model gives significantly more accurate predictions compared to a current industry practice, an unwarped kernel density estimation, and a Gaussian mixture model.

Kernel warping circumvents the need to define boundaries and boundary conditions, which are often difficult in the practice of modeling point patterns on complex spatial domains. It also captures and exploits finer-grain internal spatial structure other than boundary features, which can be prominent in various heterogeneous environments such as cities, buildings, mountains, and forests. Kernel warping is not limited to density estimation. It can be adapted to model a wide range of functions and surfaces. It can be used to perform a broad set of tasks including prediction, classification, clustering, and visualization. Inferences on uncertainty, if desired, can be obtained by assessing cross-validation variance and warping kernels to different samples of point clouds. There is much flexibility in designing the point cloud and its Laplacian. We offer some discussions on these in the context of spatial and spatio-temporal point patterns. We also offer efficient estimation of kernel bandwidth and degree of warping local to time periods and locations via cross-validation. The proposed method is straightforward to implement and easy to experiment with. The tools we have developed can be easily generalized to model a wide range of spatial

or spatio-temporal point process on complex spatial domains.

5.2 Data mining in healthcare, operations and business

Data has transformed industries such as healthcare, operations and business. New challenges arise when we use data to inform decisions in these industries. The methodologies developed in this dissertation explore solutions to three such challenges.

5.2.1 How to exploit complex dependencies in data

Much effort has focused on developing sophisticated statistical methods for independent data, or incorporating dependence specific to one dimension, such as time or space. However, in real-world data, patterns across time, space, event type, users and other features are the norm. Additionally, patterns in different dimensions often interact to give rise to highly intricate dependence structures. For example, ambulance demand exhibits daily seasonality at densely populated areas, but not in lightly populated neighborhoods. Factors such as precipitation have different impacts on ambulance demand at different times (e.g., commute hours vs. at night) and locations (where snow is promptly cleared vs. where it is not).

To capture such complex dependencies is critical for accurate prediction; these dependencies provide invaluable information that we ought to take advantage of. The challenge is to design modeling schemes that allow flexible dependence representation. Using the time-varying Gaussian mixture model proposed in Chapter 2, we improve predictive accuracy by 20-25% compared to current industry practice. Incorporating location-specific temporal dynamics alone boosts predictive accuracy by 3%.

This is an important area of research with many interesting questions. For

example, event type adds another layer of dependence; crimes of different severities, demand of different priorities, or product promotions of different types are rarely independent. Most studies that jointly model multiple types of events assume independence among event types and between event types and time, location and features. Incorporating these dependencies will provide further insights for accurate prediction. Additionally, modern data in healthcare, operations and business comes in many forms; we would also like to model how language or opinions from texts, networks, or images, change with time, space, and other factors.

5.2.2 How to overcome information overload

Many predictions are made from proprietary black box models. We pour in all the data we have, with little understanding of how relevant each data component is; we also cannot easily understand how predictions are made. In contrast, it is very appealing to be able to draw out the most relevant aspects of data and represent them in an interpretable form for accurate inference. Predictions can be improved via dimension reduction. More importantly, domain experts such as doctors or business managers can easily understand which factors play central roles in prediction. There has been a lot of effort on ranking the importance of features (e.g., decision trees, association rule mining). In spatio-temporal predictions, I have shown that spatial and temporal information in data can provide important pathways to constructing interpretable models.

For example, using the spatio-temporal weighted kernel density estimation proposed in Chapter 3, we use a spatio-temporal weight function to represent exactly how important each historical data is to the current predictive task. We can potentially also represent which covariates are the most helpful in any predictive task. Domain experts in EMS can easily learn insights and act accordingly.

Spatial and temporal characteristics in the data are easy to visualize and can

be very powerful. It would be interesting to explore how they can provide interpretable and informative basis for other problems such as those in clustering and classification, by ranking data and/or features.

5.2.3 How to overcome sparsity in data

Ironically, we frequently encounter data sparsity despite having massive amounts of data. This is a result of our desire to predict at high resolution. We would like to predict product demand at Amazon for every product, shipping method and destination city, and predict ambulance demand for every hour, every 1 km^2 region, and every priority class. For many of these requirement combinations, we have little data. If we naively restrict our attention to within each combination, we can do no better than very noisy predictions, while avoiding always predicting zero events.

Temporal and spatial characteristics in data often provide us with opportunities to borrow observations or information from outside of the target requirement combination. Here we have explored clustering, hierarchical structure, Bayesian modeling, mixture models or kernel methods, and pooling / combining data across time and/or space. Most of these methods are aided by understanding patterns and dynamics in time and space.

Subsequent interesting questions to ask include how to intelligently disaggregate uncertainty back to each combination. Second, mainstream predictive performance measures tend not to behave very well on sparse data. Many measures become undefined with sparse data, and always predicting zero events frequently has the highest performance. The quest is to look for measures robust and fair for sparse data. Finally, we eventually have to answer the foundational question of what is predictable and what is not.

BIBLIOGRAPHY

[1] C. C. Aggarwal. A framework for diagonosing changes in evolving data streams. In *ACM SIGMOD International Conference on Management of Data*, pages 575 – 586, 2003.

[2] S. Aldor-Noiman, P. Feigin, and A. Mandelbaum. Workload forecasting for a call center: Methodology and a case study. *Annals of Applied Statistics*, 3:1403–1447, 2009.

[3] C. A. Aldrich, J. C. Hisserich, and L. B. Lave. An analysis of the demand for emergency ambulance service in an urban area. *American Journal of Public Health*, 61:1156–1169, 1971.

[4] F.J. Anscombe. Contribution of discussion paper by h. hotelling 'new light on the correlation coefficient and its transforms'. *Journal of the Royal Statistical Society: Series B*, 15:229 – 230, 1953.

[5] A. Baddeley and R. Turner. spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12:1–42, 2005.

[6] S. Banerjee, A. E. Gelfand, and B. P. Carlin. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall, New York, 2003.

[7] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.

[8] M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56:209 – 239, 2004.

[9] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning*, 7:2399 – 2434, 2006.

[10] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 1975.

[11] J.M. Bernardo. Expected information as expected utility. *Annals of Statistics*, 7:686–690, 1979.

[12] C. Berzuini and D. Clayton. Bayesian analysis on survival on multiple time scales. *Statistics in Medicine*, 13:823–838, 1994.

[13] J. Besag, J. C. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43:1–59, 1991.

[14] J. E. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–225, 1974.

[15] S. Brooks, A. Gelman, G. Jones, and X. Meng, editors. *Handbook of Markov Chain Monte Carlo*. Handbooks of Modern Statistical Methods. Chapman and Hall/CRC, New York, 2011.

[16] C. Brunsdon, J. Corcoran, and G. Higgs. Visualising space and time in crime patterns: a comparison of methods. *Computers, Environment and Urban Systems*, 31:52 – 75, 2007.

[17] T. Cacoullos. Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18:197 – 189, 1966.

[18] R. T. Cadigan and C. E. Bugarin. Predicting demand for emergency ambulance service. *Annals of Emergency Medicine*, 18:618–621, 1989.

[19] O. Cappé, C. Robert, and T. Rydé. Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers. *Journal of the Royal Statistical Society: Series B*, 65:679–700, 2003.

[20] A. Chakraborty and A. E. Gelfand. Analyzing spatial point patterns subject to measurement error. *Bayesian Analysis*, 5:97–122, 2010.

[21] N. Channouf, P. L'Ecuyer, A. Ingolfsson, and A. Avramidis. The application of forecasting techniques to modeling emergency medical system calls in calgary, alberta. *Health Care Management Science*, 10:25–45, 2007.

[22] A.P. Dawid. Statistical theory: the prequential approach. *Journal of the Royal Statistical Society: Series A*, 147:278–292, 1984.

[23] P. J. Diggle. *Statistical Analysis of Spatial Point Patterns*. Arnold, London, second edition, 2003.

[24] M. Ding, L. He, D. Dunson, and L. Carin. Nonparametric Bayesian segmentation of a multivariate inhomogeneous space-time Poisson process. *Bayesian Analysis*, 7:235–262, 2012.

[25] D.L. Donoho and C. Grimes. Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. In *Proceedings of the National Academy of Sciences*, volume 102, 2005.

[26] T. Duong and M. L. Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32:485–506, 2005.

[27] L. Ertöz, M. Steinbach, and V. Kumar. Finding clusters of different sizes, shapes and densities in noisy, high dimensional data. In *Proceedings of the SIAM International Conference on Data Mining*, pages 47 – 58, 2003.

[28] M. Ester, H.P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noice. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[29] J. M. Flegal, M. Haran, and G. L. Jones. Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, 23:250–260, 2008.

[30] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972 – 976, 2007.

[31] A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457472, 1992.

[32] T. Gneiting and A. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

[33] J. B. Goldberg. Operations research methods for the deployment of emergency service vehicles. *EMS Management Journal*, 1:20–39, 2004.

[34] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society: Series B*, 14:107–114, 1952.

[35] Google Maps. Map of melbourne, australia. Web, 2015.

[36] A. G. Gray and A. W. Moore. Nonparametric density estimation: toward computational tractability. In *Proceedings of the SIAM International Conference on Data Mining*, 2003.

[37] P. Green. Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[38] S. G. Henderson. Operations research tools for addressing current challenges in emergency medical services. In J. J. Cochran, L. A. Cox, P. Keskinocak, J. P. Kharoufeh, and J. C. Smith, editors, *Wiley Encyclopedia of Operations Research and Management Science*. Wiley, New York, 2009.

[39] R. Ibrahim and P. L'Ecuyer. Forecasting call center arrivals: fixed-effects, mixed-effects, and bivariate models. *Manufacturing and Service Operations Management*, 15:72–85, 2013.

[40] J. B. Illian, A. Penttinen, H. Stoyan, and D. Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley, Chichester, England, 2008.

[41] E. M. Jansenberger and P. Staufer-Steinnocher. Dual kernel density estimation as a method for describing spatio-temporal changes in the upper Austrian food retailing market. In *AGILE Conference on Geographic Information Science*, 2004.

[42] C. Ji, D. Merl, and T. B. Kepler. Spatial mixture modeling for unobserved point processes: Examples in immunofluorescence histology. *Bayesian Analysis*, 4:297–315, 2009.

[43] G. L. Jones, M. Haran, B. S. Caffo, and R. Neath. Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101:1537–1547, 2006.

[44] M. C. Jones, J. S. Marron, and B. U. Park. A simple root n bandwidth selector. *Annals of Statistics*, 19:1919–1932, 1991.

[45] L. Knorr-Held. Conditional prior proposals in dynamic models. *Scandinavian Journal of Statistics*, 26:129–144, 1999.

[46] L. Knorr-Held and J. Besag. Modelling risk from a disease in time and space. *Statistics in Medicine*, 17:2045–2060., 1998.

[47] A. Kottas and B. Sansó. Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. *Journal of Statistical Planning and Inference*, 137:3151–3163, 2007.

[48] T. O. Kvalseth and J. M. Deems. Statistical models of the demand for emergency medical services in an urban area. *American Journal of Public Health*, 69:250–255, 1979.

[49] J. Marin, K. Mengersen, and C. Robert. Bayesian modeling and inference on mixtures of distribution. In *Handbook of Statistics*, pages 459–507. Elsevier, 2005.

[50] D. S. Matteson, M. W. McLean, D. B. Woodard, and S. G. Henderson. Forecasting emergency medical service call arrival rates. *Annals of Applied Statistics*, 5:1379–1406, 2011.

[51] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall / CRC, London, 2nd edition, 1989.

[52] A. C. Micheas. Hierarchical bayesian modeling of marked non-homogeneous Poisson processes with finite mixtures and inclusion of covariate information. *Journal of Applied Statistics*, 41:2596–2615, 2014.

[53] A. C. Micheas, C. K. Wikle, and D. R. Larsen. Random set modeling of three-dimensional objects in a hierarchical Bayesian context. *Journal of Statistical Computation and Simulation*, 84:107–123, 2012.

[54] A.C. Micheas. Hierarchical Bayesian random sets with applications to growth models. In *Proceedings of the Joint Statistical Meetings, Bayesian Statistical Science Section*, Miami Beach, FL, 2011.

[55] J. Møller and R. P. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC, London, 2004.

[56] T. Nakaya and K. Yano. Visualising crime clusters in a space-time cube: an exploratory data analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, 14:223 – 239, 2010.

[57] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2001.

[58] S. M. Omohundro. Five balltree construction algorithms. *International Computer Science Institute Technical Report*, 1989.

[59] C. M. Procopiuc and O. Procopiuc. Density estimation for spatial data streams. *Advances in Spatial and Temporal Databases*, 3633:109 – 126, 2005.

[60] T. Ramsay. Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B*, 64:307 – 319, 2002.

[61] C. E. Rasmussen and C. K. I. Williams. *Gaussian Process for Machine Learning*. MIT, Boston, 2006.

[62] R.G. Regis and C.A. Shoemaker. A stochastic radial basis function method for the global optimization of expensive functions. *INFORMS Journal on Computing*, 19:497 – 509, 2007.

[63] R.G. Regis and C.A. Shoemaker. Parallel stochastic global optimization using radial basis functions. *INFORMS Journal on Computing*, 21:411 – 426, 2009.

[64] S. Richardson and P. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B*, 59:731–792, 1997.

[65] B.D. Ripley and J.P. Rasson. Finding the edge of a Poisson forest. *Journal of Applied Probability*, 14:483 – 491, 1977.

[66] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2nd edition, 2004.

[67] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[68] R. Sainudiin, G. Teng, J. Harlow, and D. Lee. Posterior expectation of regularly paved random histograms. *ACM Transactions on Modeling and Computer Simulation*, 23, 2013.

[69] D. W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiley and Sons, New York, 1992.

[70] H. Setzler, C. Saydam, and S. Park. EMS call volume predictions: A comparative study. *Computers & Operations Research*, 36:1843–1851, 2009.

[71] H. Shen and J. Z. Huang. Forecasting time series of inhomogeneous Poisson process with application to call center management software. *Annals of Applied Statistics*, 2:601–623, 2008.

[72] H. Shen and J. Z. Huang. Intraday forecasting and interday updating of call center arrivals. *Manufacturing and Service Operations Management*, 10:391–410, 2008.

[73] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

[74] K. F. Siler. Predicting demand for publicly dispatched ambulances in a metropolitan area. *Health Services Report*, 10:254–263, 1975.

[75] B.W. Silvermann. *Density estimation for statistics and data analysis*. Chapman & Hall, London, 1986.

[76] V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 824 – 831, 2005.

[77] A.J. Smola and R. Kondor. Kernels and regularization on graphs. In *Learning Theory and Kernel Machines, Lecture Notes in Computer Science*, pages 144–158, 2003.

[78] M. Stephens. Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Annals of Statistics*, 28:40–74, 2000.

[79] A. Swersey. The deployment of police, fire, and emergency medical units. In *Handbooks in Operations Research and Management Science*, volume 6, pages 151–200. North-Holland, Amsterdam, 1994.

[80] M. A. Taddy. *Bayesian nonparametric analysis of conditional distributions and inference for Poisson point processes*. PhD thesis, University of California, Santa Cruz, 2008.

[81] M. A. Taddy. Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime. *Journal of the American Statistical Association*, 105:1403–1417, 2010.

[82] M. A. Taddy and A. Kottas. Mixture modeling for marked Poisson processes. *Bayesian Analysis*, 7:335–362, 2012.

[83] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540, 1987.

[84] J.B. Tenebaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319 – 2323, 2000.

[85] L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1762, 1994.

[86] W. Tych, D. Pedregal, P. Young, and J. Davies. An unobserved component model for multi-rate forecasting of telephone call demand: the design of a forecasting support system. *International Journal of Forecasting*, 18:673–695, 2002.

[87] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. Dimensionality reduction: a comparative review. *Journal of Machine Learning Research*, 10:66–71, 2009.

[88] J. L. Vile, J. W. Gillard, P. R. Harper, and V. A. Knight. Predicting ambulance demand using singular spectrum analysis. *Journal of the Operations Research Society*, 63:1556–1565, 2012.

[89] M. P. Wand and M. C. Jones. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9:97–116, 1994.

[90] J. Weinberg, L. D. Brown, and J. R. Stroud. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association*, 102:1185–1199, 2007.

[91] J. W. Wilesmith, M. A. Stevenson, C. B. King, and R. S. Morris. Spatio-temporal epidemiology of foot-and-mouth disease in two counties of great britain in 2001. *Preventive Veterinary Medicine*, 61:157 – 170, 2003.

[92] S.N. Wood, M.V. Bravington, and S.L. Hedley. Soap film smoothing. *Journal of the Royal Statistical Society: Series B*, 70:931 – 955, 2008.

[93] Z. Zhang, D. Chen, W. Liu, J. S. Racine, S. H. Ong, Y. Cheng, G. Zhao, and Q. Jiang. Nonparametric evaluation of dynamic disease risk: a spatio-temporal kernel approach. *PLoS ONE*, 6, 2011.

[94] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schoelkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, 2003.

[95] Z. Zhou and D.S. Matteson. Predicting ambulance demand: a spatio-temporal kernel approach. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page to appear, 2015.

[96] Z. Zhou and D.S. Matteson. Predicting ambulance demand using kernel warping. submitted, 2015.

[97] Z. Zhou and D.S. Matteson. Temporal and spatio-temporal models for ambulance demand. In H. Yang and E.K. Lee, editors, *Healthcare Data Analytics, Wiley Series in Operations Research and Management Science*. Wiley, New York, 2015.

[98] Z. Zhou, D.S. Matteson, D.B. Woodard, S.G. Henderson, and A.C. Micheas. A spatio-temporal point process model for ambulance demand. *Journal of the American Statistical Association*, 110:6–15, 2015.

[99] X. Zhu, J. Kandola, Z. Ghahramami, and J. Lafferty. Nonparametric transforms of graph kernels for semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2005.