Predicting Health and Safety

Essays in Machine Learning for Decision Support in the Public Sector

Dylan Fitzpatrick

Dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Machine Learning and Public Policy

August 28, 2020

Accepted by the Dissertation Committee and Approved by the Dean:

Daniel B. Neill (Co-Chair): Daniel B Neill Date: 8/28/2020

Wilpen Gorr (Co-Chair): Wilpen L. Gorr Date: 8/28/2020

Rayid Ghani: DocuSigned by: 17C7A292BB754F8... Date: 8/28/2020

Roni Rosenfeld: DocuSigned by: Roni Rosenfeld 122B515B2B7C4B4... Date: 8/28/2020

Department Head, Machine Learning Department

Ramayya Krishnan: DocuSigned by: Ramayya Krishnan 7D2BBEF5A516439... Date: 8/27/2020

Dean, Heinz College

Predicting Health and Safety

Essays in Machine Learning for Decision Support in the Public Sector

Dylan Fitzpatrick

Committee: Daniel B. Neill (Co-chair)
Wilpen Gorr (Co-chair)
Rayid Ghani
Roni Rosenfeld

Doctoral Dissertation
In partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Machine Learning and Public Policy

August 2020

For Tucker, my constant distraction and devoted friend.

Acknowledgements

Thank you to my mom, for always being available for calls and impromptu visits when I needed to recharge. Thank you Dad, for your advice and insight on all things *academic*, and for making my meandering path through grad school seem possible and worthwhile. To Sarah, my perpetual role model, who gave me the best piece of PhD advice I ever received ("Get a dog!"). And to Persis, for making the future seem so bright.

Thank you Rahul, for your unexpected friendship. You injected life and humanity into an experience that was at times profoundly isolating. I am grateful to be leaving Pittsburgh with a lifelong brother-in-spirit.

And finally, an enormous thank you to Daniel Neill, for providing mentorship and guidance seven years ago to a Master's student in a different department. I was still finding my footing in an unfamiliar field, and you showed me that I could forge a career out of what felt like disconnected interests in machine learning and social policy issues. Thank you for your constant dedication to the success of your students and to principled, impactful research.

Table of contents

List of figures ix

List of tables xi

Machine Learning for Decision Support in the Public Sector: Introduction xiii

1 Support Vector Subset Scan for Spatial Pattern Detection 1

1.1 Introduction 1

1.2 Support vector subset scan (SVSS) 5

1.2.1 Background: Penalized fast subset scan (PFSS) 5

1.2.2 Background: Support vector machines 7

1.2.3 SVSS optimization problem 8

1.2.4 Ranking disconnected regions 11

1.2.5 Tuning parameters 12

1.3 Evaluation and Results 14

1.3.1 Detecting letter-shaped simulated patterns 14

1.3.2 Detecting disease clusters 19

1.3.3 Detecting crime hot-spots 22

1.3.4 Detecting pothole clusters 23

1.3.5 Discussion of real-world case studies 28

1.4 Conclusions 28

Appendix 1.A Implementation Details 29

2 Assessing Risk of Opioid Misuse from Prescription Drug Monitoring Data 31

2.1 Introduction 31

2.2 Trajectory Clustering for Early Risk Assessment 33

2.2.1 Shape-based time series clustering 34

2.2.2 Early risk assessment of partial time series 37

2.3 Recurrent Generative Adversarial Networks for Semi-Supervised Learning of Opioid Misuse . . . 39

2.3.1 Adapting recurrent generative adversarial networks for semi- supervised learning . . . 41

2.3.2 Evaluation on medical time series data . . . 44

2.3.3 Evaluation on opioid time series . . . 46

2.4 Discussion . . . 59

2.5 Conclusions . . . 60

3 Policing Chronic and Temporary Hot Spots of Violent Crime: A Controlled Field Experiment 62

3.1 Introduction . . . 62

3.2 Empirical Comparison of Prediction Methods for Crime Forecasting . . . 66

3.2.1 Data . . . 66

3.2.2 Methodology . . . 67

3.2.3 Evaluation of prediction methods . . . 70

3.3 Hot Spot Field Experiment in Pittsburgh, PA . . . 73

3.3.1 Experimental design . . . 74

3.3.2 Results . . . 79

3.4 Discussion . . . 90

3.5 Conclusions . . . 94

Appendix 3.A Estimates of Costs Avoided from Observed Crime Reductions . . 95

References 97

List of figures

1.1 Refinement of the detected pattern (shown in dark blue) across iterations of SVSS. On the left, the pattern detected by the first iteration of the Fast Subset Scan includes many points outside the true affected region. In the second (middle) and third (right) iterations, points outside the SVM decision boundary are penalized and the detected pattern improves, rapidly approaching the true affected region. 11

1.2 (a) Binary classification with kernel SVM on final iteration of SVSS. SVM decision boundary shown in black. (b) Penalty surface learned on final iteration based on distance to separating hyperplane. 12

1.3 Anomalous patterns detected by two methods in the presence of both hot spots and cold spots. (a) True labels of spatial locations, with hot spots shown in red and cold spots shown in green. (b) Pattern detected by SVSS. (c) Pattern detected by the circular scan. The presence of a cold spot forces the circular scan to identify only a portion of the true affected region, while SVSS is able to closely approximate the spatial extent of the affected region. 13

1.4 Precision and recall of six scanning algorithms in simulated experiments. . 17

1.5 Overlap of six scanning algorithms in simulated experiments. 18

1.6 Clusters of West Nile Virus detected by six pattern detection algorithms in Chicago, IL. 21

1.7 Clusters of street crime detected by six pattern detection algorithms in Portland, OR. 24

1.8 Clusters of potholes detected by six pattern detection algorithms in Pittsburgh, PA. 27

2.1 Cluster centroids from shape-based clustering applied to smoothed morphine milligram equivalent time series. 35

2.2 Centroid and sample of patient time series showing characteristic shape for example cluster. 36

2.3 Proportion of patients triggering red flags by cluster for three indicators of unsafe prescriptions. 37
2.4 Diagram of early detection lead time for a single patient with increasing morphine milligram equivalents and a red flag indicating an unsafe prescription. 40
2.5 Data pipeline for training of recurrent generative adversarial network for semi-supervised classification. 44
2.6 Receiver Operating Characteristic curves showing evaluation of binary classifiers on prediction of heart rate becoming critical 46
2.7 Precision-recall curves showing evaluation of binary classifiers on prediction of heart rate becoming critical in ICU medical time series data 47
2.8 Receiver Operating Characteristic curve and Precision-Recall curve showing evaluation of binary classifiers on prediction of morphine milligram equivalents becoming critical for all patients. 53
2.9 Receiver Operating Characteristic curve and Precision-Recall curve showing evaluation of binary classifiers on prediction of morphine milligram equivalents becoming critical for patients not critical prior to prediction period 54
2.10 Comparison of prediction performance from prescription timeline features provided as conditional inputs to SS-RGAN on opioid prediction task. . . . 58
2.11 Comparison of prediction performance for two labeling procedures in training data (positive and negative labels known vs. only positive labels known). 59

3.1 Prediction tradeoff curves for 10 forecasting models on one-week-ahead predictions of Part 1 violent crime over evaluation period. 73
3.2 Composite tradeoff curve representing a combined chronic-temporary hot spot program. 76
3.3 Field trial partitions used to separate treatment from control cells each week. 78
3.4 Hot spot footprint over 12 months for chronic and temporary hot spots. . . . 89

List of tables

1.1 Summary statistics of detected patterns for simulated regions across three signal intensities. 16
1.2 Summary statistics of top West Nile Virus clusters. 20
1.3 Summary statistics of top street crime clusters. 25
1.4 Summary statistics of top pothole clusters. 26
1.5 Location-specific contributions to the score function for expectation-based statistics in the exponential family. 30
1.6 Parameter values for Support Vector Subset Scan on real world data sets. . 30

2.1 Summary of shape-based patient clusters grouped on smoothed morphine milligram equivalent time series. 38
2.2 Summary statistics of red flag lead times for high-risk alerts. 40
2.3 Summary statistics of red flag lead times for random alerts. 41
2.4 Description of four medical time series data sets. 45
2.5 Performance of binary classifiers on prediction of ICU measurements becoming critical for all patients. 47
2.6 Performance of binary classifiers on prediction of ICU measurements becoming critical for patients that were non-critical during observation period. . . 48
2.7 Description of prescription opioid time series data sets for predicting unsafe levels of morphine milligram equivalents. 50
2.8 Performance of binary classifiers on prediction of opioid time series crossing critical threshold for all patients in prescription drug monitoring program . 52
2.9 Performance of binary classifiers on prediction of opioid time series crossing critical threshold for patients not critical prior to prediction period 53
2.10 Performance of binary classifiers on prediction of opioid time series becoming critical during prediction period under two resource constraint scenarios. 56
2.11 Statistical comparison of SS-RGAN performance metrics to best performing alternative 56

2.12 Comparison of performance from prescription timeline features provided as conditional inputs to SS-RGAN on opioid prediction task under two resource constraint scenarios. 58

3.1 Counts of APRS crime incidents for all Part 1 offenses from June 1, 2011 through June 1, 2016 in Pittsburgh, PA. 67

3.2 Leading indicator variables provided as inputs to multivariate prediction models. 68

3.3 Partial area under the tradeoff curve and hot spot selection entropy for a 1% target area from 10 forecasting models over evaluation period. 74

3.4 Percent of hot spot selections occupied by frequently-selected locations. . . 76

3.5 Regression results estimating dose-dependent treatment effect on Part 1 violent crime counts. 83

3.6 Regression results estimating dose-dependent treatment effect on Part 1 property crime counts. 84

3.7 Regression results estimating non-dose-dependent treatment effect on Part 1 violent crime counts. 86

3.8 Regression results estimating non-dose-dependent treatment effect on Part 1 property crime counts. 87

3.9 Counts and standard deviations of P1V crimes in treatment and control hot spots across cell-weeks. 88

3.10 Observed differences in Part 1 violent crime counts between control treatment areas over 12 months of proactive hot spot patrols. 91

3.11 Observed differences in Part 1 property crime counts between control treatment areas over 12 months of proactive hot spot patrols. 91

3.12 Per-offense estimates of costs to society from Part 1 crime types. 95

Machine Learning for Decision Support in the Public Sector: Introduction

Recent advances in the production, collection, and curation of data have resulted in a new and complex set of resources available to guide operational decisions at all levels of government. At the same time, improvements in the accessibility of advanced statistical and computational tools have expanded the capacity of governments to conduct rigorous empirical analyses and develop decision support tools that are grounded in observed data. Machine learning has been applied with widespread success for a variety of revenue-generating tasks in the commercial sector such as demand forecasting and prediction of individual consumer preferences. This success suggests that machine learning techniques could be similarly applied in the public sector to benefit quality of life and improve social conditions, particularly in urban settings where density of available data is highest.

The movement towards data-driven and computational approaches for decision-making represents a significant paradigm shift in many policy contexts that have historically depended solely on human expertise. n the health domain, clinicians and public health practitioners traditionally rely on their own expert judgment to determine whether patients are at-risk of abusing prescribed drugs. In metropolitan police departments, command staff are routinely placed in charge of allocating patrols within a district of the city on a daily or weekly basis. Machine learning offers a set of tools to bolster human intelligence with insights summarized from large administrative data sets that would be impossible for a human to process without computational support.

Machine learning methods also represent a fundamentally different approach from causal inference studies that have traditionally been the focus of public policy research. Machine learning methods excel at finding correlations or detecting patterns in observational data. Consequently, these approaches are well-suited for settings where prediction or characterization of patterns is useful as a support for decision-making even when the underlying causal forces may not be clear. For example, public health agencies may be concerned with predicting the future spread of a disease outbreak in order to allocate healthcare resources and make safety recommendations to affected citizens. The causal mechanisms underlying the spread of the disease may be difficult to understand and untangle quickly in a crisis situation, while accurately predicting new cases at the population level can help efficiently deploy personnel and medical supplies where they are likely to be needed most.

Naive application of machine learning methods also has the potential to exacerbate the very problems they are intended to solve. Machine learning algorithms learn from patterns in historical data, and model outputs will therefore reflect any biases inherent in the data on which they were trained. If interventions based on these outputs affect how new data is generated or collected, then models can perpetuate or worsen historical biases. Shifts in the underlying data distribution, whether due to a policy intervention or some other cause, may mean that models that were once accurate and unbiased are no longer meeting these objectives.

For any machine learning system, designing a proper evaluation framework is critical for understanding whether the system will work as intended when deployed. Rigorous and periodic evaluation is particularly important when model outputs underlie decisions that affect social conditions and the well-being of human lives. Regular vetting of operational models by both (1) analysts with technical proficiency in predictive modeling and (2) subject matter experts with knowledge of underlying domain and affected populations can ensure that a deployed system remains effective over time. In many machine learning problems related to public health and safety, such as disease outbreak detection or prediction of drug misuse, proper evaluation is difficult because ground truth observations are scarce or impossible to obtain. The absence of ground truth makes evaluation of new methods particularly hard in settings where predictions or other model outputs can affect the data-generating process.

In this thesis, we present three case studies in which we propose and evaluate novel machine learning approaches to inform operational decisions in the domains of public health and safety. These studies showcase different approaches for evaluation of new machine learning methods when ground truth data is limited or not available. These studies all represent examples where public service experts identified a specific problem where administrative data was available but limited, and methodology was designed with both the problem and available data in mind.

In Chapter 1, we introduce the support vector subset scan (SVSS), a new method for detecting localized and irregularly shaped anomalous patterns in spatial data. SVSS alternately maximizes a penalized log-likelihood ratio over subsets of locations to obtain an anomalous pattern, and learns a high-dimensional decision boundary between locations included in and excluded from the anomalous subset. On each iteration of the algorithm, we assign location-specific penalties to the log-likelihood ratio based on distance to the high-dimensional decision boundary, encouraging patterns which are spatially compact but potentially highly irregular in shape. As ground truth labels are not available in many pattern detection settings, we highlight the performance of SVSS relative to competing methods for spatial cluster detection on detection of randomly generated patterns in simulated experiments. Using publicly available data sets, we also demonstrate the real-world utility of SVSS in three policy domains: disease surveillance, crime hot spot detection, and pothole cluster detection.

In Chapter 2, we develop new methods to assess risk of opioid misuse based on individual-level opioid timelines generated from prescription drug monitoring program (PDMP) data. We first introduce a shape-based clustering framework to evaluate risk of misuse in new individuals when patient outcomes are completely unknown outside of observed prescription drug histories. By identifying “red flag" behaviors which are indicative of opioid misuse, we evaluate the shape-based clustering approach on the task of early risk assessment, and find that the detection method achieves statistically and practically significant lead times with respect to red flags triggered by the PDMP. We also address the setting where labeled instances of unsafe drug use are available but sparse, developing a new method for semi-supervised learning using recurrent generative adversarial networks (RGANs) and designed to assess risk of opioid misuse in new patients based on these labels. The RGAN architecture provides a natural framework for incorporating conditional inputs to both the generator and the discriminator. We incorporate red flag indicators and shape-based cluster assignments as conditional inputs in addition to the opioid trajectories, as this additional information has the potential to improve (1) the generator’s ability to generate realistic time series conditioned on high risk, and (2) the discriminator’s performance on classifying new patients.

Lastly, in Chapter 3 we discuss findings from an empirical comparison of crime forecasting methods and a randomized field experiment evaluating a hot-spot-based predictive policing program in Pittsburgh, PA. We compare the performance of several place-based forecasting models on predicting historical crime data and select those that demonstrated high predictive accuracy and spatial dispersion of forecasted areas in Pittsburgh. We evaluate an operational hot spot program using a controlled crossover study, with areas exposed to targeted patrols changing on a weekly basis. We observe statistically and practically significant reductions in crime counts within hot spots treated with foot patrols, and find no evidence of crime displacement resulting from increased patrols to predicted hot spots. We also investigate potential harms from over-policing in hot spots, and find minimal evidence of arrests due to hot spot patrols during the field study.

Chapter 1

Support Vector Subset Scan for Spatial Pattern Detection[1]

1.1 Introduction

Detecting anomalous patterns in spatial data has applications across a wide variety of policy domains. Public health agencies may be interested in characterizing spatial regions with high prevalence of disease, indicating a possible outbreak. In large cities, police analysts are interested in detecting and characterizing flare-ups of violent crime in order to dispatch patrols effectively. Identifying spatial clusters of citizen complaints can help agencies responsible for city services such as road maintenance or sanitation to prioritize projects and efficiently address complaints. In this paper, we present an approach to address such examples where decision makers must identify spatial patterns to design and target policy interventions. In real world settings, we may expect patterns to be highly irregular in shape, as spatial clustering is often influenced by environmental or social factors such as transportation patterns, built infrastructure, land use, or natural features. Our proposed method allows for precise localization of spatial clusters regardless of shape, addressing the need for flexible detection of spatial clusters for intervention.

Anomalous patterns which are *spatially compact* are often preferable for identifying situations in need of intervention or for guiding operational decisions in policy applications. An anomalous cluster of locations is spatially compact if member locations are situated close to each other in space and non-anomalous locations are sparse within the boundaries of the cluster. Spatially compact clusters may be preferable for targeting intervention because they are more likely to correspond to a single structural cause (e.g., virus-carrying mosquitoes

[1] This chapter is based on the research paper of the same title, co-authored by Yun Ni and Daniel B. Neill.

breeding in a pool of stagnant water), or because locations in these clusters can be efficiently targeted for mitigation efforts due to their physical proximity (e.g., a cluster of potholes on a highly-trafficked road can be repaired by a single maintenance crew). Yet simpler *spatial scan* approaches such as Kulldorff (1997), which search for spatially compact clusters of a fixed shape, may fail to correctly identify the spatial extent of the cluster, and have reduced detection power when the cluster is elongated or irregular in shape.

This work builds on the *subset scan* approach to pattern detection, which finds anomalous patterns by performing a constrained scan over subsets of data points. In this framework, the *anomalousness* of fixed subsets can be evaluated using a predefined score function, such as the log-likelihood ratio statistics applied in Kulldorff (1997), Neill et al. (2005), Neill (2009), and Neill (2012a). The subset scan approach has demonstrated high power to detect both localized and global patterns, unlike 'bottom-up' approaches which identify and aggregate individual anomalies, and 'top-down' approaches which localize anomalous patterns detected in aggregated data (Neill, 2009, 2012a). Outlier detection methods such as one-class SVM (Schölkopf et al., 2001) are likely to pick out individually anomalous data records with high counts (often due to chance) and thus fail to detect the regions of interest, while density-based clustering methods such as DBSCAN (Ester et al., 1996) can find anomalous regions but are dramatically outperformed by our proposed method (as shown in Section 1.3.1).

Subset scanning poses a significant computational challenge, as there exist 2^N possible subsets to consider when searching for the most anomalous subset for a data set containing N elements. Several approaches have been proposed to reduce the computation needed to search over the entire data set. One approach is to restrict the search space by considering only regions of a specific shape, such as circles (Kulldorff, 1997), ellipses (Kulldorff et al., 2006), or rectangles (Neill and Moore, 2004). Other approaches reduce the number of subsets under consideration by enforcing connectivity constraints between elements included in a subset (Costa and Kulldorff, 2014; Duczmal and Assuncao, 2004; Duczmal et al., 2007; Patil and Taillie, 2004; Speakman et al., 2015; Takahashi et al., 2008; Yiannakoulias et al., 2007). These methods enable efficient computation of anomalous patterns but sacrifice both detection power and spatial accuracy in comparison to unconstrained methods which do not restrict the search space (Neill, 2012a).

One alternative to subset scanning for anomalous pattern detection is to fully model dependence across spatially-distributed point observations using a geostatistical model, such as a spatial generalized linear mixed model (SGLMM). Introduced in Diggle et al. (2002), SGLMMs are a form of generalized linear model in which spatial dependence is modeled with Gaussian processes across the spatial extent. While SGLMMs represent a powerful tool for modeling spatial data, standard sample-based inference approaches on these models are computationally expensive and slow to converge (Haran, 2011). Further, SGLMMs do not define a decision boundary around anomalous spatial regions, which is a practically useful output for characterizing the extent of an affected region and thus enabling targeted interventions.

Identifying patterns with arbitrary shape in the subset scanning framework is non-trivial given the high number of patterns to consider. Methods that search over subsets with a fixed geometric shape or impose connectivity constraints are not likely to accurately characterize affected regions with irregular shapes or multiple disconnected components. Connectivity may also be difficult to determine in contexts where no inherent graph structure is obvious. Underconstrained patterns which are too disconnected or sparse may be similarly unrealistic. Thus, recent developments in spatial scanning have focused on encouraging patterns which are spatially compact while still allowing for detection of irregular shapes. Duczmal et al. (2006) and Yiannakoulias et al. (2007) propose penalized score functions that discourage highly irregular shapes based on measures of non-compactness or non-connectivity, but do not provide a statistical framework for interpreting the penalized versions of the score functions. Other approaches have applied multi-objective optimization algorithms to simultaneously maximize a score function and minimize a geometric penalty function (Cancado et al., 2010; Duarte et al., 2010; Duczmal et al., 2012; Moreira et al., 2015). These multi-objective methods result in a set of non-dominated candidate patterns which must then be ranked by a single objective function to obtain the most anomalous pattern. This ranking step presents both computational and theoretical difficulties, as the set of candidates may be large and the desired tradeoff between multiple objectives could be ill-defined across candidate patterns.

Neill (2012a) presents the fast subset scan (FSS), demonstrating that the most anomalous unconstrained subset across an entire data set can be found both efficiently and exactly for a family of score functions satisfying the Linear Time Subset Scan property. In practice, the FSS framework may detect patterns which are spread across the spatial extent of the study area and sparsely distributed among non-anomalous points. Several approaches to imposing hard spatial constraints on FSS have been proposed, such as searching only over local neighborhoods consisting of each location and its $k-1$ neighbors (Neill, 2012a), or searching over locations connected by an underlying graph structure (Speakman et al., 2015). Speakman et al. (2016) provides a structured approach to incorporating soft constraints into the FSS framework with the penalized fast subset scan (PFSS), showing that one can apply additive penalties and still maximize the penalized score function efficiently and exactly. While PFSS gives us a framework for incorporating soft constraints, the question of how to define penalty terms to encourage spatial compactness in detected patterns remains open.

In this work, we present the support vector subset scan (SVSS), which detects anomalous patterns in spatial data that are spatially coherent but potentially highly irregular in shape. SVSS integrates PFSS with a kernel support vector machine (SVM) to encourage compact subsets of locations. The SVM provides a natural solution to the problem of specifying element-specific penalties for PFSS such that detected patterns are geometrically compact but unconstrained in size, shape, or connectivity. SVSS benefits from the ability of PFSS to detect subtle but significant anomalous patterns, while leveraging the SVM to identify coherent spatial regions with a high density of anomalous points. This novel combination of two proven methods results in a new approach for anomalous pattern detection that outperforms each of the individual component methods. SVSS imposes soft constraints on FSS, which encourage spatial compactness at the cost of a lower anomalousness score. In comparison to the sparse patterns returned by FSS, SVSS finds compact patterns that are more suitable for targeted intervention.

The SVSS algorithm proceeds iteratively, alternating between efficiently maximizing a penalized log-likelihood ratio (LLR) over subsets of locations, and learning a high-dimensional decision boundary between locations included in and excluded from the anomalous subset. Location-specific penalties are computed according to distance to the decision boundary and added to the LLR score function, resulting in anomalous patterns which are spatially compact and irregular in shape. This iterative method is guaranteed to converge to a locally-optimal subset with respect to the biconvex SVSS objective function (Gorski et al., 2007). We apply multiple random restarts to approach the global optimum of the SVSS objective.

In Section 1.2, we provide the statistical background motivating SVSS, then define the SVSS optimization problem and algorithm. In Section 1.3.1, we evaluate SVSS on the task of detecting letter-shaped anomalous patterns in simulated data and find that it significantly outperforms competing methods at finding patterns which closely approximate the true affected region. Using publicly available data sets, we demonstrate the method in three real world contexts where spatial pattern detection is useful for guiding operations and policy decisions in Sections 1.3.2-1.3.3. In the domain of disease surveillance, we apply SVSS to West Nile Virus test results to identify disease clusters throughout the city of Chicago, IL. For crime surveillance, we apply SVSS to characterize hotspots of street crime in Portland OR. Finally, we apply SVSS to detect clusters of potholes in Pittsburgh, PA, demonstrating the utility of the method for city services and management. We end with concluding remarks in Section 1.4.

1.2 Support vector subset scan (SVSS)

In this section, we describe a parametric scan statistic approach for spatial pattern detection under weak distributional assumptions.

1.2.1 Background: Penalized fast subset scan (PFSS)

Consider the setting in which data set D includes a set of spatial coordinates $\mathbf{x}_i$ for locations $(i = 1,...,N)$. Let $\alpha \in \{0,1\}^N$ be a vector specifying a subset of locations, with $\alpha_i = 1$ if location i is included in the subset and $\alpha_i = 0$ otherwise. Maximizing a score function over subsets is performed by searching over values of the vector α and maximizing some score function $F(\alpha)$. Neill et al. (2005) proposes a class of score functions called *expectation-based scan statistics*, in which data set D also includes observed values (or “counts”) **c** and expected values (or “baselines”) **b** of a random field indexed at locations. These location-specific observations and expected values provide the basis for defining $F(\alpha)$ and determining whether a subset is anomalous.

Let $H_1(\alpha)$ be an alternative hypothesis that assumes an event occurring in the subset defined by α causing increased values at those locations, and let H_0 be the null hypothesis that assumes no event occurring in the data set (or equivalently, that $\alpha_i = 0$ for all i). Following Kulldorff (1997) and Neill et al. (2005), we define our score function as a log-likelihood ratio (LLR) statistic $F(\alpha) = \log(Pr(D|H_1(\alpha))/Pr(D|H_0))$. The expectation-based scan statistics assume that under alternative hypothesis $H_1(\alpha)$, values c_i are drawn with mean qb_i inside of the region defined by α and mean b_i outside of that region for some multiplicative constant factor $q > 1$ known as *relative risk*. The expectation-based scan statistic is formulated as

$$F(\alpha) = \max_{q>1} \sum_{i=1}^{N} \alpha_i \left[\log Pr(c_i|qb_i) - \log Pr(c_i|b_i)\right] \tag{1.1}$$

Speakman et al. (2016) introduce the Penalized Fast Subset Scan (PFSS), observing that for a fixed value of relative risk q, the LLR for the exponential family of expectation-based scan statistics can be expressed as an additive set function over all locations included in a subset:

$$F(\alpha|q) = \sum_{i=1}^{N} \alpha_i \lambda_i(q)$$

$$F(\alpha) = \max_{q>1} F(\alpha|q) = \max_{q>1} \sum_{i=1}^{N} \alpha_i \lambda_i(q)$$

where λ_i terms depend only on observed values c_i, baselines b_i, and relative risk q. Because $\lambda_i(q)$ expressions can be derived for a variety of expectation-based scan statistics, this additive score function is flexible in the underlying data distribution, making the assumption that observed values c_i are drawn from a distribution in the exponential family with finite first moments. Further, the additive property of the score function enables addition of location-specific penalty terms to the LLR, denoted as Δ_i. Each Δ_i can be interpreted as the prior log-odds that location i is included in the affected subset. We express the penalized score function as

$$F_{pen}(\alpha) = \max_{q>1} \sum_{i=1}^{N} \alpha_i \big(\lambda_i(q) + \Delta_i\big) \tag{1.2}$$

where $\lambda_i(q) + \Delta_i$ represents the total contribution of location i to the score function. Conditioning on a fixed value of relative risk q, the penalized score function can be optimized over all subsets by including all and only those locations with a positive total contribution $\lambda_i(q) + \Delta_i$. The score functions for expectation-based scan statistics can be optimized efficiently by considering at most $2N$ distinct values of q (Speakman et al., 2016).

Specifically, Speakman et al. (2016) show that $\gamma_i(q) = 0$ for at most two values of q, and we can compute a q_i^{min} and q_i^{max} for each location i such that $\gamma_i(q_i^{min}) = \gamma_i(q_i^{max}) = 0$ and $\gamma_i(q) > 0$ for all $q_i^{min} < q < q_i^{max}$. We sort the set $\{q_1^{min}, ..., q_N^{min}, q_1^{max}, ..., q_N^{max}\}$, remove any duplicate values of q, then consider the disjoint intervals formed by consecutive values of the sorted q. For each interval, we find the subset of locations with positive $\gamma_i(q)$ over that entire interval, then evaluate the score of that subset using the maximum likelihood estimate for q given the subset. We then can optimize the penalized or unpenalized score function by considering one value of q per interval,

PFSS thus provides an extremely flexible and computationally efficient framework for scanning over subsets and incorporating soft constraints to encourage patterns with desirable attributes. Yet the PFSS framework is not sufficient to obtain spatial coherence or compactness in detected patterns. Because the penalty terms Δ_i must be decided for each element before optimizing the penalized LLR, there is no natural way of assigning element-wise bonuses or penalties if we do not already know where the anomalous subset is likely to be. PFSS is also limited because the penalties must be location-specific, precluding application of an arbitrary prior over subsets to encourage more coherent regions. If the penalties depended on other locations in the subset, we would not be able to perform the scan efficiently and would have to exhaustively enumerate subsets. Thus, we must incorporate additional tools in order to specify location-specific Δ_i terms which promote compactness.

1.2.2 Background: Support vector machines

To formulate Δ_i terms for the PFSS framework, we turn to the support vector machine (SVM), a popular algorithm for binary classification first proposed by Cortes and Vapnik (1995). The SVM is trained on data elements consisting of feature vectors $\mathbf{x}_i$ and positive or negative class labels y_i, finding the separating hyperplane between classes which maximizes the margin between classes, or the distance between the hyperplane and the nearest data point on either side.

A soft-margin SVM introduces slack variables ξ_i and tuning parameter C to address the case when the two classes are not linearly separable. Learning an SVM is formulated as the following optimization problem, where weight vector $\mathbf{w}$ and intercept term w_0 define the separating hyperplane and ϕ is a nonlinear transformation which maps $\mathbf{x}$ to a high-dimensional feature space and allows a nonlinear decision boundary in the original space:

$$\begin{aligned} \min_{\xi,\mathbf{w},w_0} &\ \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to } &\ \xi_i \geq 0, \forall i = 1,...,N \\ &\ y_i(\mathbf{w}\cdot\phi(\mathbf{x}_i) - w_0) \geq 1 - \xi_i, \forall i = 1,...,N \end{aligned}$$

This problem is typically optimized through its Lagrangian dual using algorithms for efficiently solving quadratic programs. The dual formulation allows us to avoid the costly computation of ϕ by defining an easy-to-compute kernel function K such that $K(\mathbf{x}_i,\mathbf{x}_j) = \langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle$. The SVM dual is then expressed as

$$\begin{aligned} \max_{\mu} &\ \sum_{i=1}^{N}\mu_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mu_i\mu_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) \\ \text{subject to } &\ \sum_{i=1}^{N}\mu_i y_i = 0 \\ &\ 0 \leq \mu_i \leq C, \forall i = 1,...,N \end{aligned}$$

The distance from any data element $\mathbf{x}_j$ to the hyperplane in high dimensional space is expressed as $\mathbf{w}\cdot\phi(\mathbf{x}_j) - b$. Although we cannot evaluate the weight vector $\mathbf{w}$ after solving the dual SVM problem, we can easily compute the distance from $\mathbf{x}_j$ to the hyperplane defined by $\mathbf{w}$ and b in high dimensional space:

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^{N} \mu_i y_i \phi(\mathbf{x}_j) \\ \mathbf{w} \cdot \phi(\mathbf{x}_j) - b &= \sum_{i=1}^{N} \mu_i y_i \phi(\mathbf{x}_i) \phi(\mathbf{x}_j) - b \\ &= \sum_{i=1}^{N} \mu_i y_i K(\mathbf{x}_i, \mathbf{x}_j) - b \end{aligned}$$

The SVM with Gaussian kernel results in a decision boundary with a potentially highly irregular shape and multiple disconnected components, tending to demarcate high-density regions of each class in the original feature space. As a supervised method, the SVM requires locations to be labeled as belonging to one class or the other. In the context of spatial pattern detection, it is not obvious how to assign labels for the SVM. One straightforward approach is to apply a threshold to some function of location-specific counts and baselines to assign class labels (included for comparison in Section 1.3.1). However, particularly for subtle signals, a high proportion of points will initially be mislabeled, leading to an extremely noisy classification problem and resulting poor performance. These considerations motivate our SVSS approach which alternates between PFSS and SVM optimization steps: we iteratively pick good thresholds for class labeling using the penalized score function from PFSS, then specify location-specific penalty terms as distances to a SVM hyperplane within the PFSS framework. The resulting patterns are spatially coherent but irregular in shape due to the nonlinear decision boundary given by the kernel SVM.

1.2.3 SVSS optimization problem

With the kernel SVM, we now have the tools necessary to specify Δ_i penalty terms for PFSS in the context of unsupervised subset scanning. Given a fixed α which defines a subset of locations, let $y_i = 2\alpha_i - 1$ for all locations. Thus, our class labels $y_i \in \{-1, 1\}$ represent inclusion or exclusion from the given subset defined by α, so that the SVM learns a decision boundary to separate included from excluded locations. We formulate SVSS as a modified version of the SVM optimization problem, while also minimizing over subsets α and including the unpenalized LLR score function $F(\alpha)$ as an additional regularization term. Alternatively, we could view this as a maximization of the penalized scan statistic, optimizing $F(\alpha)$ with penalties from the SVM slack variables and the width of the margin. We now have two tuning parameters C_0 and C_1, controlling the relative importance of these three factors.

$$\begin{aligned}
\min_{\alpha,\xi,\mathbf{w},w_0} \ & \frac{1}{2}||\mathbf{w}||^2 + C_0 \sum_{i=1}^{N} \xi_i - C_1 F(\alpha) \\
\text{s.t. } & \alpha_i \in \{0,1\}, \forall i = 1,...,N \\
& \xi_i \geq 0, \forall i = 1,...,N \\
& (2\alpha_i - 1)(\mathbf{w} \cdot \phi(\mathbf{x}_i) - w_0) \geq 1 - \xi_i, \forall i = 1,...,N
\end{aligned}$$

Equivalently, we can express slack variables as a function of α_i to obtain the SVSS optimization problem.

$$\begin{aligned}
\min_{\alpha,\xi,\mathbf{w},w_0} \ & \frac{1}{2}||\mathbf{w}||^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i) - C_1 F(\alpha) \\
\text{s.t. } & \alpha_i \in \{0,1\}, \forall i = 1,...,N \\
\xi_i(\alpha_i) = & \max(0, 1 - (2\alpha_i - 1)(\mathbf{w} \cdot \phi(\mathbf{x}_i) - w_0))
\end{aligned} \tag{1.3}$$

This optimization problem is not convex, making computation of the global optimum non-trivial. We optimize the SVSS objective by alternately (1) fixing the anomalous subset α and optimizing $\mathbf{w}$ and w_0 by training the SVM, then (2) fixing $\mathbf{w}$ and w_0 and learning an optimal subset α through a search over subsets to maximize the score function. This alternating approach to minimization is guaranteed to give a convergent sequence for the biconvex SVSS objective function, but does not necessarily find the global optimum (Gorski et al., 2007). Thus, we use multiple restarts to randomly initialize the Δ_i penalty terms and take the optimal subset across all restarts (as measured by the combined objective) as our anomalous pattern. SVSS iterates over two computationally efficient algorithms (PFSS and SVM). Each iteration of PFSS is an $O(N \log N)$ operation. Computational complexity of the RBF-kernel SVM scales between $O(N^2)$ and $O(N^3)$ and is dependent on the specific data set and amount of regularization applied (Bottou and Lin, 2007). In practice, across a variety of data sets, only a small number of iterations are needed for the algorithm to converge to a local optimum. Algorithm 1 outlines the SVSS algorithm using T_{max} random restarts.

For a given hyperplane specified by a fixed $\mathbf{w}$ and w_0, we can optimize the SVSS objective using the PFSS algorithm. Optimizing the SVSS objective for fixed $\mathbf{w}$ and w_0 is equivalent to

$$\operatorname*{argmax}_{\alpha} F(\alpha) - \frac{C_0}{C_1} \sum_{i=1}^{N} \xi_i(\alpha_i)$$

Algorithm 1 Support Vector Subset Scan

```
procedure SVSS($\mathbf{c}, \mathbf{b}, \mathbf{x}, T_{max}, C_0, C_1$)        ▷ Values c, expectations b,
    $min\_score \leftarrow \infty$ and coordinates x
    for $t := 1$ to $T_{max}$ do                                        ▷ $T_{max}$ random restarts
        $\xi_i(\alpha_i) \leftarrow \text{Uniform}(-C_0, C_0), \forall i = 1, ..., N$
        while $\alpha$ is changing do
            $\alpha \leftarrow \underset{\alpha}{\text{argmax}}\, F(\alpha) - (C_0/C_1) \sum_{i=1}^{N} \xi_i(\alpha_i)$        ▷ Optimize over $\alpha$
            $\xi, \mathbf{w}, w_0 \leftarrow \underset{\xi, \mathbf{w}, w_0}{\text{argmin}} \frac{1}{2}||\mathbf{w}||^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i)$        ▷ Optimize over $\mathbf{w}, w_0$
        end while
        $score \leftarrow \frac{1}{2}||\mathbf{w}||^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i) - C_1 F(\alpha)$
        if $score < min\_score$ then
            $min\_score \leftarrow score$
            $\alpha_{min} \leftarrow \alpha$
        end if
    end for
    return $\alpha_{min}$
end procedure
```

where

$$\xi_i(\alpha_i) = \begin{cases} \max(0, 1 - \mathbf{w} \cdot \phi(\mathbf{x}_i) + w_0), & 2\alpha_i - 1 = +1) \\ \max(0, 1 + \mathbf{w} \cdot \phi(\mathbf{x}_i) - w_0), & 2\alpha_i - 1 = -1) \end{cases}$$

Without changing the optimal solution, we can solve a modified problem with penalty terms that are non-zero only for points included in the subset defined by a fixed α:

$$\underset{\alpha}{\text{argmax}}\, F(\alpha) - \frac{C_0}{C_1} \sum_{i=1}^{N} \alpha_i \Delta_i \tag{1.4}$$

where

$$\begin{aligned} \Delta_i =& \max(0, 1 - \mathbf{w} \cdot \phi(\mathbf{x}_i) + w_0) - \max(0, 1 + \mathbf{w} \cdot \phi(\mathbf{x}_i) - w_0) \\ =& \begin{cases} \mathbf{w} \cdot \phi(\mathbf{x}_i) - w_0 + 1, & \mathbf{w} \cdot \phi(\mathbf{x}_i) - w_0 \geq 1 \\ 2(\mathbf{w} \cdot \phi(\mathbf{x}_i) - = w_0), & \mathbf{w} \cdot \phi(\mathbf{x}_i) - w_0 \in (-1, 1) \\ \mathbf{w} \cdot \phi(\mathbf{x}_i) - w_0 - 1, & \mathbf{w} \cdot \phi(\mathbf{x}_i) - w_0 \leq -1 \end{cases} \\ =& [\mathbf{w} \cdot \phi(\mathbf{x}_i) - w_0 > -1](\mathbf{w} \cdot \phi(\mathbf{x}_i) - w_0 + 1) + \\ & [\mathbf{w} \cdot \phi(\mathbf{x}_i) - w_0 < 1](\mathbf{w} \cdot \phi(\mathbf{x}_i) - w_0 - 1) \end{aligned}$$

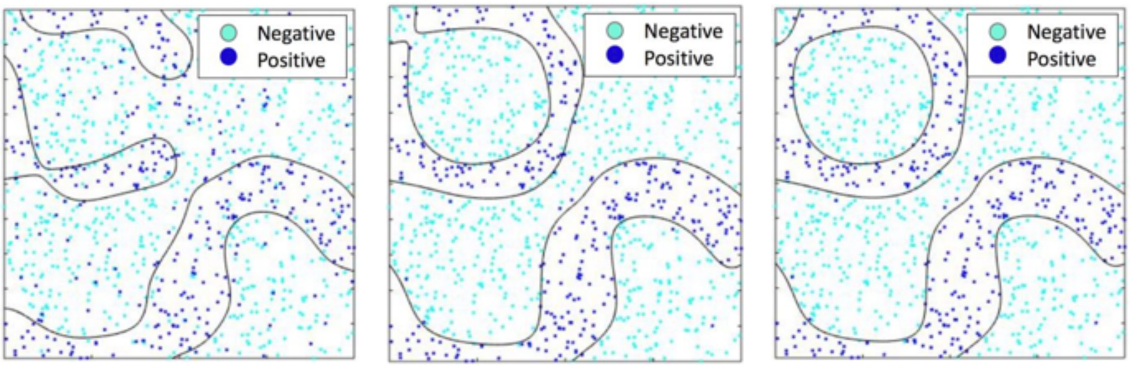


Fig. 1.1 Refinement of the detected pattern (shown in dark blue) across iterations of SVSS. On the left, the pattern detected by the first iteration of the Fast Subset Scan includes many points outside the true affected region. In the second (middle) and third (right) iterations, points outside the SVM decision boundary are penalized and the detected pattern improves, rapidly approaching the true affected region.

Because each penalty term Δ_i depends only on spatial coordinates from location i and not other locations, we can efficiently optimize (1.4) using the PFSS algorithm. Specifically, for a fixed relative risk q, we include only those locations with a positive total contribution to the objective function, and we maximize the objective over linearly many values of q as discussed in Section 1.2.1.

Refinement of the detected pattern across iterations of the SVSS algorithm is demonstrated in Figure 1.1. As the algorithm progresses, points outside of the SVM decision boundary are penalized, resulting in patterns with spatial coherence. Figure 1.2 shows the values of the penalty term Δ generated by SVSS on the final iteration of the algorithm across the spatial region surrounding a simulated anomalous pattern. Figure 1.3 shows the patterns returned by SVSS and circular scanning windows in the presence of both hot spots (with increased counts relative to baseline) and cold spots (with decreased counts). The presence of a cold spot contained *within* the hot spot does not affect the ability of SVSS to detect the surrounding hot spot, but the cold spot forces the circular scan to identify only a small portion of the true hot spot. While this work focuses on applying SVSS for detecting hot spots with elevated values, the method can also be applied for cold spot detection with a minor change to the log-likelihood ratio specification.

1.2.4 Ranking disconnected regions

As previously noted, the decision boundary learned by an SVM may result in multiple disconnected components, allowing the SVSS algorithm to return anomalous patterns with multiple disjoint regions. In the subset scanning framework, it is reasonable to consider

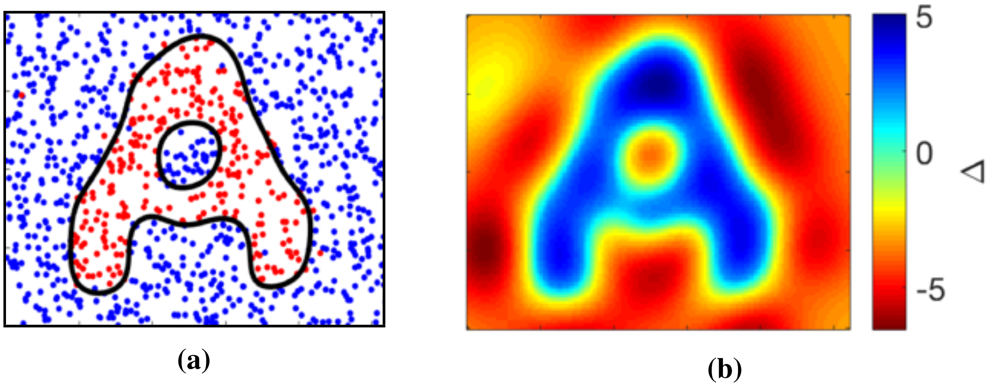


Fig. 1.2 (a) Binary classification with kernel SVM on final iteration of SVSS. SVM decision boundary shown in black. (b) Penalty surface learned on final iteration based on distance to separating hyperplane.

these regions as a single anomalous pattern, because the problem formulation assumes a constant relative risk q across the entire pattern. However, for some applications, we may seek to further search over components of our pattern to find the most anomalous component across disconnected regions of our pattern. To accomplish this, we optimize the penalized LLR $F_{pen}(\alpha)$ over components of the final SVM decision boundary as a post-processing step. We can also consider the convex hull of a connected component in the grid in order to evaluate geometric characteristics of the region, such as the compactness measure discussed in Section 1.3. We demonstrate this ranking approach to connected components in Sections 1.3.2 and 1.3.3.

After running SVSS to obtain an optimal pattern, we take the following approach to find disjoint components of the pattern. We first overlay a grid of equally-spaced points over our spatial extent, and classify points using the SVM classifier learned on the final iteration of SVSS. We then find the connected components within the grid belonging to the positive (anomalous) class. Locations in the optimal pattern detected by SVSS are assigned to the connected component of the nearest point from the grid overlay. The resolution of the grid overlay is selected such that any disjoint components separated by less distance than the grid resolution can be practically considered a single pattern component.

1.2.5 Tuning parameters

The SVSS optimization problem includes several parameters which must be selected ahead of time. C_0 is a regularization parameter which controls the impact of misclassification on the overall objective function during the SVM step. With higher C_0, the SVM learns a more complex decision boundary to avoid misclassifications, giving patterns that are more

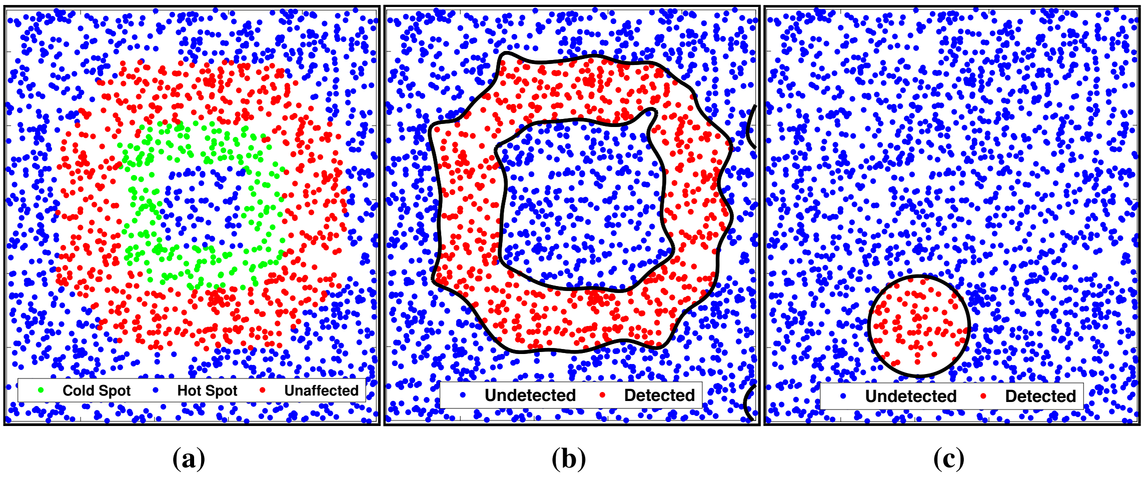


Fig. 1.3 Anomalous patterns detected by two methods in the presence of both hot spots and cold spots. (a) True labels of spatial locations, with hot spots shown in red and cold spots shown in green. (b) Pattern detected by SVSS. (c) Pattern detected by the circular scan. The presence of a cold spot forces the circular scan to identify only a portion of the true affected region, while SVSS is able to closely approximate the spatial extent of the affected region.

irregular in shape. Similarly, the kernel function chosen for the SVM step may have a tuning parameter which affects the shape of the decision boundary such as the bandwidth parameter for a Gaussian kernel. C_1 should be chosen in relation to the value of C_0, as the ratio $\frac{C_0}{C_1}$ controls the scale of the penalty terms relative to the LLR in the PFSS.

In practice, the choice of parameters can have a significant impact on the shape and size of the patterns returned by SVSS. High values of C_0 and low values of Gaussian kernel bandwidth parameter can result in highly irregular and elongated patterns that are unrealistic and likely capture noise in the data rather than true anomalous patterns. A procedure is needed for selecting parameter values that avoids overfitting to noise in the data while still enabling SVSS to capture truly irregular affected regions. To tune the SVSS parameters, we perform 10-fold cross-validation and choose the set of parameters that results in the highest average anomalousness score on held-out data. Specifically, we choose the parameters which maximize the average *unpenalized* LLR for points classified as anomalous by the SVM trained in the final iteration of SVSS, since LLR on the held-out data corresponds to how well the identified SVSS decision boundary (for particular parameter settings) captures the latent risk surface. By optimizing on multiple held-out data folds, we prevent overfitting to noise and the complexity of the resulting patterns reflects the true underlying spatial distribution. For a fixed data set, we observe minimal variation in the optimal values of C_0 and C_1 selected across multiple random restarts, providing evidence that the optimal parameter choices are a function of patterns in the true underlying data distribution. We therefore reduce computation time by completing the parameter tuning step a single time rather than tuning parameters separately for each restart.

1.3 Evaluation and Results

Evaluation of pattern detection methods on real world data can be difficult, given that we often do not know the true affected region that we hope to capture with the detected patterns. We evaluate the performance of SVSS and other pattern detection methods on simulated experiments where ground truth is known, then demonstrate SVSS in three real world pattern detection settings using real data.

1.3.1 Detecting letter-shaped simulated patterns

To evaluate our method in an experimental setting, we generate patterns of varying size, shape, and intensity in simulated data. On each run of the simulation, we draw 2000 locations uniformly at random across a rectangular study area. To generate patterns of irregular shape, we insert an *affected region* within the study area with shape matching a letter from the English alphabet. Each location has an observed count c_i drawn from the Poisson distribution, with counts outside the affected region drawn $c_i \sim Poisson(100)$ and counts inside the region drawn $c_i \sim Poisson(100 + intensity)$. Each location has a fixed baseline $b_i = 100$. We report average performance across 1300 simulations (50 simulations for each of the 26 letters in the uppercase English alphabet) for each pattern size under consideration, ranging from 1% to 20% of the study area. We tune parameters for SVSS using the cross-validation procedure outlined in Section 1.2.5. For all data sets considered, we observe minimal variation in the subsets returned and the LLR of optimal subsets across multiple, randomly initialized restarts for SVSS, and we therefore fix the number of restarts at 10 for all experiments. We compare the performance of SVSS with five other methods for spatial pattern detection: the *circular scan statistic* (Kulldorff, 1997), *upper level set scan statistic (ULS)* (Patil and Taillie, 2004), the *fast subset scan* (FSS) (Neill, 2012a), *DBSCAN* with thresholding (Ester et al., 1996), and the *Kernel Support Vector Machine* (kSVM) with thresholding. Implementation details for these methods are included in Appendix 1.A. All experiments were run in MATLAB R2016a.

For the circular scan, ULS, FSS, and SVSS, we apply the expectation-based Poisson scan statistic to formulate the LLR. We evaluate the performance of all methods at capturing the true affected region with the top pattern returned using *precision* and *recall.* Precision is defined as the proportion of points in the top pattern that lie in the true affected region, or the number of true positives divided by the number of true and false positives. Recall (or true positive rate) is defined as the proportion of points in the true affected region that are included in the top pattern, or the number of true positives divided by the number of true positives and false negatives.

We first report summary statistics from the six pattern detection methods on individual samples from three different signal intensities. For a pattern S returned by one of the scanning algorithms, we report the number of locations included in the pattern (n_S) and the unpenalized log-likelihood ratio (LLR) and the maximum likelihood estimate of the relative risk q_{MLE} as two measures of anomalousness. We also adopt a measure of geometric compactness presented in Duczmal et al. (2006). For a zone z, the geometric compactness $K(z)$ is defined as the area of z divided by the area of the circle with the same perimeter as the convex hull of z. This measure of compactness is highest for circles ($K(z) = 1$), and low for shapes that are highly irregular in shape. $K(z)$ depends only on the shape of the zone but is independent of its size. K is highest for circles ($K = 1$) and low for shapes that are highly irregular in shape. We only report K for the circular scan and SVSS, as this measure evaluates compactness of shapes and cannot be computed over sets of points returned by FSS and ULS. We also introduce an alternate measure of compactness, K_{point}, which operates on sets of points and allows us to compare compactness across all six detection methods. To compute K_{point}, we first find the Voronoi polygons for all spatial locations in the data set, then clip these polygons to the convex hull of the pattern under evaluation and dissolve any shared edges between polygons belonging to points in the pattern. K_{point} is then computed as the area of the polygons covering our pattern divided by the area of the circle with the same total perimeter as these polygons, giving a point-based measure analogous to K. K_{point} is close to 1 for patterns that are roughly circular in shape and not dispersed among points excluded from the pattern. Patterns which are elongated or spread out among excluded points have a low compactness as measured by K_{point}.

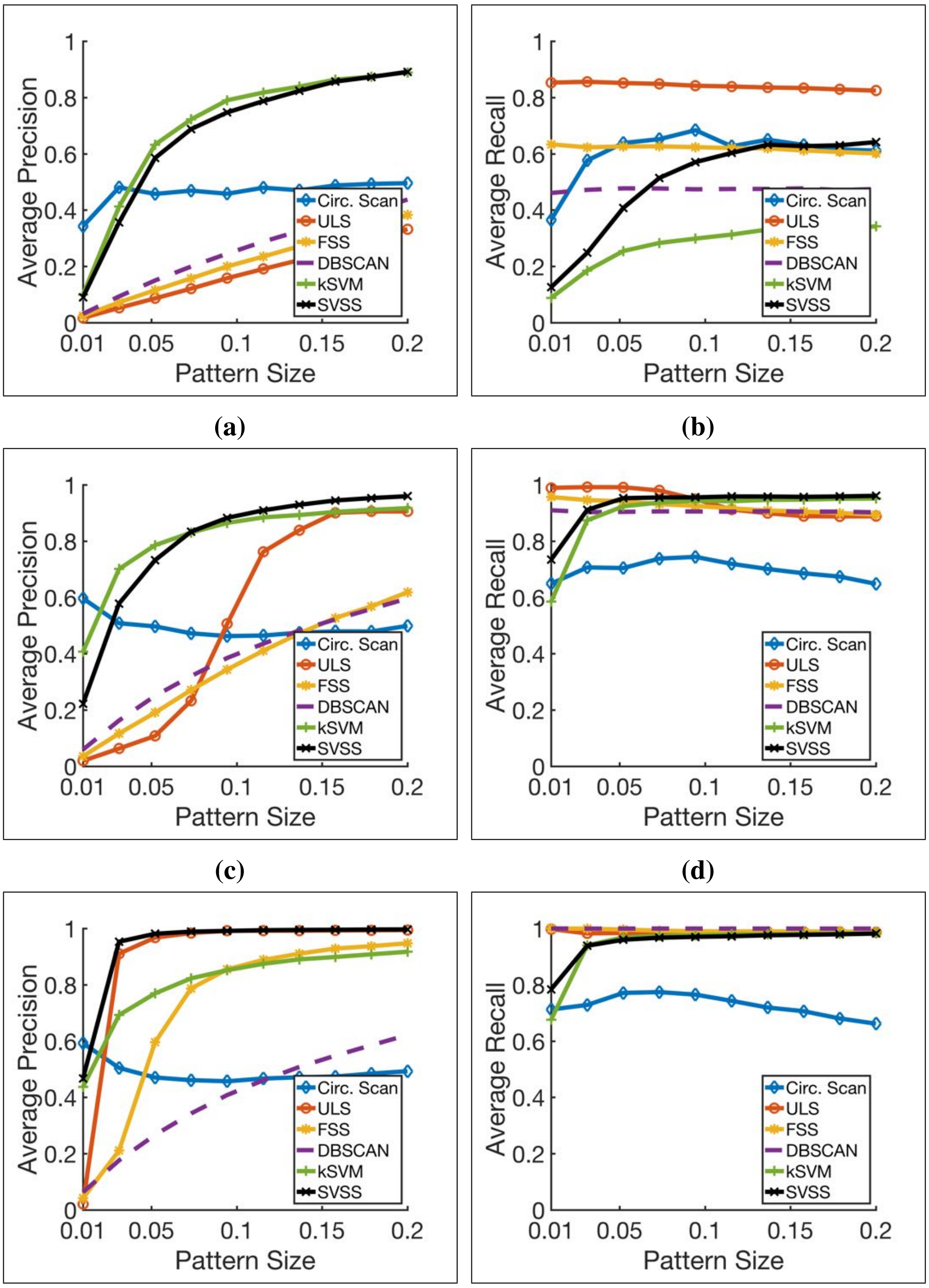
Pattern characteristics of the top patterns returned by all methods for three samples are reported in Table 1.1. Across samples of varying pattern intensity, SVSS scores highest on compactness metrics K and K_{point}. While methods such as FSS and ULS tend to find patterns with higher LLR and q_{MLE}, these methods score poorly on compactness, indicating that the detected patterns are sparse and may be sensitive to observations that are elevated due to random noise. With respect to computation time, SVSS is faster than the circular scan and ULS across all pattern intensities, but is slower than DBSCAN, FSS, and kSVM, indicating that the ability to detect spatially compact patterns comes at the expense of an increase in computation time relative to less-constrained detection methods.

Table 1.1 Summary statistics of detected patterns for simulated regions across three signal intensities: affected regions have 10%, 25%, and 50% increase in expected counts relative to unaffected regions. Statistics are shown for individual samples from each intensity with affected region shaped like the letter "A".

	n_S			CPU Time (sec)		
	10%	**25%**	**50%**	**10%**	**25%**	**50%**
Circular scan	302	402	496	17.6	17.9	18.2
ULS	744	320	373	55.6	53.7	52.1
FSS	564	511	392	0.29	0.22	0.23
DBSCAN	78	259	377	0.21	0.24	0.2
kSVM	413	550	614	0.55	0.52	0.45
SVSS	236	356	372	15.8	13.3	10.6

	LLR			q_{MLE}		
	10%	**25%**	**50%**	**10%**	**25%**	**50%**
Circular scan	74.8	398.8	1579.8	1.07	1.14	1.26
ULS	433.2	1157.1	4017.1	1.09	1.28	1.50
FSS	616.8	1346.6	4041.1	1.18	1.24	1.49
DBSCAN	252.7	1092.6	3734.6	1.27	1.30	1.48
kSVM	540.0	1301.4	3626.8	1.17	1.22	1.36
SVSS	147.3	1067.9	3925.4	1.11	1.25	1.49

	K			K_{point}		
	10%	**25%**	**50%**	**10%**	**25%**	**50%**
Circular scan	1.00	1.00	1.00	0.83	0.86	0.88
ULS	-	-	-	0.01	0.03	0.12
FSS	-	-	-	0.01	0.01	0.08
DBSCAN	-	-	-	0.01	0.01	0.06
kSVM	-	-	-	0.01	0.01	0.01
SVSS	0.45	0.48	0.48	0.17	0.13	0.15



(a)

(b)

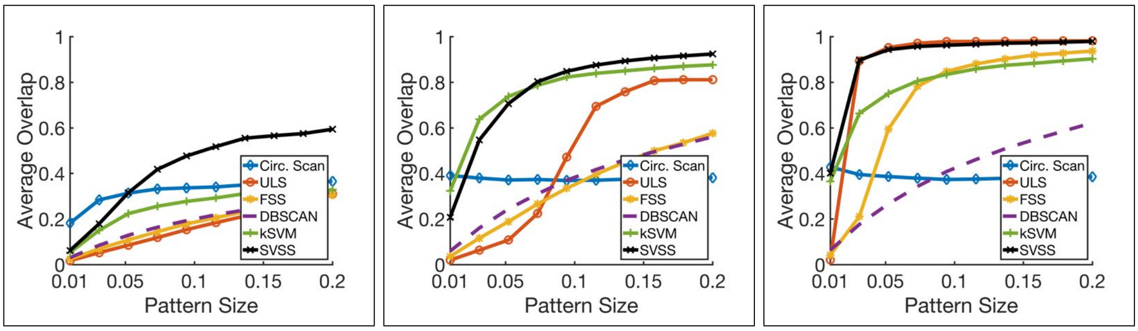
(c)

(d)

(e)

(f)

Fig. 1.4 Average precision (left) and recall (right) of six scanning algorithms on detection of letter-shaped patterns of varying size (proportion of study area) in simulated data. Results are shown for three different signal intensities: points in the affected region have a 10% (top), 25% (middle), and 50% (bottom) increase in expected counts.



(a) (b) (c)

Fig. 1.5 Average overlap of six scanning algorithms on detection of letter-shaped patterns of varying size (proportion of study area) in simulated data. Results are shown for three different signal intensities: points in the affected region have a 10% (left), 25% (middle), and 50% (right) increase in expected counts.

For the circular scan, ULS, FSS, and SVSS, we apply the expectation-based Poisson scan statistic to formulate the LLR. We evaluate the performance of all methods at capturing the true affected region with the top pattern returned using *precision*, *recall*, and *overlap*. Precision is defined as the proportion of points in the top pattern that lie in the true affected region, or the number of true positives divided by the number of true and false positives. Recall (or true positive rate) is defined as the proportion of points in the true affected region that are included in the top pattern, or the number of true positives divided by the number of true positives and false negatives. Overlap is a measure of similarity between the top pattern and the true affected region, defined as the number of points in the intersection divided by the number of points in the union of the pattern and true affected points.

Precision and recall for patterns with three different signal intensities are reported in Figure 1.4. For patterns with a 25% increase in expected counts relative to unaffected points, we find that both SVSS and kSVM significantly outperform the other methods on precision for the majority of pattern sizes under consideration, indicating that points included in the top SVSS pattern are very likely to be in the true affected region for all but the smallest patterns considered. SVSS outperforms kSVM on pattern sizes larger than 7.5% of the study area. SVSS demonstrates high recall for patterns large and small, outperforming kSVM and the circular scan across all pattern sizes and outperforming all methods on pattern sizes larger than approximately 10% of the study area. Recall diminishes slightly for the circular scan, FSS, and ULS as patterns increase in size, but SVSS maintains a recall close to 1 even as patterns grow large. Although kSVM demonstrates comparable performance to SVSS with a 25% signal intensity, recall of kSVM drops significantly on patterns with weaker signals.

On weaker patterns with a 10% signal intensity, SVSS and kSVM still outperform competing methods on precision for most of the range of pattern sizes under consideration. Recall of kSVM drops dramatically on the weaker signal relative to signal intensity of 25%. SVSS is beaten by ULS on recall, but significantly outperforms kSVM on recall on the 10% signal intensity. These results suggest that even on relatively weak signals, locations returned by SVSS are very likely to be in the true affected region. The high precision of SVSS on weak signals comes at the expense of suboptimal recall, but the drop in recall is smaller than for other high-precision methods like kSVM. With stronger signals (e.g., 50% signal intensity), both SVSS and ULS demonstrate high performance across the range of pattern sizes considered and across both evaluation metrics.

Overlap for patterns with three different signal intensities is reported in Figure 1.5. On weaker patterns with a 10% signal intensity, SVSS shows a clear advantage over all competing methods for all but the smallest patterns. The circular scan outperforms SVSS for the smallest patterns, as patterns made up of only small number of clustered locations are naturally well-approximated by a circular scanning window. On patterns with a 25% signal intensity, kSVM and SVSS perform similarly with respect to overlap, with SVSS slightly outperforming kSVM on larger patterns. Finally, on stronger patterns with a 50% signal intensity, SVSS and ULS both perform extremely well, with the top pattern from both methods overlapping the true affected region almost perfectly across all but the smallest patterns. Of particular note is that kSVM does not perform as well as SVSS or ULS for patterns with the strongest intensities. While kSVM and ULS both give comparable performance on specific signal intensities, SVSS outperforms other methods on overlap across the full range of signal intensities considered. These results provide compelling evidence that SVSS represents a flexible pattern detection method that can applied for identification of both subtle and strong anomalous patterns. In comparison, competing methods either demonstrate comparable performance across all signal intensities (e.g., circular scan, DBSCAN), or perform relatively well on a narrow range of signal intensities (e.g., kSVM, ULS, FSS).

1.3.2 Detecting disease clusters

In the domain of disease surveillance, we demonstrate detection of disease clusters in mosquito pools tested for West Nile Virus (WNV), using data made publicly available by the Chicago Department of Public Health (CDPH) through the City of Chicago Data Portal. Measuring presence of WNV in mosquitoes, a relatively short-lived vector for infection, gives a useful approach to identifying spatial and temporal trends in disease risk throughout a susceptible region (Lampman et al., 2013). Patterns returned by SVSS and other scanning methods indicate the spatial clusters where the proportion of positive test results were

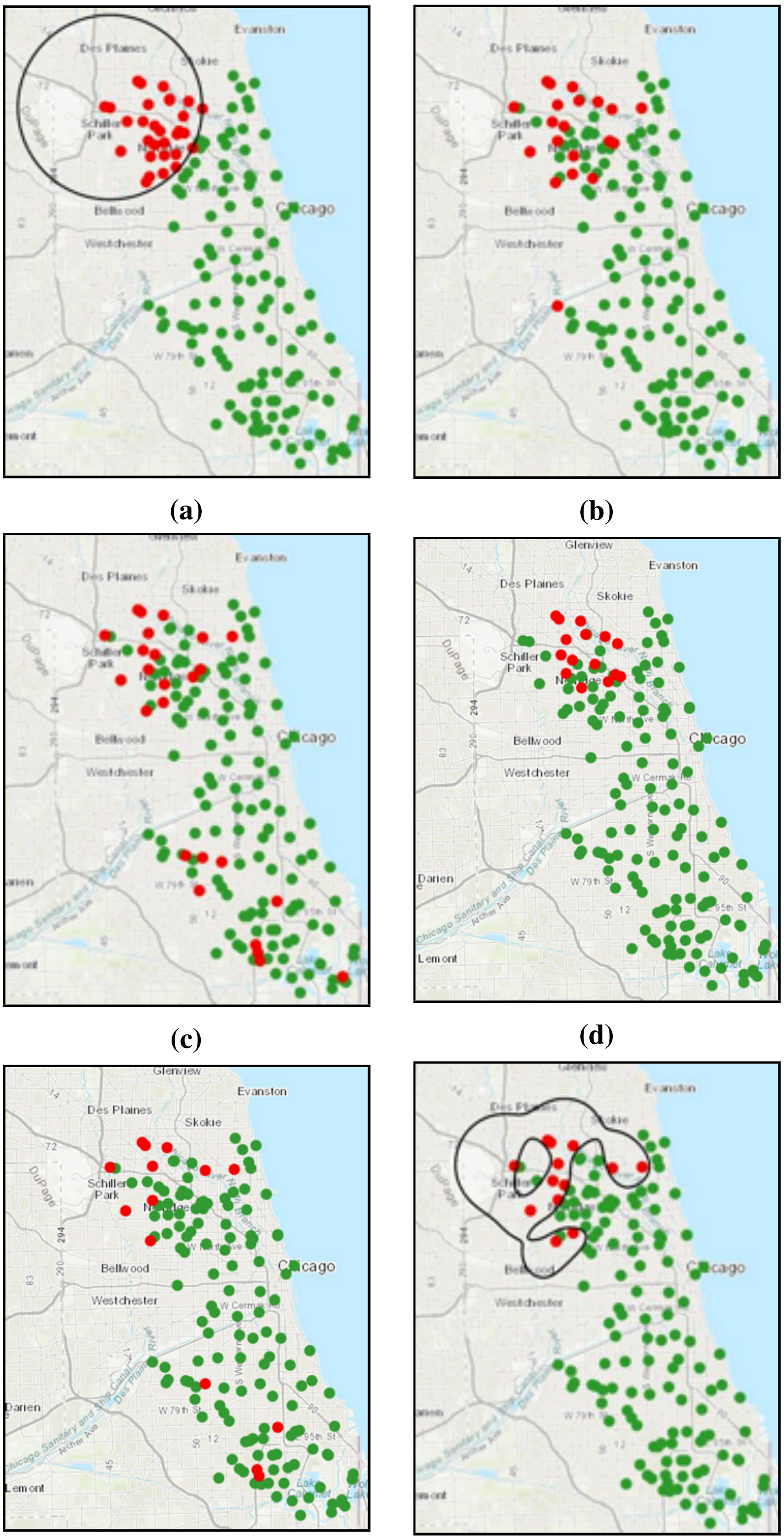
Table 1.2 Summary statistics of top West Nile Virus clusters.

	n_S	CPU Time (seconds)	*LLR*	q_{MLE}	q_{CV}	K	K_{point}
Circular scan	30	0.64	70.8	1.66	1.21	1.00	0.86
ULS	20	0.34	91.8	1.76	-	-	0.10
FSS	25	0.08	116.9	1.84	-	-	0.06
DBSCAN	15	0.16	59.2	1.75	-	-	0.15
kSVM	14	0.16	105.5	1.95	1.32	-	0.10
SVSS	13	7.40	97.3	1.87	1.36	0.62	0.17

elevated with respect to the citywide average over this period, which can help the CDPH target mosquito control programs. Mosquito management is typically implemented through the use of chemical pesticides. Accurately characterizing the spatial regions where the disease is most prevalent in mosquitoes and the risk of transmission to humans is highest can minimize the application of mosquito control measures which may have harmful effects on the ecological health of the treated areas.

Mosquito pools throughout the city are tested regularly for presence of WNV by the CDPH, with individual locations often tested multiple times a year over the course of several years. The expectation-based binomial scan statistic is appropriate in this setting due to the number of total tests varying across spatial locations. Each location thus has an observed count of positive test results c_i, an expected number of positive tests b_i, and a total number of tests n_i. We aggregate observed counts and total number of tests at each test location for a period of over 11 years from June 1, 2007 through September 30, 2018. For the expected number of positive tests, we compute an overall rate of positive test results by aggregating tests across the entire city and the entire study period, then multiply this average rate by the number of total tests n_i at each location. We thus assume a uniform rate of positive test results across test locations under the null hypothesis.

Figure 1.6 shows the top patterns detected by six detection algorithms under comparison. The circular scan is constrained in shape and approximates the shape of the true affected region, either with an overly large circle surrounding the affected locations, or with an overly small one identifying only a piece of the affected region. In comparison, SVSS has improved power to detect disease clusters that are elongated or irregular in shape. For example, the top WNV cluster detected by SVSS (Figure 1.6f) roughly conforms to sections of two major rivers in North Chicago, overlapping significant portions of the forest preserves adjacent to these rivers. FSS and ULS find patterns that are spread more widely throughout the study area and interspersed with non-anomalous points.



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Fig. 1.6 Clusters of West Nile Virus detected by six pattern detection algorithms in Chicago, IL. (a) Circular scan. (b) Upper level set scan. (c) Fast subset scan. (d) DBSCAN with thresholding. (e) Kernel support vector machine with thresholding. (f) Support vector subset scan.

The top patterns returned by each method are characterized in Table 1.2. SVSS finds a pattern with a higher *LLR* and relative risk q_{MLE} than the circular scan and DBSCAN. For additional validation of the detected patterns, we also compute the *held-out relative risk* q_{CV} by holding out points from the pattern detection methods through 10-fold cross validation, and computing the relative risk of all points that fall within the anomalous pattern decision boundary produced by running the detection method on the other 9 folds. The held-out relative risk values provide evidence that the patterns discovered are meaningful with respect to unseen data or locations not provided to the detection method. For both all methods for which we can compute q_{CV} based on the detected decision boundary, we find relatively high values that provide out-of-sample validation of the detected patterns, with SVSS outperforming the circular scan and kSVM on this out-of-sample validation measure.

As measured by K_{point}, the SVSS pattern is more compact than the patterns found by all methods except the circular scan, while still maintaining high relative risk and *LLR* comparable to ULS. FSS finds the unconstrained subset with the highest *LLR* but at the cost of low compactness. While other methods trade off compactness for high *LLR* or vice versa, the pattern returned by SVSS scores highly on both objectives. The higher spatial compactness of the SVSS pattern comes at the cost of higher computation time relative to other methods under comparison.

This analysis applied SVSS in order to detect spatial patterns over a single fixed time window, but the method can be easily extended to track changes in disease hot spots over time by updating the observed and expected values at each location as new data is received. Baseline values can be computed based on pre-outbreak levels, or continually updated based on recent trends to assess where new hot spots are occurring or where existing ones are spreading. This flexibility in definition of baseline values makes SVSS well-suited for problem settings where it is necessary to characterize the changes in anomalous patterns over time.

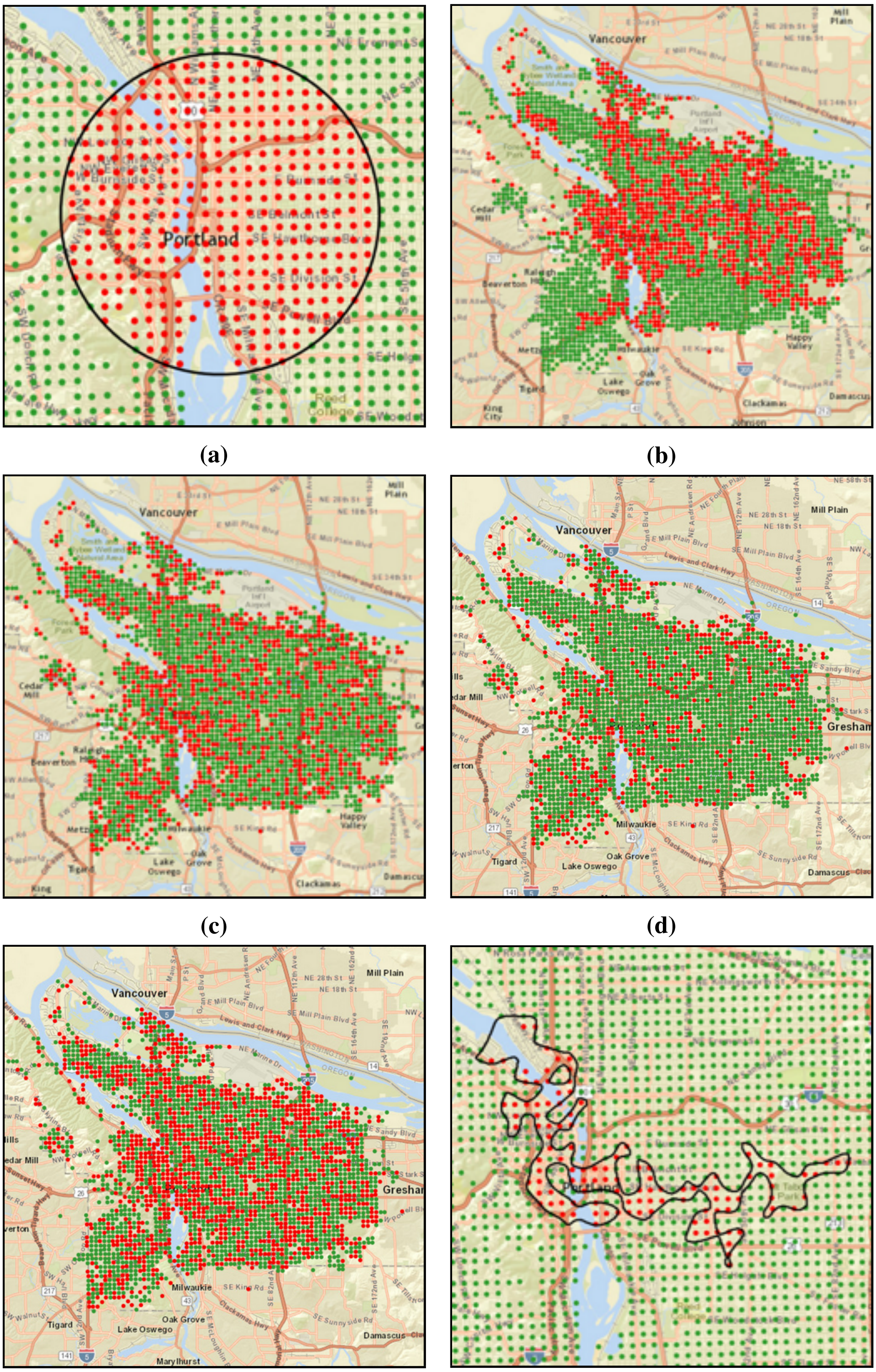
1.3.3 Detecting crime hot-spots

Next, we apply SVSS in the context of crime surveillance using calls-for-service records from Portland, OR. These records were made publicly available by the Portland Police Bureau (PPB) for the National Institute of Justice's Real-Time Crime Forecasting Challenge. We restrict our analysis to calls-for-service relating to "street crime" as categorized by the PPB, which includes assaults, robberies, shootings, stabbings, and vice-related crimes, among other crime types. We aggregate geotagged CFS records to 1000 foot square grid cells, and estimate location-specific expected counts using the time series for each cell. Specifically, we compute expected counts as an an average annual count of street crimes for each grid cell using data from the three year period from March 2012 through February 2015. Observed counts are aggregated over the following year, from March 1, 2015 through February 29, 2016. We use the expectation-based Poisson scan statistic for all six methods. Patterns returned by SVSS and other scanning methods indicate spatial regions where observed crime in the most recent year of data was elevated relative to expected counts estimated from the previous three years. Such regions could indicate newly emerging hot-spots of crime, e.g., due to changing neighborhood composition, new patterns of gang or other criminal activity, crime attractors such as bars or liquor stores, or other structural changes. While police departments are typically aware of neighborhoods with chronically high levels of crime, they may not be aware of newly emerging hot-spots which could be effectively targeted for crime prevention.

The crime patterns from six pattern detection algorithms are displayed in Figure 1.7, with summary statistics reported in Table 1.3. The circular scan finds a circular pattern covering much of Downtown and East Portland on either side of the Willamette River. The SVSS pattern is situated in roughly the same area of Southeast Portland as the circular scan pattern, but is highly irregular in shape and extends eastward to to encompass the Hawthorne District, a popular commercial strip known for its bohemian vibe and vintage clothing stores. While not a particularly high-crime area as compared to downtown Portland, the high foot traffic and store density in this area provide ample opportunity for larceny that could be prevented through targeted police patrols. The SVSS pattern has higher *LLR* and relative risk when compared with the circular scan. FSS and ULS both result in large patterns that span most of the city, with higher *LLR* than SVSS but extremely low relative compactness. The large size and relative sparsity of these patterns indicate that FSS and ULS are badly overfitting. The methods are not sufficiently constrained to produce coherent subsets, so they just pick out many individual points throughout the study region with high counts due to chance. As an additional evaluation metric, we report the count of street crimes per cell for the year following the test period, from March 1, 2016 through February 28, 2017. We find that the SVSS pattern resulted in the highest crimes per cell across all six methods in the year following the test period. Even though FSS and ULS pick out points with high relative risk in the training data (comparable to SVSS), the points chosen by SVSS have much higher crime rate in the subsequent year's data and thus seem to be a much better target for proactive police patrols.

1.3.4 Detecting pothole clusters

For our final application, we apply SVSS in the domain of city services and management to detect clusters of pothole complaints in Pittsburgh, PA. Our data set for this analysis consists



(a) (b) (c) (d) (e) (f)

Fig. 1.7 Clusters of street crime detected by detected by six pattern detection algorithms in Portland, OR. (a) Circular scan. (b) Upper level set scan. (c) Fast subset scan. (d) DBSCAN with thresholding. (e) Kernel support vector machine with thresholding. (f) Support vector subset scan.

Table 1.3 Summary statistics of top street crime clusters.

	n_S	CPU Time (seconds)	LLR	q_{MLE}	Next-year crimes/cell	K	K_{point}
Circular scan	347	36.4	257.4	1.26	29.9	1.00	0.910
ULS	1102	171.7	986.3	1.39	17.0	-	0.005
FSS	945	0.3	1687.5	1.77	11.0	-	0.002
DBSCAN	536	0.2	1311.2	2.69	4.8	-	0.003
kSVM	1252	6.6	1310.3	1.48	13.7	-	0.003
SVSS	115	379.6	420.0	1.62	32.4	0.23	0.027

of publicly available call records from Pittsburgh's 311 system. People living in Pittsburgh can call the 311 telephone center to notify the city of any non-emergency issues, including requests for service related to road deterioration. Potholes represent one of the most common issues reported to the city, with pothole reports making up 13.1% of all 311 calls between 2016 and 2018. Detecting clusters in these reports has the potential to help public works agencies in Pittsburgh and other cities identify and efficiently respond to emerging clusters of potholes.

We aggregate counts of pothole reports to city blocks, using a two-year period from January 1, 2016 through December 31, 2017 to estimate a city-wide average annual count of potholes. As in the disease outbreak detection context, we thus assume a uniform baseline rate of pothole reports under the null hypothesis. We find observed counts for each city block from January 1, 2018 through December 31, 2018, and apply the expectation-based Poisson scan statistic to search for spatial regions with elevated counts of potholes in 2018 in comparison to the previous two years. Such clusters could indicate newly emerging regions in need of attention due to weather events or recent shifts in traffic patterns contributing to road surface deterioration, helping public works agencies plan and prioritize future road maintenance projects.

For this analysis, we demonstrate an alternate approach to finding irregular patterns with SVSS that may have multiple disconnected regions. In many real-world use cases for pattern detection, multiple affected regions exist in the same data and we therefore would benefit from a method for both detecting and prioritizing over many anomalous clusters. If desired for operational purposes, SVSS allows users to rank the disconnected regions by the unpenalized log-likelihood ratio statistic and choose k components to include in order to retrieve an anomalous pattern of the desired scale (Higher k leading to a larger pattern consisting of more disconnected but individually compact regions). Instead of selecting the

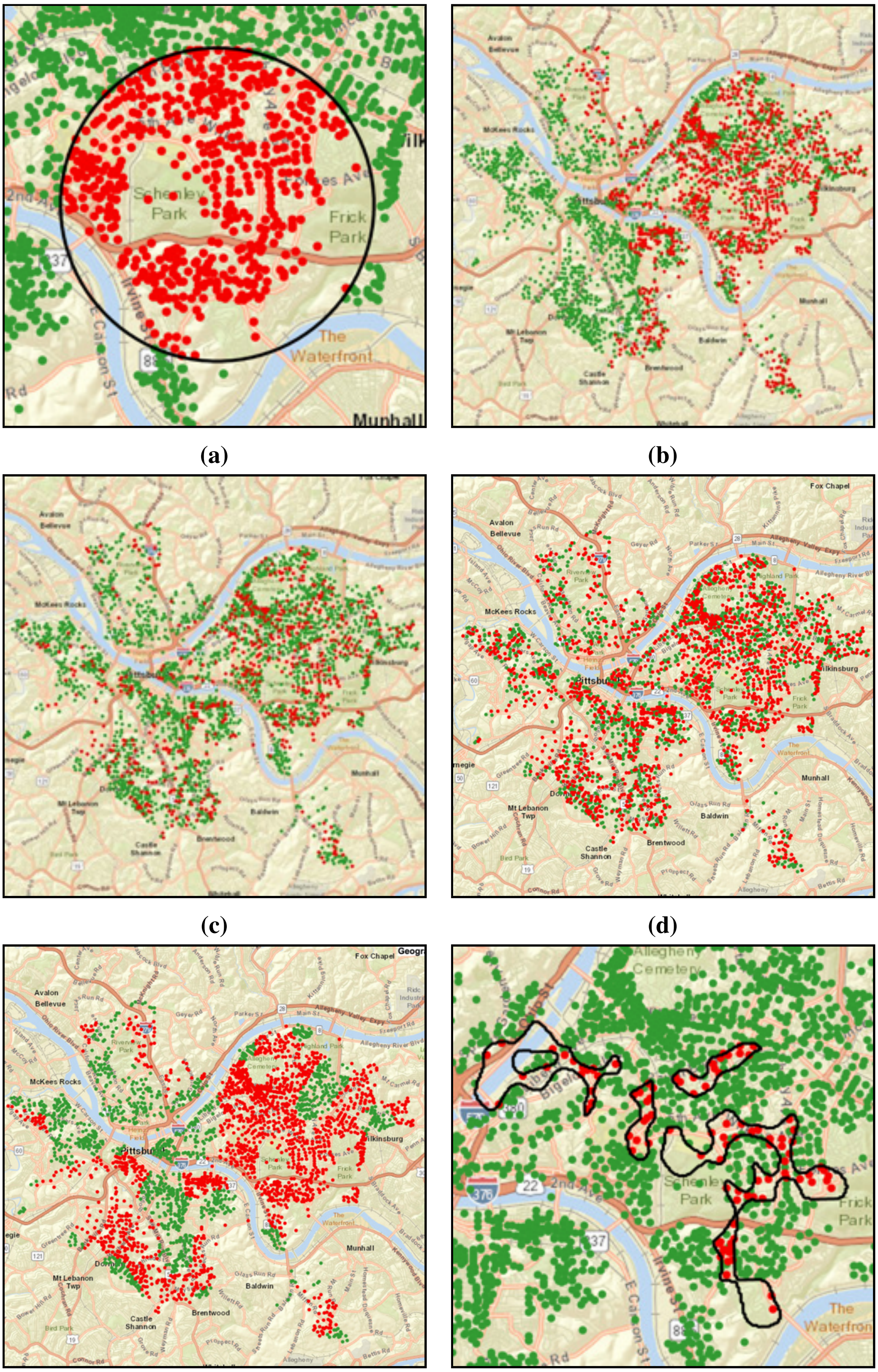
Table 1.4 Summary statistics of top pothole clusters.

	n_S	CPU Time (seconds)	*LLR*	q_{MLE}	q_{CV}	K	K_{point}
Circular scan	497	28.3	2038.0	4.48	4.15	1.00	0.876
ULS	1096	140.3	6128.3	5.22	-	-	0.006
FSS	642	0.3	9182.2	8.78	-	-	0.003
DBSCAN	1607	0.2	7635.4	4.81	-	-	0.003
kSVM	1805	2.0	4242.2	3.48	3.26	-	0.076
SVSS	111	131.4	2272.3	10.91	4.12	0.42*	0.030*

* denotes average over top five disjoint components.

single top component from the connected components of the SVM decision boundary as discussed in Section 1.2.4, here we include the top 5 disconnected components of the pattern returned by SVSS. Public works agencies could scale a proposed infrastructure project up or down based on operational constraints by increasing or decreasing the number of disjoint components to include.

Figure 1.8 displays the top pothole clusters returned by six pattern detection methods, and Table 1.4 provides summary statistics for these patterns. For the compactness measures, we report the average compactness across top 5 components for SVSS. As discussed above, SVSS returns a pattern consisting of multiple disconnected regions. This pattern has the highest relative risk among the detection methods under comparison, and higher *LLR* than the circular scan pattern. The held-out relative risk q_{CV} of both SVSS and the circular scan are comparable and higher than that of kSVM. The individual components of SVSS correspond to highly trafficked roads and intersections throughout Pittsburgh that are subject to high rates of wear and degradation, with 4 of the 5 components overlapping one or more public bus routes. The disconnected regions which make up the SVSS pattern are elongated due to the underlying spatial structure of the road network. Yet these regions are still individually compact, as indicated by the high average geometric compactness measures relative to the sparse and underconstrained patterns found by ULS and FSS. As in the previous two applications, SVSS scores relatively highly on *both* compactness and measures of anomalousness, resulting in patterns that are highly anomalous but still spatially coherent.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 1.8 Clusters of potholes detected by six pattern detection algorithms in Pittsburgh, PA. (a) Circular scan. (b) Upper level set scan. (c) Fast subset scan. (d) DBSCAN with thresholding. (e) Kernel support vector machine with thresholding. (f) Support vector subset scan.

1.3.5 Discussion of real-world case studies

In all three of the above case studies, the literature reveals multiple distinct environmental factors that can drive West Nile Virus, crime, or potholes respectively. Thus, these factors do not clearly indicate which part of the city to target with public health, law enforcement, or road maintenance interventions respectively, while our approach precisely localizes a spatial area that can benefit from targeted intervention.

For West Nile Virus, *Culex* species mosquitos which transmit the virus can breed in a variety of stagnant water sources, including low places with poor drainage, urban catch basins, roadside ditches, sewage treatment lagoons, and manmade containers around houses (Ruiz et al., 2007). A variety of other factors including temperature, humidity, rainfall, surface permeability, and bird migration patterns were also identified as predictive (Hernandez et al., 2019). Human WNV cases in a 2002 outbreak in Chicago were found to be associated with higher percentages of vegetation in a census tract, and areas in Chicago's inner suburbs were found to have higher human WNV rates than either the outer suburbs or the urban center (Ruiz et al., 2007). Thus the prior literature supports our identification of certain forest preserve and river areas as WNV hot spots but does not necessarily point to these particular areas in North Chicago. Similarly, the literature on crime prediction reveals that chronic hot spots of crime are often found in large commercial areas and nearby residential areas (Fitzpatrick et al., 2019), and while the Hawthorne District is one well-known commercial district of Portland, there was no reason to expect a priori that this particular strip would exhibit a flare-up of property crimes in the particular year of data under analysis. Finally, predictive factors for pothole formation include weather (temperature and freeze-thaw cycles), pavement condition, and traffic loads (Sadeghi et al., 2016). An analysis by the Metropolitan Transportation Commission (MTC, June 2011) estimates that buses and other large vehicles create thousands of times more physical stress on pavements per trip as compared to passenger vehicles, supporting our discovery of spatial clusters of potholes in particular, heavily trafficked bus routes in Pittsburgh.

1.4 Conclusions

In this chapter, we introduce the support vector subset scan (SVSS), a novel method for detecting anomalous patterns in spatial data that are spatially compact and irregular in shape. SVSS integrates soft spatial constraints into the fast subset scan, rewarding patterns with spatial coherence. As demonstrated above in the contexts of disease outbreak detection, crime surveillance, and city services and management, SVSS provides a flexible framework for spatial pattern detection in a variety of problem settings where detection and characterization of coherent anomalous patterns in spatial data has demonstrable real-world benefits. Characteristics of patterns returned by SVSS may also be helpful as features in predictive models related to the spatial data in question. For example, grid cells returned by SVSS as part of crime clusters reported more crime in the following year than those returned by other pattern detection methods. In future work, the authors plan to further evaluate how inclusion of SVSS cluster attributes can improve prediction models in areas of public health and safety.

Appendix 1.A Implementation Details

As discussed in Section 2.1, Speakman et al. (2016) provide the expressions for the log-likelihood ratio statistics $\lambda_i(q)$ for the expectation-based binomial scan statistic (EBB, used in Section 3.2), the expectation-based Poisson scan statistic (EBP, used in Sections 3.1, 3.3, and 3.4) and others in the exponential family. We include the expressions for $\lambda_i(q)$ in Table 1.5 for ease of reproducibility. We also report optimized parameter values for the three SVSS tuning parameters used in the experiments in Sections 3.2-3.4 in Table 1.6.

To evaluate our method in an experimental setting, we generate patterns of varying size, shape, and intensity in simulated data, and compare precision and recall of SVSS with five other methods for spatial pattern detection:

- The *circular scan statistic*, which searches over N^2 total circles and returns the circle with the highest log-likelihood ratio (LLR). For each location, we evaluate the N circles of increasing radius centered at the location, such that each successive circle grows to include one additional neighboring point (Kulldorff, 1997).

- The *upper level set scan statistic (ULS)*, which searches over connected components of all possible upper level sets with respect to the ratio of observed values to baselines. ULS searches over tessellated cells rather than points, so we construct a Voronoi tessellation from points in space as a pre-processing step (Patil and Taillie, 2004).

- The *fast subset scan (FSS)*, which returns the subset of locations which maximizes the unpenalized LLR (Neill, 2012a).

- *DBSCAN with thresholding*, a clustering algorithm that finds high-density clusters of arbitrary shape (Ester et al., 1996). Only with locations with count-to-baseline ratios above a fixed threshold are clustered. The threshold and DBSCAN parameters are selected to optimize anomalousness (LLR) of the top cluster. The single cluster with highest LLR is considered as the pattern returned by DBSCAN.

- *Kernel Support Vector Machine (kSVM) with thresholding*, which applies a threshold to the ratio of counts to baselines for each location, then trains an SVM with a Gaussian kernel to learn a nonlinear decision boundary between points above and below the threshold. The threshold and SVM parameters are chosen using 10-fold cross-validation to optimize the anomalousness score (LLR).

Table 1.5 Location-specific contributions to the score function for expectation-based statistics in the exponential family. See (Speakman et al., 2016) for full derivations.

Distribution	$\boldsymbol{\lambda_i(q)}$
Poisson	$c_i \log q + b_i(1-q)$
Gaussian	$c_i b_i \frac{(q-1)}{\sigma_i^2} + b_i^2(\frac{1-q^2}{2\sigma_i^2})$
exponential	$\frac{c_i}{b_i}(1-\frac{1}{q}) - \log q$
binomial	$c_i \log q + (n_i - c_i)\log(\frac{n_i - qb_i}{n_i - b_i})$
negative binomial	$c_i \log q + (r_i + c_i)\log(\frac{r_i + b_i}{r_i + qb_i})$

Table 1.6 Parameter values for Support Vector Subset Scan on real world data sets.

Data Set	**Gaussian kernel bandwidth**	$\mathbf{C_0}$	$\mathbf{C_1}$
Chicago West Nile	0.09	50	100
Portland street crime	0.03	100000	200000
Pittsburgh potholes	0.03	1000	2000

Chapter 2

Assessing Risk of Opioid Misuse from Prescription Drug Monitoring Data[1]

2.1 Introduction

Prescription drug misuse has rapidly become one of the most common forms of illicit drug use in the United States, with an estimated 1.7 million people suffering from prescription pain reliever use disorder and an estimated 9.9 million people misusing prescription pain relievers at least once in 2018 (Lipari and Park-Lee, 2018). Rates of drug overdose deaths caused by prescription opioids have been increasing year-over-year since the 1990s, reaching a peak of 13.4 deaths[2] per 100,000 people in 2017, the most recent year reported by the Centers for Disease Control and Prevention (CDC) (Hedegaard et al., 2018). 75% of heroin users report that their first experience with opioids was a prescription pain reliever (Cicero et al., 2014). Opioids nonetheless represent an important option for clinicians in the mitigation of chronic pain. In a 2015 literature review, Vowles et al. found that rates of addiction averaged between 8-12% for patients being prescribed opioids for chronic pain, suggesting an outstanding need for accurate assessment of risk for patients presenting with symptoms of chronic pain to clinicians.

The rise in abuse of prescription pain relievers coincided with an increase in legitimate prescription of opioids in the 1990s and early 2000s for addressing chronic pain (Kuehn, 2007). Rigg et al. (2010) identify a set of practices contributing to the abuse and diversion of prescription opioids in this period. So-called “pill mills” enabled patients to request specific medications, accepted cash as the only form of payment, and engaged in aggressive

[1] This chapter is based on joint research with Daniel B. Neill. We wish to acknowledge Fan Xiong (Kansas Department of Health) for providing data and domain expertise for this study.

[2] Excluding deaths from heroin and methodone overdose.

advertising campaigns to attract customers. On-site pharmacies simplified the process of obtaining prescription drugs for patients and facilitated obtaining prescriptions from multiple doctors. Liberal prescribing habits meant that patients could easily obtain a prescription for a much higher dose or a stronger drug than what was needed for their symptoms. Sponsored drug diversion was the practice of individuals sharing the cost of a doctor's visit and prescription and splitting the drugs afterwards. A lack of regulatory oversight enabled doctor and pharmacy shopping, in which patients obtained prescriptions from to multiple prescribers simultaneously or sought out those with liberal prescribing tendencies. Falsification of symptoms was common among patients seeking to obtain a prescription without any legitimate source of chronic pain. These practices all contributed to a historic rise in the abuse of prescription opioids, leading to corresponding increases in cases of substance abuse disorder and overdose deaths from opioids (Hedegaard et al., 2018).

In recent years, federal and state governments have taken measures to curb the unsafe prescribing practices that proliferated in the past three decades. In 2017, the U.S. Department of Justice formed the Opioid Fraud and Abuse Detection Unit, with the express goal of increasing surveillance of opioid-related healthcare fraud and prosecuting individuals contributing to the prescription opioid epidemic (DOJ, 2017). At the state level, prescription drug monitoring programs (PDMPs) have been deployed in 49 out of 50 U.S. states. These programs maintain a statewide electronic database that tracks all prescriptions of controlled substances within the state, requiring that pharmacists enter prescription information into the system before dispensing drugs.

In most states, clinicians are required to query PDMPs before writing a new prescription for a controlled substance, thus providing them a natural checkpoint for assessing a patient's recorded history of prescription drugs. In addition, several question-based risk assessment tools have been developed to provide additional support to clinicians interacting with patients in need of pain treatment, as surveyed by Ducharme and Moore (2019). These risk assessment tools all rely on self-reporting and require that patients answer questions truthfully about their symptoms and history of drug use. Ducharme and Moore identify this weakness in all existing screening tools for risk assessment, and recommend supplementing these tools with additional resources for determining whether prescription of opioids or other controlled substances is safe and warranted.

The widespread adoption of statewide PDMPs in the U.S. has laid the groundwork for much better regulatory oversight of prescription practices on the part of prescribers, pharmacies and patients. These programs represent a significant improvement in data infrastructure available to help physicians and pharmacists make decisions about which patients can safely be prescribed opioids. Access to these new resources also raises questions about how best to use the records stored in these databases to promote safe practices.

In this work, we discuss approaches for leveraging prescription drug monitoring data to assess risk of opioid misuse based on patient-level opioid time series obtained from a statewide PDMP. The majority of previous studies applying predictive models to predict risk of adverse outcomes related to opioid misuse rely on medical records that may be difficult for public health agencies to access outside of a clinical setting (Hylan et al., 2015). Ferris et al. (2019) linked PDMP data to patient overdose deaths and predicted fatal overdose using a multivariate logistic regression. Hastings et al. (2020) use Medicaid claims data to predict risk of poor outcomes from a wide array of variables related to a patient's history in state-maintained databases, such as demographics, insurance claims, arrests, and payments received from social welfare programs, and find strong predictive power. To our knowledge, no previous studies have investigated whether risk of misuse can be predicted using only signals present in PDMP data, which represent the most widely available data to state and federal public health agencies.

In the remainder of this chapter, we develop two approaches for leveraging patient timelines generated from a PDMP database for individual-level assessment of opioid risk. In Section 2.2, we introduce a shape-based clustering framework to evaluate risk of misuse in new patients when no ground truth data is linked to prescription timelines. In Section 2.3, we move to a semi-supervised setting for predicting risk of opioid misuse, proposing and evaluating a novel approach for prediction that leverages recurrent generative adversarial networks (RGANs) for risk assessment in new patients when a small pool of trajectories are linked to known cases of unsafe drug use. We close with a discussion of implications and future research directions in Section 2.4.

2.2 Trajectory Clustering for Early Risk Assessment

The primary difficulty in using PDMP data for patient-level prediction stems from the lack of explicit patient outcomes linked to prescriptions in the data set. In this section, we discuss an approach for early risk assessment of new patients based on supplementing a shape-based clustering analysis with noisy signals of opioid misuse that are observed in prescription records.

2.2.1 Shape-based time series clustering

As part of a collaboration with epidemiologists at the Kansas Data-Driven Prevention Initiative, we were provided access to de-identified records from the Kansas state PDMP for a three-year study period ending in 2015. Using duration and quantity of prescribed opioids aggregated across prescribers for a given patient, we generate timelines of morphine milligram equivalents (MME) for individual patients. Given this aggregate measure of total opioids being prescribed across different opioid types and dosages, we pose the following research question: can we identify early indicators in patient MME timelines that are predictive of later opioid misuse?

As an initial preprocessing step, we apply a 14-day moving average to MME timelines to smooth out spikes occurring when prescriptions overlap. We also align patients at their first day of non-zero MME, and keep only patients with at least two years of overlap with the study period after their first opioid prescription. After preprocessing, 387,023 patients remain in the data set, whom we randomly allocate into a training and test set using a 75-25 split.

To understand what common patterns appear across patients, we apply the k-shape algorithm (Paparrizos and Gravano, 2015) to group patients together who have similar patterns in their smoothed MME time series. k-shape is an algorithm for partitional clustering that proceeds similarly to the popular k-means method for clustering; the algorithm alternates between updating cluster members according to the closest cluster centroid according to a shape-based distance metric, then updates cluster centroids based on changes to cluster membership. This method is particularly well suited for the task of clustering MME time series, because we would like to group patients together that have the same *characteristic shape* in their prescription opioid timeline, regardless of differences in scale or translation. For example, if two patients both experience a steady increase in MME at some point in their timeline, we would like to group those patients together even if the increases begin at different times or the absolute dosage levels vary.

Figure 2.1 shows the cluster centroid and a sample of member time series for 8 clusters identified using the k-shape algorithm. Figure 2.1 shows the cluster centroid and four randomly selected patient time lines for an example cluster. The characteristic shape is apparent in the centroid as well as the samples; the patients experience a dip or period of relatively flat MME, followed by a steady increase over time. As the shape-based distance is translation invariant, the increases in MME are not aligned across patients.

Using the characteristic shape of the cluster centroids and member time series, we can begin to infer which clusters may be associated with higher-risk patients, but additional verification is needed based on other signals available in the prescription records. Aided by

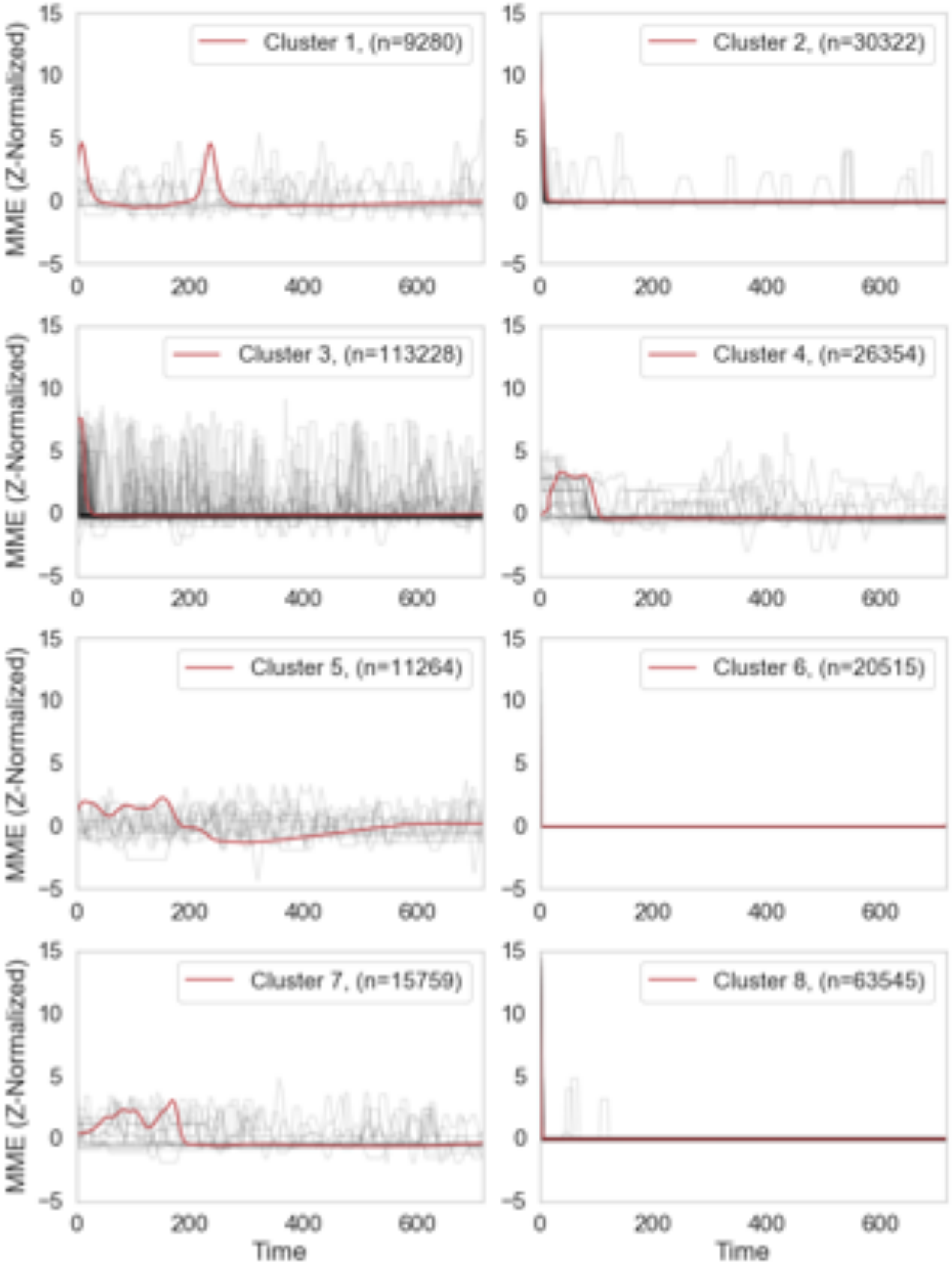


Fig. 2.1 Cluster centroids from shape-based clustering applied to smoothed morphine milligram equivalent (MME) time series. Sample of member time series in training set shown in gray behind cluster centroids (red).

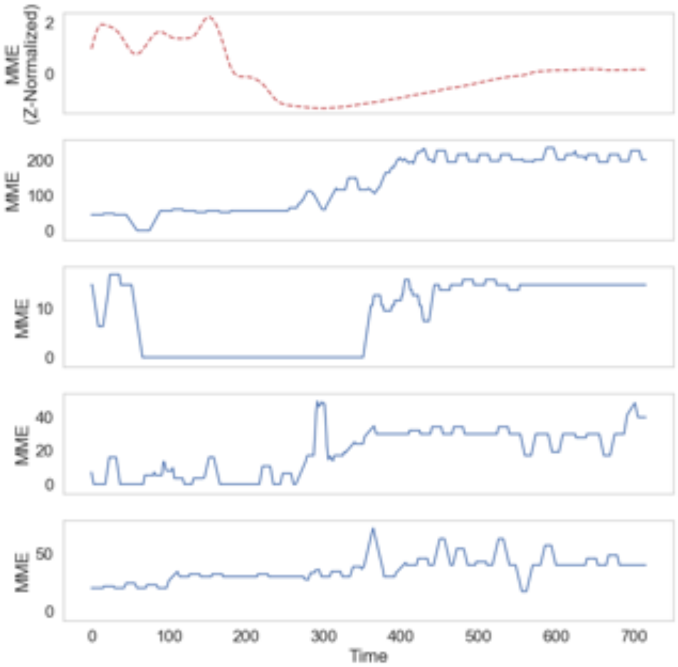
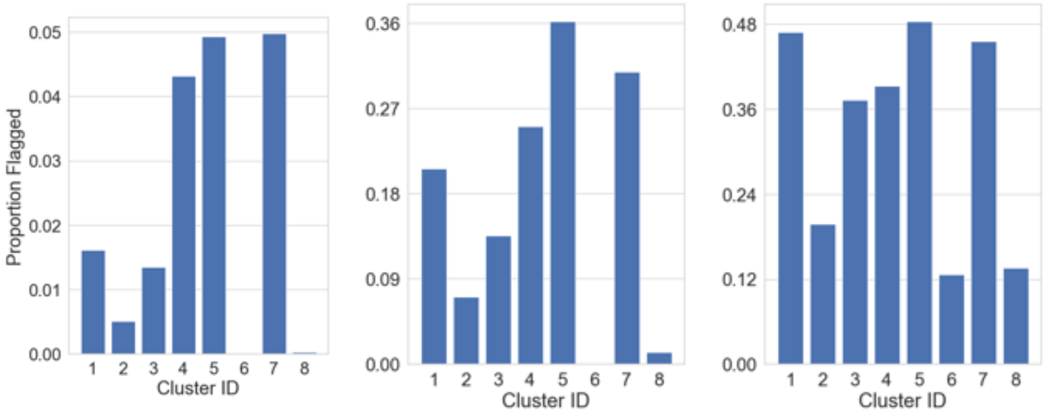


Fig. 2.2 Cluster centroid (top) and sample of four patient time series (bottom) showing characteristic shape for example cluster (Cluster 5).

epidemiologists at the Kansas Department of Health and Environment, we identified three red flag indicators in PDMP data that suggest unsafe prescription drug practices: (1) greater than two simultaneous opioid types, (2) greater than one simultaneous opioid prescriber, and (3) benzodiazepine and opioid prescribed simultaneously. We show a visual comparison of red flag rates across clusters in Figure 2.3 and report red flag rates by cluster in Table 2.1. Although the PDMP data does not contain explicit information on patient outcomes, the rates of red flag indicators enable us to associate some groups of patients with a higher risk of prescription opioid misuse.

Based on characteristics of the cluster centroid, representative samples, and red flag rates across clusters, we determined that Clusters 1, 5, and 7 represent a high-risk group of patients relative to patients belonging to other clusters. Individuals from Clusters 1, 5, and 7 make up approximately 12.5% of the patients in the training set and account for 23.1% of total red flags. Cluster 1 tends to include patients with periodic spikes in daily MME above safe levels. Cluster 5 includes patients with a steady increase over the two-year observation period (See Figure 2.2). Cluster 7 similarly includes patients with a period of plateauing or steadily increasing MME, but also includes many patients with a drop to zero MME after these periods. It is possible that this drop to zero could represent the patient stopping their prescriptions or moving out of state and thus dropping out of the monitoring program.



(a) (b) (c)

Fig. 2.3 Proportion of patients triggering red flags by cluster for three indicators of unsafe prescriptions. Red flag proportions shown for (a) greater than two simultaneous opioid types, (b) greater than one simultaneous opioid prescriber, and (c) benzodiazepine and opioid prescribed simultaneously.

However, the high rate of red flags for this group suggests other explanations related to prescription drug use, such as overdose or a shift to other forms of illicit drugs that are not monitored by the PDMP. Cluster 4 was also considered for inclusion in the high-risk group, but examination of the cluster centroid and sample patient time series indicate that this cluster is associated with patients with a short-term increase in MME that returns to zero after a handful of prescriptions.

The lowest-risk clusters tend to group patients with only a brief history recorded in the PDMP. Clusters 2, 3, 6, and 8 appear to group patients with MME dropping to zero quickly after the initial prescription, with only occasional increases above zero observed in Clusters 2, 3, and 8.

For the purposes of this analysis, we define a high-risk group of patients solely based on cluster membership. Other relevant patient characteristics (e.g., average 30-day MME, demographic characteristics, opioid type) could be combined with cluster membership to define more granular subgroups for consideration as high-risk subsets of individuals.

2.2.2 Early risk assessment of partial time series

Once a high-risk group has been identified, a natural approach suggests itself for assigning cluster membership for new patients without extensive prescription records in PDMP data, who may have relatively short opioid time series present in the data set. Adapting the shape-based distance measure from Paparrizos and Gravano (2015), we propose assigning

Table 2.1 Summary statistics of patient clusters grouped with shape-based clustering approach on smoothed morphine milligram equivalent (MME) time series.

Cluster	% of Patients	% Flagged, > 2 Simult. Opioid Type	% Flagged, > 1 Simult. Prescriber	% Flagged, Opioid+Benzo
1	3.2	1.62	20.66	46.9
2	10.5	0.51	7.10	19.8
3	39.0	1.34	13.58	37.3
4	9.1	4.31	25.07	39.3
5	3.9	4.93	36.18	48.4
6	7.1	4.87	0.09	12.7
7	5.4	4.97	30.90	45.5
8	21.9	0.03	1.22	13.6

partial time series to the nearest cluster centroid and detecting a patient as high-risk if they are assigned to one of the high-risk clusters. To find the nearest cluster centroid to a partial time series of general length, we first extend the definition of shape-based distance to accept sequences with different lengths.

Let $CC_{\mathbf{xy}}$ be the cross-correlation or *sliding inner product* between vectors $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$ for a sequence of shifts $w \in \{1,2,...,m+n-1\}$. The cross-correlation is defined as:

$$CC_{\mathbf{xy}}(w) = R_{\mathbf{xy}}(w-m) \tag{2.1}$$

where

$$R_{\mathbf{xy}}(k) = \begin{cases} \sum_{i=1}^{n-k} x_{m-n+k+i} \cdot y_i & k > n-m \\ \sum_{i=1}^{m} x_i \cdot y_{n-m-k+i} & 0 \le k \le n-m \\ \sum_{i=1}^{m+k} x_i \cdot y_{n-m-k+i} & k < 0 \end{cases} \tag{2.2}$$

The shape-based distance $SBD(\mathbf{x},\mathbf{y})$ is then defined by finding the maximum over the cross-correlation sequence.

$$SBD(\mathbf{x},\mathbf{y}) = 1 - \max_{w} \left(\frac{CC_{\mathbf{xy}}(w)}{||\mathbf{x}||_2 ||\mathbf{y}||_2} \right) \tag{2.3}$$

Note that we apply the same coefficient normalization approach proposed by Paparrizos and Gravano (2015) to maintain the property that this distance is in the interval [-1, 1]. Note that the endpoints of this interval will not be reached for $m \neq n$ with this normalization approach, but we still obtain a meaningful metric for comparing relative distance to cluster centroids each having equal lengths.

Simulating observation of a new prescription drug trajectory over time, we find the nearest cluster centroid at each time step with respect to the SBD distance metric. We apply this detection approach to all patients in our test set, updating the cluster assignment at each observed time step. We mark the **detection time** as the earliest time at which a patient is assigned to a high-risk cluster after a minimum observation period of 14 days.

As visualized in Figure 2.4, we define *lead time* as the difference between the time of the earliest red flag trigger and the earliest high-risk detection time. A positive lead time indicates that the cluster-assignment approach was successful for early risk detection relative to red flag indicators. In practice, a large and negative lead time is no worse than a lead time of zero. Similarly, there is likely to be an upper bound on the practically useful lead time for early detection. We therefore bound lead times below by zero and above by a maximum threshold of 30 days. Lead times for all three red flag indicators are summarized in Table 2.2.

We obtain p-values for “percent detected with positive lead time” and “average lead time” statistics through permutation testing. In each simulation, we fix the observed detection times but randomly permute which patients are detected. We therefore test the null hypothesis that we observed these statistics by chance alone, conditioned on the observed detection times. We find that both “percent detected with positive lead time” and “average lead time” are statistically significant at $\alpha = 0.01$ for all three red flag indicators. Lead time summary statistics under randomly-permuted alerts over 1000 simulations are provided in Table 2.3.

These results indicate that early risk assessment of partial time series using a cluster-assignment method can provide significant improvements over a system of red flags based on simple indicators triggered by PDMP records. We detect 39.9% of all the “greater than two simultaneous opioid types” red flags with an average lead time of 10.0 days. While the average lead times are shorter for the other two red flags, they still provide several days of advance notice before the simple indicator is triggered, which may provide a crucial window for a clinician deciding whether it is safe to prescribe additional drugs for a new patient.

2.3 Recurrent Generative Adversarial Networks for Semi-Supervised Learning of Opioid Misuse

In the previous section, we discussed an approach for assessing risk of unsafe prescribing practices when data on patient outcomes are unavailable outside of prescription information

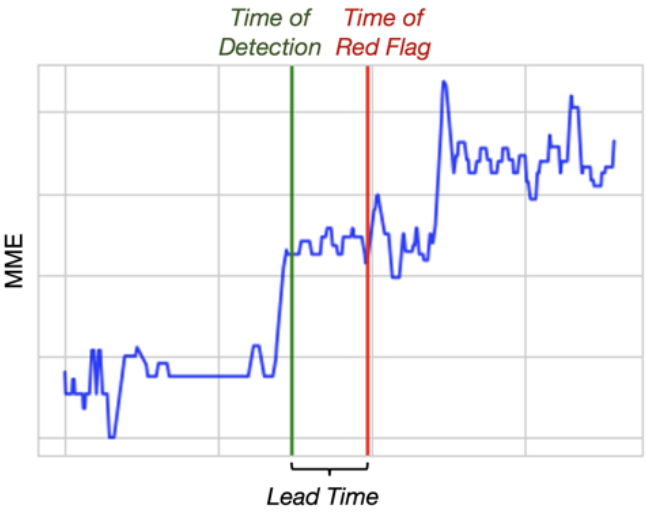


Fig. 2.4 Diagram of early detection lead time for a single patient with increasing morphine milligram equivalents and a red flag indicating an unsafe prescription.

Table 2.2 Summary statistics of red flag lead times for high-risk alerts.

Red Flag	% of Patients	% Detected with Positive Lead Time	Average Lead Time (Days)
>2 Simultaneous Opioid Types	1.51	39.9*	10.04*
>1 Simultaneous Opioid Prescriber	12.5	24.2*	5.42*
Benzo+Opioid	29.8	13.9*	3.15*

* Significant at $\alpha = 0.01$. P-values obtained through permutation testing.

monitored by the PDMP. Data available to public health agencies on opioid-related health outcomes is often limited to a small pool of patients for whom extreme adverse events are known (e.g., overdose death), or to patients for whom health outcomes can be discerned from signals present in the PDMP data (e.g., high and increasing MME or prescription of pharmacological treatments for opioid dependence). Ideally, public health agencies would like to learn from both the small set of patients for which they have observed poor health outcomes, as well as the much larger pool of unlabeled patients monitored by the PDMP, in order to learn patterns that are predictive of adverse outcomes. This combination of labeled and unlabeled data suggests that a semi-supervised classification approach, in which a prediction method leverages both a small set of labeled samples and a large pool of unlabeled samples to train a classifier, may be beneficial.

In this section, we will present a novel approach for semi-supervised classification of time series using recurrent generative adversarial networks (RGANs). We first discuss

Table 2.3 Summary statistics of red flag lead times for randomly permuted alerts, averaged over 1000 simulations.

Red Flag	% of Patients	% Detected with Positive Lead Time	Average Lead Time (Days)
>2 Simultaneous Opioid Types	1.51	10.3	2.9
>1 Simultaneous Opioid Prescriber	12.5	9.0	2.6
Benzo+Opioid	29.8	6.0	1.7

performance of semi-supervised recurrent generative adversarial networks (SS-RGANs) on four medical time series classification tasks, which provide a set of benchmark tasks for comparison against competing classification methods. Then, we present performance of SS-RGAN and competing methods on prediction of unsafe levels of prescribed opioids using PDMP data, incorporating red flag indicators and high-risk cluster assignment time series from Section 2.2 as conditional inputs to the SS-RGAN model.

2.3.1 Adapting recurrent generative adversarial networks for semi- supervised learning

Since their introduction by Goodfellow et al. (2014), generative adversarial networks (GANs) have attracted significant attention for their ability to generate realistic images that are often indistinguishable from real ones, even for human faces or other complex imagery. GANs learn to simulate the underlying data distribution by pitting a generative neural network and a discriminative neural network against each other as the models are trained simultaneously. The generator attempts to generate realistic samples that are indistinguishable from real data, while the discriminator attempts to classify samples as real or fake. Hyland et al. (2017) propose an extension to the original GAN framework, the recurrent GAN (RGAN), in which both the generator and discriminator are replaced with a recurrent neural network, resulting in a framework for generating realistic, real-valued multivariate time series.

In the RGAN optimization procedure, the discriminator minimizes the average cross-entropy between predictions and the labels of a sequence, averaged across time steps. Let $RNN(X)$ represent the vector of T outputs from a recurrent neural network (RNN) taking a sequence of T input vectors $\{\mathbf{x}_t\}_{t=1}^{T}$ with each $\mathbf{x}_t \in \mathbb{R}^d$, and let $CE(\mathbf{a},\mathbf{b})$ be the average cross-entropy between two sequences $\mathbf{a}$ and $\mathbf{b}$. Given a pair of sequence outputs and labels $\{X_i, \mathbf{y}_i\}$

with $X_i \in \mathbb{R}^{T \times d}$ and $\mathbf{y}_i \in \{0,1\}^T$, the loss function for discriminator RNN_D is expressed as

$$D_{loss}(X_i, \mathbf{y}_i) = CE(RNN_D(X_i), \mathbf{y}_i)) \tag{2.4}$$

where $\mathbf{y}_i$ is a sequence of 1s for real samples and a sequence of 0s for generated samples. The input Z_i to the generator RNN_G is a sequence of T points sampled independently from the latent noise space $\mathbf{Z}$. As the generator is attempting to produce sequences the discriminator cannot distinguish from real ones, the generator loss is formulated as the average cross-entropy between the discriminator's predictions on generated sequences and the sequence of 1s (indicating the "true" class label).

$$G_{loss}(Z_i) = CE(RNN_D(RNN_G(Z_i)), \mathbf{1}) \tag{2.5}$$

Hyland et al. (2017) show that alternating updates to the generator and discriminator based on these models results in a generator that can successfully simulate data distributions across multiple domains, such as sine waves, smooth functions, handwritten digit sequences, and medical time series.

We propose a modification to this method to adapt RGANs for the task of semi-supervised time series classification. Specifically, we split the discriminator model into two separate unsupervised and supervised discriminators with shared feature weights, where the unsupervised discriminator still attempts to distinguish real time series from fake ones, and the supervised discriminator attempts to classify from among K classes on the actual classification task of interest. We follow the efficient implementation for this dual-discriminator approach proposed in Salimans et al. (2016), in which the supervised discriminator is first defined with a softmax output activation over K classes corresponding to the classification task of interest. The unsupervised discriminator takes the input to the softmax function from the supervised model at each time step and passes it through a logit-exponential-sum activation function defined as

$$a(x) = \frac{\sum_k \exp[l_k(x)]}{1 + \sum_k \exp[l_k(x)]} \tag{2.6}$$

where $l_k(x)$ is the logit input to the softmax function for class k. This activation function outputs values close to 0 for small or negative activations, and close to 1 for large and positive activations. The result of this stacked discriminator approach is that the supervised model is encouraged to make a clear class prediction on real time series but not on fake ones. Updates to both discriminators follow the approach proposed in the original RGAN framework, such that the average cross-entropy between RNN outputs and label sequences is minimized.

As discussed above, the discriminator model in the original RGAN algorithm is trained by applying a vector of all 1s for real time series and a vector of all 0s for generated, “fake” time series. In contrast, the label vectors provided to the supervised discriminator need not be assigned to a single value across time steps. In many classification settings, labels may vary across time, particularly if they are indicating occurrence of a particular event related to the input sequences. If labels vary across time, the time-variant label sequence can be provided to the supervised discriminator instead of a uniform label vector. For example, consider a patient with certain health measurements being recorded once per minute in an intensive care unit. If an adverse health event occurs for 30 minutes from times $t = 1000$ to $t = 1090$ out of a total of 1440 measurements, then the label sequence indicating the adverse event then consists of 0s from $t = 0$ to $t = 999$, 1s from $t = 1000$ to $t = 1029$, and 0s again through the end of the sequence. In settings where labels of interest do not vary across time, a uniform label vector is provided to both the supervised and unsupervised discriminators.

Figure 2.5 illustrates the data pipeline for training of semi-supervised RGANs (SS-RGANs) for classification when labeled time series are sparse. On each training iteration, the unsupervised discriminator and generator are updated exactly as they were in the original RGAN framework. The supervised discriminator attempts to classify labeled sequences from the small pool of labeled time series. In settings where only positive labels are known and the class distribution is heavily weighted towards the negative class, (e.g., a small set of patient timelines linked to drug overdoses), samples from the unlabeled sequences can be provided to the supervised discriminator as “noisy” negative class examples.

The proposed SS-RGAN requires implementation and training of three separate recurrent neural network models: the generator, the supervised discriminator, and the unsupervised discriminator. We implement all three of these models as long short-term memory networks (LSTMs), first described in Hochreiter and Schmidhuber (2011). An LSTM is a recurrent neural network made up of layers composed of memory cells. Three regulatory gates control the extent to which new information can flow into and out of each memory cell. The *input gate* controls whether new values are allowed into a cell; the *forget gate* controls whether a value is retained in a cells memory; and the *output gate* controls whether the value in a cell’s memory is used in computing the output activation of the memory cell. Together, the interactions of these regulatory gates with the flow of information passing through the network during training allows the LSTM to learn arbitrary long-term dependencies in the input sequences. LSTMs are trained through backpropagation, and are particularly well-suited for addressing the vanishing gradient problem that hinders training of traditional recurrent neural networks.

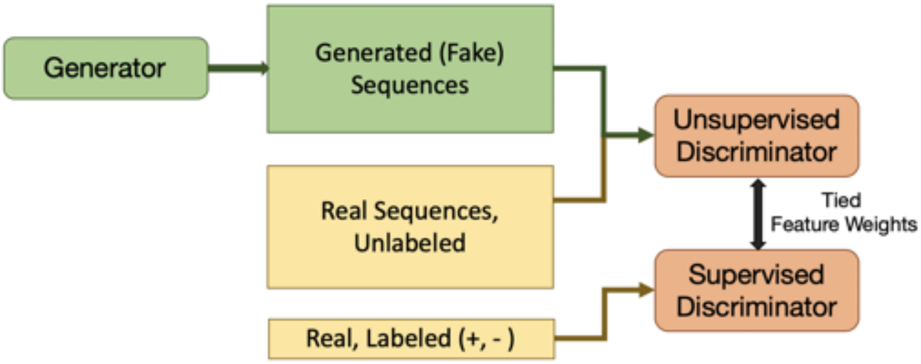

Generator
Generated (Fake) Sequences
Real Sequences, Unlabeled
Real, Labeled (+, -)
Unsupervised Discriminator
Tied Feature Weights
Supervised Discriminator

Fig. 2.5 Data pipeline for training of recurrent generative adversarial network for semi-supervised classification.

2.3.2 Evaluation on medical time series data

Before evaluating semi-supervised RGANs (SS-RGANs) on prediction of opioid misuse, we first define a set of benchmark classification tasks to iterate on model architecture and assess performance relative to other classification methods.

Data

From the publicly available Philips eICU database (Pollard et al., 2019), we select four variables measured by bedside monitors in Intensive Care Units (ICUs): heart rate (HR), respiratory rate (RR), oxygen saturation (SpO2), and mean arterial pressure (MAP). These four sets of time series provide a complex range of patterns on which to evaluate SS-RGANs on classification tasks. We obtained data on 192,831 total patients in the eICU database, downsampling to a single measurement every fifteen minutes for all four variables. Due to the size of the data set, we opted to drop any patients with missing values after downsampling. To help frame our benchmark classification tasks, we consider critical thresholds for each variable which indicate a potential adverse event in the ICU. We define the first 24 hours of a patient's stay as the observation period, and the subsequent 4 hours (hours 25-28 inclusive) as the prediction period. Binary labels for classification are determined for each patient based on whether they cross the critical threshold in the prediction period.

Evaluation Framework

We conduct two benchmarks for each ICU variable. The first benchmark task includes all patients, regardless of whether the critical threshold was crossed in the observation period. Note that this presents a relatively easy classification task for many positive examples, as a

Table 2.4 Description of four time series data sets measured by bedside monitors in hospital Intensive Care Units (ICU). Positive labels indicate critical threshold crossed during four-hour prediction period.

		All Patients		Non-Critical in 24-Hour Observation Period	
Measurement	**Critical Threshold**	**# Patients**	**% Positive Label**	**# Patients**	**% Positive Label**
Heart Rate	> 100	52576	33.0	22237	4.65
Respiratory Rate	< 13	36045	19.0	20555	4.60
Oxygen Saturation	< 95	32862	47.2	8535	11.8
Mean Arterial Pressure	> 110	4863	10.8	3472	4.20

patient who crossed the critical threshold one or more times in the observation period is much more likely to cross the threshold again in the prediction period. The second benchmark is restricted to only patients who do not cross the critical threshold in the observation period. This makes prediction more difficult, but provides a more realistic evaluation of model performance in a deployed setting. Table 2.4 shows a summary of each time series data set for the two benchmark tasks. The proportion of patients with positive labels varies considerably across variables and benchmark tasks, with significant class imbalance weighted towards the negative class in most cases.

As a point of comparison, we select three methods which have proven successful for supervised time series classification: the long short-term memory network (LSTM), the random forest (RandF), and the support vector machine with global alignment kernel (SVM-GAK) introduced in Cuturi (2011). For each benchmark task, we divide patients into a training and test set using a 75-25 split. To simulate the semi-supervised setting, we hold back 80% of the training set from supervised methods (but make the unlabeled time series from this held-back data available to the SS-RGAN). Hyperparameters for all methods are tuned through 10-fold cross-validation on the labeled training set. We evaluate all methods on area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC).

Results

ROC curves and precision-recall curves are shown for the heart rate benchmark tasks on all patients in Figures 2.6 and 2.7. Due to the imbalanced nature of the data sets, the precision-recall curves provide an appropriate assessment of prediction performance on the positive class. On the heart rate benchmark with all patients, SS-RGAN slightly outperforms the

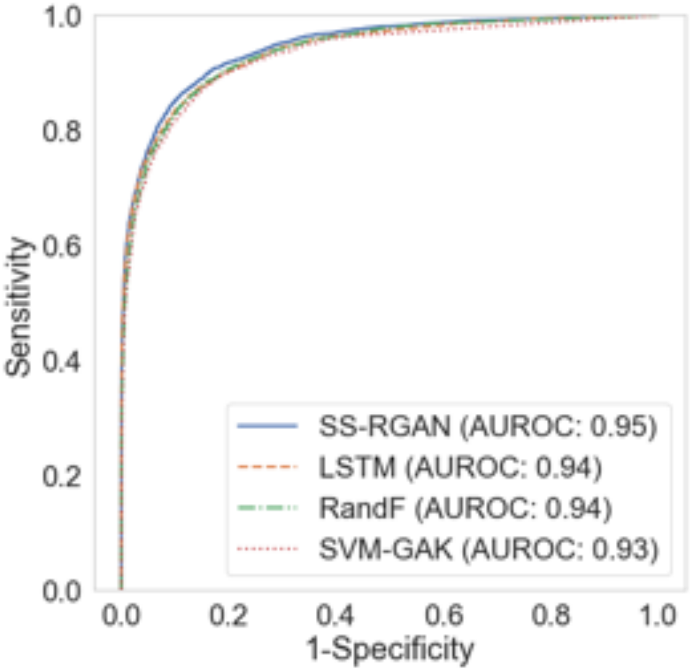


Fig. 2.6 Receiver Operating Characteristic (ROC) curves showing evaluation of binary classifiers on prediction of heart rate becoming critical, for all patients.

LSTM and RandF classifiers. On the heart rate benchmark restricted to non-critical patients in the observation period, the SS-RGAN demonstrates lower precision than LSTM and RandF at lower ranges of recall, but performs comparably to these methods at high recall ranges.

Table 2.5 and Table 2.6 provide AUROC and AUPRC across all benchmark tasks. For all patients, we observe that SS-RGAN gives slightly better performance than other methods on three of the four ICU variables considered, with the random forest winning on the mean arterial pressure experiment. For patients non-critical in the observation, SS-RGAN performs comparably to the best-performing methods, and performs particularly well relative to other methods on the “SpO2 < 95” task. Across all benchmarks, the SS-RGAN beats the SVM-GAK classifier and performs similarly to the LSTM and random forest.

2.3.3 Evaluation on opioid time series

This preliminary benchmarking analysis method on four medical time series data sets provides promising evidence that SS-RGAN is able to achieve classification performance comparable to benchmark supervised methods for time series classification. We next proceed with a full evaluation of SS-RGANs on the task of predicting unsafe levels of opioid prescription in Kansas PDMP data. We follow the same experimental framework outlined in Section 2.3.2 for medical time series classification, in which we generate binary labels from a critical

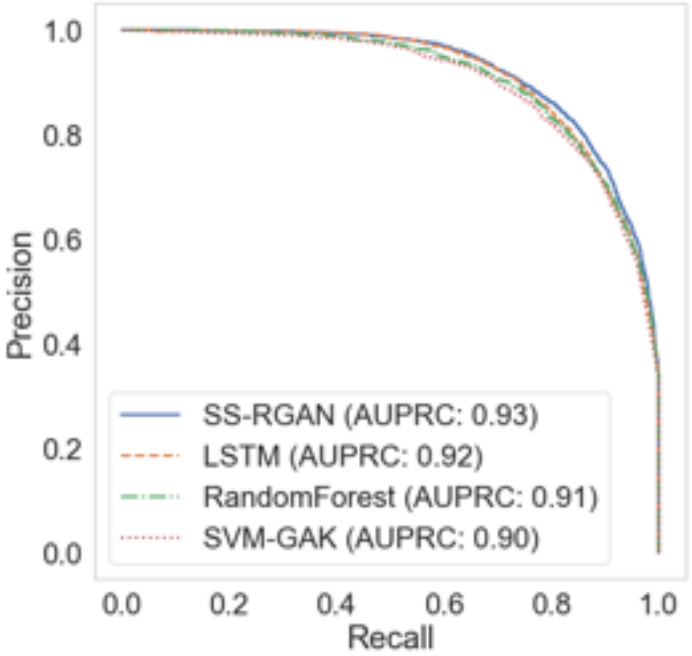


Fig. 2.7 Precision-recall curves showing evaluation of binary classifiers on prediction of heart rate becoming critical in ICU medical time series data.

Table 2.5 Performance of binary classifiers on prediction of ICU measurements becoming critical for all patients. Model performance is evaluated on area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC).

	HR > 100		**RR < 100**	
Classifier	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
SS-RGAN	**0.95**	**0.93**	**0.88**	**0.74**
LSTM	0.94	0.92	0.87	0.71
RandF	0.94	0.91	**0.88**	**0.74**
SVM-GAK	0.93	0.90	0.86	0.69
	SpO2 < 95		**MAP > 110**	
Classifier	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
SS-RGAN	**0.88**	**0.87**	0.80	0.47
LSTM	0.87	**0.87**	0.80	0.46
RandF	0.87	0.86	**0.82**	**0.48**
SVM-GAK	0.86	0.84	0.80	0.45

Best-performing classifier for each column denoted in bold face.

Table 2.6 Performance of binary classifiers on prediction of ICU measurements becoming critical for patients non-critical during observation period. Model performance is evaluated on area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC).

	HR > 100		**RR < 100**	
Classifier	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
SS-RGAN	0.81	0.23	0.65	0.09
LSTM	**0.82**	**0.27**	**0.68**	0.11
RandF	0.81	0.25	**0.68**	**0.12**
SVM-GAK	0.80	0.22	0.66	0.09
	SpO2 < 95		**MAP > 110**	
Classifier	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
SS-RGAN	**0.72**	**0.29**	0.61	0.09
LSTM	0.69	0.24	0.64	**0.13**
RandF	**0.72**	**0.29**	**0.67**	0.11
SVM-GAK	0.65	0.21	0.60	0.09

Best-performing classifier for each column denoted in bold face.

threshold applied to a prediction period late in patient timelines, simulating a small pool of labeled patients known to have poor outcomes.

Data

For evaluation of SS-RGAN on the predicting unsafe levels of opioid prescriptions, we rely on the same data set described in Section 2.2. MME timelines for patients represented in the Kansas PDMP provide multi-year time series for individuals that represent aggregate measure of total opioids being prescribed across different drugs and dosages. As in the shape-based clustering analysis, we preprocess the raw MME timelines by applying a 14-day moving average to smooth out spikes occurring when prescriptions overlap.

Evaluation Framework

The CDC recommends that clinicians employ extra caution when prescribing opioids in dosages greater than or equal to 50 MME per day, and to altogether avoid prescription of opioids for pain relief in dosages above 90 MME per day. A daily MME of 90 is the equivalent of nine 10/325 tablets of hydrocodone/acetaminophen, two 30 mg tablets of sustained-release oxycodone, or four 5 mg tablets of methadone. CDC guidelines note that sustained intake of opioids above this threshold introduces serious risk of addiction and overdose, without providing additional benefits in the form of chronic pain relief (Dowell et al., 2016). We therefore use this high-risk threshold to generate binary labels for classification that indicate unsafe levels of prescribed opioids. Daily MME observations in observation periods ranging from three to twelve months represent the primary inputs to prediction models. Binary labels are then defined based on whether a patient crosses the critical MME threshold of 90 MME per day in a three-month prediction period subsequent to the observation period.

To generate labels, the critical MME threshold is applied *after* MME time series have been smoothed with a 14-day moving average. Crossing the critical threshold therefore indicates that a patient has sustained an unsafe level of prescribed opioids over at least two weeks. Overlaps in prescriptions that may result in spikes in the raw MME timelines are unlikely to occur in lengths exceeding one week, thus the positive labels are likely to capture only those patients who are actually consuming unsafe levels of opioids across prescriptions or diverting prescription opioids for illicit resale.

We conduct two sets of experiments based on different criteria for patient inclusion in the analysis. For the first set of experiments, we include all patients in the MME time series data set, which includes all individuals present in the PDMP data set for at least two years after their first opioid prescription. These patients may have crossed the 90 MME/day critical threshold one or many times in the observation period, and therefore this initial set of experiments likely presents an easier prediction task for those individuals with sustained, unsafe MME levels in both the observation and prediction periods. The second set of experiments is constrained to only include patients who do not cross the 90 MME/day threshold in the three months before the prediction period. The constrained experimental setting provides a more difficult prediction task, but represents a more realistic evaluation of model performance in detection of patients who may be close to crossing the critical threshold but have not already been identified by rule-based red flag filters in the PDMP.

To accurately evaluate the performance of classifiers in a realistic setting, we create four separate pairs of training and test sets from the original MME data and report average performance across all test sets. Each test set is generated from a non-overlapping period following its corresponding training set, and prediction periods in the four test sets do not overlap each other in time. Training and test sets include all MME timelines in the original data set with an opioid prescription in the first three months of the observation period, thus all data sets include patients with at least nine months of observations after their first opioid prescription.

The four time series data sets used for opioid prediction experiments are described in Table 2.7. In Data Set (1), the training set is generated from observations in months 1 through

Table 2.7 Description of time series data sets designed to evaluate prediction unsafe levels of morphine milligram equivalents. Positive labels indicate critical threshold crossed during three-month prediction period. Statistics are provided for two groups of patients: (1) all patients with an opioid prescription in the first three months of the observation period, and (2) the subset of those patients who do not cross the critical threshold in the three months before the prediction period.

	All Patients			**Non-Critical Patients**		
Data Set	n_{train}	n_{test}	% Positive	n_{train}	n_{test}	% Positive
(1)	97,488	99,163	2.6%	92,706	94,281	2.0%
(2)	98,712	96,571	2.5%	93,739	91,789	1.9%
(3)	101,509	92,836	2.1%	95,937	86,456	1.8%
(4)	98,813	85,586	2.3%	93,125	81,805	1.9%

12 of the three-year study period inclusive, and the test set is generated from observations in months 13 through 24 inclusive. In Data Set (2), the training set is generated from months 4 through 15, and the test set is generated from months 16 through 27. In Data Set (3), the training set is generated from months 7 through 18, and the test set is generated from months 19 through 30. Finally, in Data Set (4), the training set is generated from months 10 through 21, and the test set is generated from months 22 through 33. For all training and test sets, critical threshold labels are generated from the three months following the end of the observation period.

As in the previous section evaluating SS-RGAN on four medical time series prediction tasks, we compare prediction performance against three supervised methods that have demonstrated high performance on time series classification tasks: LSTM, RandF, and SVM-GAK. To simulate a semi-supervised setting, we provide 5% of the original training set to the three supervised methods. SS-RGAN, as a semi-supervised approach, has access to both the labeled time series (5% of original training data set) and the unlabeled time series (95% of original data set). Hyperparameters for all methods are tuned through 10-fold cross-validation on the labeled time series. We first report results for all methods on traditional performance metrics for binary classifiers in the machine learning literature: area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC).

We additionally consider a realistic setting for evaluation that aligns more closely with how prediction models would be implemented by public health agencies to identify and intervene on high-risk patients. We consider two resource-constrained scenarios, in which health agencies must choose a subset of patients to prioritize and target with assistance programs, such as routine phone check-ins, referrals to medication-assisted treatment programs for substance use disorder, or referrals to mental health treatment programs. We report precision and recall under both low capacity and high capacity constraint scenarios, which indicate how well the prediction models under consideration can identify high-risk patients across a spectrum of realistic outreach program scales.

Results

AUROC and AUPRC are reported in Table 2.8 for all four classification methods on the unconstrained set of patients in the PDMP data set. Individuals in the raw PDMP naturally have histories of differing length depending on how long they have been administered prescription drugs in the state of Kansas. Patients who have started opioid prescriptions only recently or who just moved into the state will have few records with which to generate MME timelines. Ideally, a classification method would not require a long observation period in order to make accurate predictions about risk of unsafe opioid prescription. We report performance on observation periods of three different lengths (three months, six months, and twelve months), averaged across the four test sets, to illustrate how performance of the methods considered varies according to the length of the time series provided as model inputs.

ROC curves and AUROC illustrate predictive performance on both the positive and negative classes, but are strongly weighted towards performance on the majority class in heavily imbalanced settings. Due to the high class imbalance inherent to this prediction task, precision-recall curves and the AUPRC provide the clearest picture for performance in predicting the positive class (unsafe levels of opioid prescription). We find that across all three observation period lengths, SS-RGAN outperforms competing methods on AUPRC. The LSTM, RandF, and SVM-GAK models achieve comparable performance to each other, with RandF performing particularly well in settings with a shorter observation period (three months). As measured by AUPRC, the performance of SS-RGAN is not hindered by shortened observation periods, and in fact the AUPRC is slightly higher on the shorter observation periods than the twelve month observation period. With respect to AUROC, the LSTM performs well across all three observation periods, but its performance on precision-recall indicates that the LSTM is doing worse at identifying the low-frequency positive class members. ROC curves and precision-recall curves on one of the four test sets considered (Test Set (4)) are shown in Figure 2.8, where all models are provided sequences from a twelve month observation period.

Table 2.9 shows average AUROC and AUPRC for the four classification methods on the subset of patients who did not cross the 90 MME/day critical threshold in the three-month

Table 2.8 Performance of binary classifiers on prediction of opioid time series crossing critical threshold for all patients in prescription drug monitoring program. Model performance is evaluated on area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) and averaged across four test sets.

	3 month Obs. Period		**6 month Obs. Period**		**12 month Obs. Period**	
Classifier	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
SS-RGAN	0.93	**0.86**	**0.92**	**0.87**	**0.94**	**0.85**
LSTM	**0.94**	0.83	**0.92**	0.85	0.92	0.83
RandF	0.90	0.85	0.89	0.85	0.88	0.81
SVM-GAK	0.89	0.83	0.87	0.85	0.85	0.81

Best-performing classifier for each column denoted in bold face.

period leading up to the prediction period. As expected, prediction performance is lower across the board when compared with the easier prediction task that includes all patients. Still, SS-RGAN outperforms all other methods with respect to AUPRC on the constrained set of patients, with an average area under the precision-recall curve ranging from 0.43 to 0.46 across the three observation periods. The RandF performs well given three months of input data, tying the SS-RGAN on AUPRC. As in the previous setting, the LSTM performs well on AUROC, but suffers on the more relevant metric of AUPRC. ROC curves and precision-recall curves on Test Set (4) are shown in Figure 2.9.

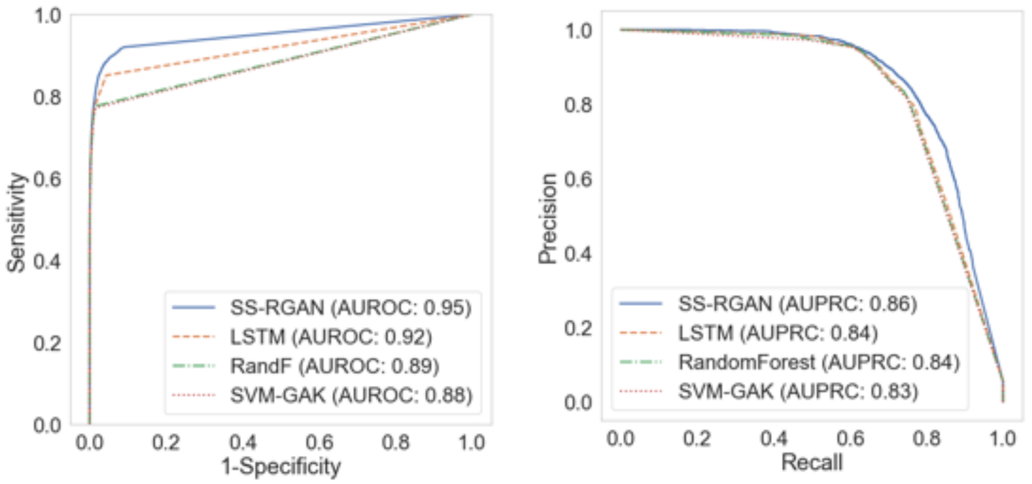
The length of the observation period did not appear to have a significant impact on the relative performance of competing methods, as AUROC and AUPRC typically varied by 1-2 percentage points across period lengths. We therefore focus on the 12-month observation period for the remaining analysis, and note that results and relative performance is unlikely to be meaningfully different for observation periods of shorter length. We also focus on the constrained subset of patients who did not cross the critical threshold in the three months leading to the prediction period, as this provides the more useful benchmark for real-world use cases.

While ROC and precision-recall curves can provide an appealing and interpretable visual comparison of prediction methods, they do not necessarily reflect relative performance of methods in realistic settings where humans use predictions to target intervention. For example, public health agencies may have limited resources with which to engage high-risk patients with voluntary assistance programs or outreach. In this setting, the top-ranked predictions from each model are therefore much more relevant to actual real-world performance than all other predictions made by the model. Performance curves which assess *all* predictions in a single curve or area metric may obscure the performance of the top-ranked predictions.

Table 2.9 Performance of binary classifiers on prediction of opioid time series crossing critical threshold for patients that were not critical in the three months prior to the prediction period. Model performance is evaluated on area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) and averaged across four test sets.

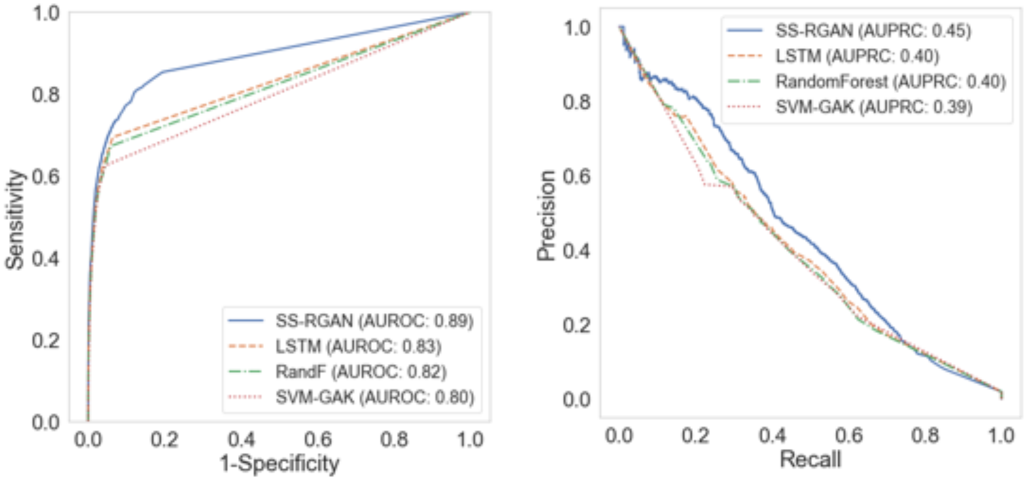
	3 month Obs. Period		**6 month Obs. Period**		**12 month Obs. Period**	
Classifier	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
SS-RGAN	0.85	**0.45**	**0.85**	**0.46**	**0.87**	**0.45**
LSTM	**0.86**	0.43	**0.85**	0.44	0.85	0.41
RandF	0.79	**0.45**	0.81	0.45	0.80	0.42
SVM-GAK	0.78	0.42	0.80	0.44	0.79	0.40

Best-performing classifier for each column denoted in bold face.



(a) **(b)**

Fig. 2.8 Receiver Operating Characteristic curve and Precision-Recall curve showing evaluation of binary classifiers on prediction of morphine milligram equivalents becoming critical for all patients with 12 months of observations in prescription drug monitoring program.



(a) **(b)**

Fig. 2.9 Receiver Operating Characteristic curve and Precision-Recall curve showing evaluation of binary classifiers on prediction of morphine milligram equivalents becoming critical for patients who were not critical in the three months prior to the prediction period

To assess the four prediction methods in a realistic setting, we report precision at k and recall at k for two different resource constraint scenarios, averaged across four test sets (Table 2.10). The low-capacity scenario simulates a setting in which resources exist for providing assistance or outreach to 100 patients, which make up approximately 0.1% of the total number of patients in the test sets considered. Out of the top 100 patients predicted by the SS-RGAN on the test set, 87 of them did in fact cross the 90 MME/day threshold during the prediction period, giving an average precision at 100 patients of 0.87 across test sets. The LSTM represents the next-best competitor, with precision of 0.83 at 100 patients averaged across test sets. This four-percentage-point difference represents an additional four patients who would have been correctly identified by the SS-RGAN as requiring particular attention due to high-risk of unsafe opioid prescription, but who would have been missed altogether by all other prediction methods considered here. We also consider recall at 100 patients under the low-capacity scenario. SS-RGAN captures a slightly higher proportion of all positive labels (patients crossing the MME threshold) relative to the other methods, and with a recall of 0.06 the SS-RGAN approaches the best possible recall under this particular resource constraint.

Under the high-capacity resource constraint scenario, we report precision and recall at the top-ranked 1000 patients for each prediction method. The four test sets include between 1500-1900 individuals with positive labels, so the high capacity scenario can be considered a benchmark for a program that is scaled up to approximately one-half to two-thirds the size of the estimated proportion of relevant individuals in the population. Under the high-capacity scenario, the SS-RGAN still outperforms all other methods, with a precision of 0.58 and recall of 0.37 for the 1000 top-predicted patients. Scaling up a targeted intervention in this way allows a public health agency to capture more of the relevant, high-risk population (higher recall), at the expense of a lower proportion of predicted individuals actually crossing the critical threshold (lower precision). Policymakers at public health agencies should carefully consider this tradeoff between precision and recall (true positive rate) when determining the appropriate scale of a targeted intervention that relies on model predictions.

The results of the low-capacity resource scenario provide particularly compelling evidence that SS-RGAN can successfully leverage large amounts of unlabeled data to outperform purely supervised classification methods on predicting unsafe levels of MME, providing a four-percentage-point improvement on precision at 0.1% of patients targeted. Still, a statistical analysis of the *difference* between SS-RGAN and other methods is useful to understand whether the method is providing a statistically significant improvement over competitors across different evaluation settings.

We follow the approach proposed in Dietterich (1998) for comparing the performance of two classifiers with a 5x2 cross-validated paired t-test. In this approach, we randomly divide the evaluation set into two splits of equal size, and train on one split while evaluating the difference between prediction methods on some performance metric (e.g., precision at k patients or recall at k patients) on the other split. The two halves are then rotated (the training set becomes the test set, and the test set becomes the training set), and the difference between methods is averaged across both rotations and variance of differences across two rotations is computed. This procedure is repeated for five iterations, giving five values for the variance in differences between methods. The test statistic is computed as the difference on the *initial* split and rotation divided by the square root of the average variance across five iterations. We assume that the resulting test statistic approximately follows a t-distribution with 5 degrees of freedom under the null hypothesis that the two prediction models have equal performance. We can then obtain p-values which indicate whether the two methods provide significantly different results on evaluation metrics considered.

Table 2.11 show the results of 5x2 cross-validation in conjunction with paired t-tests comparing SS-RGAN to the next-best classification model (LSTM) under constrained resource evaluation settings, for a single test set (Test Set (4)). We find that on both precision and recall at 100 patients, SS-RGAN provides a statistically significant improvement over the LSTM at $\alpha = 0.05$). Under the high-capacity constraint scenario, the difference between SS-RGAN and LSTM is less pronounced; we therefore find that the improvement from SS-RGAN is

Table 2.10 Performance of binary classifiers on prediction of opioid time series becoming critical during prediction period under two resource constraint scenarios. Model performance is evaluated on precision and recall at the top 100 predicted patients (low capacity) and top 1000 predicted patients (high capacity).

	Low Capacity		**High Capacity**	
Classifier	**Precision @100 Patients**	**Recall @100 Patients**	**Precision @1000 Patients**	**Recall @1000 Patients**
SS-RGAN	**0.87**	**0.055**	**0.58**	**0.37**
LSTM	0.83	0.053	0.53	0.34
RandF	0.80	0.051	0.52	0.33
SVM-GAK	0.77	0.049	0.52	0.33
Best Possible	1.0	0.063	1.0	0.63

Best-performing classifier for each column denoted in bold face.

Table 2.11 Statistical comparison of SS-RGAN performance metrics to best performing alternative, the LSTM. P-values are computed using 5x2 cross-validation with modified Student's t-test.

Performance Metric	**P-value**
Precision @ 100 Patients	0.03**
Recall @ 100 Patients	0.04**
Precision @ 1000 Patients	0.08*
Recall @ 1000 Patients	0.08*

** denotes statistical significance at $\alpha = 0.05$. * denotes statistical significance at $\alpha = 0.10$.

not significant at $\alpha = 0.05$, with p-value of 0.08 for precision and recall respectively at 1000 patients predicted.

Introducing Conditional Inputs

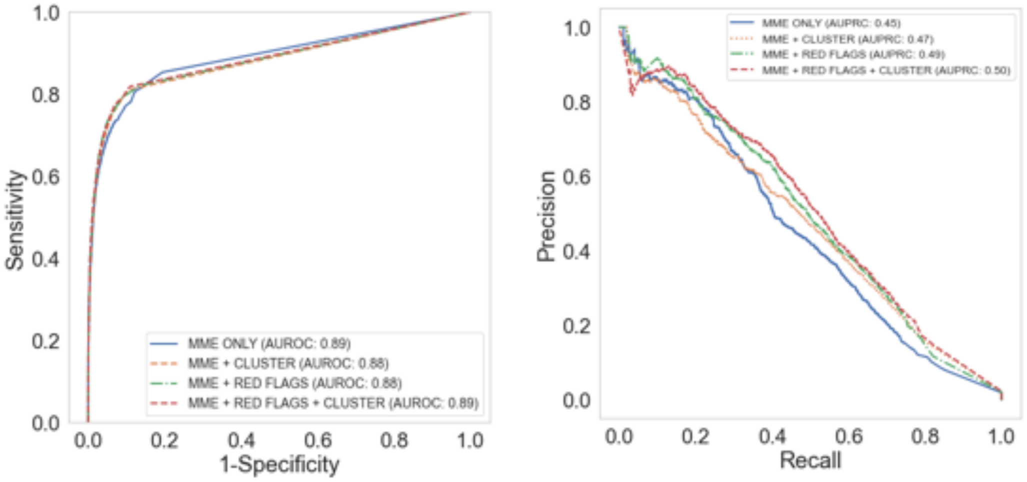
Hyland et al. (2017) describe and evaluate an approach to augmenting RGANs with additional conditional inputs. In recurrent conditional generative adversarial networks, the primary inputs to the generator and discriminator LSTM networks are augmented with conditional information at each time step. These conditional inputs are concatenated with the primary input sequences at each time step. We follow the same approach to providing additional, time-varying covariates to all three underlying networks in the SS-RGAN, so that each may leverage relevant covariates to improve on respective tasks (i.e., generating realistic time series, distinguishing fake from real time series, and classifying time series as high-risk or low-risk).

On the task of predicting unsafe levels of opioid prescription, we consider two different sources of additional information which are relevant to MME trajectories and may therefore improve prediction performance. First, we consider the set of red flag indicators suggesting unsafe prescription drug practices described in Section 2.2.1: (1) greater than two simultaneous opioid types, (2) greater than one simultaneous opioid prescriber, and (3) benzodiazepine and opioids prescribed simultaneously. We generate time series for each red flag, where each observation indicates whether the flag is raised at that particular time step. Next, we consider the set of shape-based time series clusters identified in 2.2.1, and include the sequence of cluster assignments for each patient according to the nearest cluster centroid at each point in time. We consider impacts to SS-RGAN performance when including both the red flag inputs and cluster inputs individually, and finally when including all sets of conditional inputs.

ROC and precision-recall curves for different conditional input combinations are shown in Figure 2.10. Table 2.12 provides a comparison of the baseline SS-RGAN performance (with no conditional inputs) to SS-RGAN models trained with conditional inputs provided, under the same resource constraint scenarios described in the previous section. We find that both the cluster assignments and red flag indicators improve model performance slightly over the MME-only baseline, and providing both sets of conditional inputs gives the best performance observed in any experiments thus far on both precision and recall under both constraint scenarios. These results on prediction of unsafe opioid levels provide promising evidence that the SS-RGAN can leverage time-varying conditional inputs to improve prediction performance on the classification task of interest.

Sensitivity to Single-Class Labeling

We note that in real-world public health settings, we may have access to only a small pool of patients with positive labels, and no patients that are linked to negative labels. For example, we may have a small set of patient timelines associated with known drug overdose deaths, but no reliable information on which patients dropped out of the PDMP for other reasons but did not die of drug overdose. In the case where only positive labels are known, we observe that samples from the pool of unlabeled sequences can be provided to the SS-RGAN supervised discriminator as “noisy” negative class examples. This approach assumes that the class distribution is heavily weighted towards the negative class, so the supervised discriminator will still be able to learn even with some fraction of incorrect labels in the negative class. We simulate this setting by removing all time series with negative labels from the labeled data set provided to the supervised discriminator. At each training iteration, a random set of



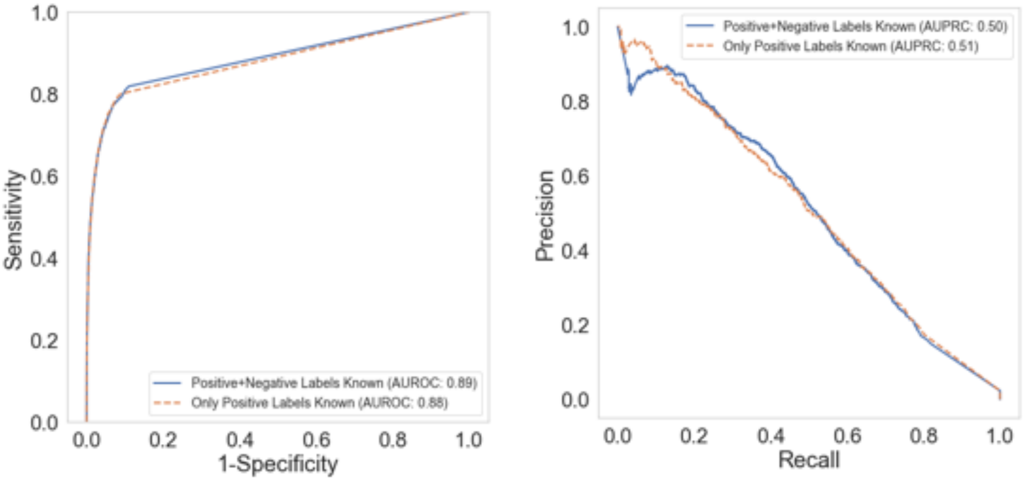
(a) **(b)**

Fig. 2.10 Comparison of prediction performance from prescription timeline features (red flag indicators and nearest MME trajectory cluster centroid) provided as conditional inputs to SS-RGAN on opioid prediction task.

Table 2.12 Comparison of performance from prescription timeline features (red flag indicators and nearest MME trajectory cluster centroid) provided as conditional inputs to SS-RGAN on opioid prediction task under two resource constraint scenarios.

	Low Capacity		**High Capacity**	
Classifier	**Precision @100 Patients**	**Recall @100 Patients**	**Precision @1000 Patients**	**Recall @1000 Patients**
MME Time Series (No Conditional Inputs)	0.87	**0.06**	0.58	0.37
+High-Risk Cluster Labels	0.87	**0.06**	0.59	0.38
+Red Flag Indicators	**0.88**	**0.06**	0.59	0.38
+All Conditional Inputs	**0.88**	**0.06**	**0.60**	**0.39**
Best Possible	1.0	0.06	1.0	0.63

Best-performing classifier for each column denoted in bold face.



(a) **(b)**

Fig. 2.11 Comparison of prediction performance for two labeling procedures in training data, simulating settings where only positive labels are known during training time.

time series are selected from the unlabeled pool of time series and assigned negative class labels. The supervised discriminator is then updated as usual. Figure 2.11 shows ROC and precision-recall curves for the baseline SS-RGAN model and the SS-RGAN trained with only positive labels. We find that this alternative labeling scheme gives almost identical results to the original labeling scheme, due to the extremely low prevalence of true positive examples in the unlabeled data pool.

2.4 Discussion

The analyses presented in presented in Section 2.2 and 2.3 provide important preliminary evidence that both unsupervised and semi-supervised machine learning methods can help human decision makers learn from patterns in PDMP data and guide decisions around which patients to target with additional intervention for reducing risk of opioid misuse. In settings where PDMP data represent the only source of information at hand, clustering approaches can be used to improve upon rule-based red flag filters that are currently used by public health agencies to identify high-risk patients. Targeted interventions can additionally benefit from prediction models even when patient outcomes are known for only a small set of individuals.

Prescription drug monitoring programs represent a rich source of information on prescription practices that have not yet been leveraged to their full potential for identifying and intervening on high-risk patients. PDMPs already provide a critical source of information to clinicians and other prescribers of opioids, representing the primary means by which doctors can avoid over-prescription of opioids to patients with obvious indicators of unsafe drug use. Outside of the clinical setting, PDMPs can provide public health agencies with valuable information in targeting assistance programs designed to monitor and provide help to the highest-risk patients. For example, routine phone check-ins on high risk patients represent a low-cost intervention that may have significant impacts on patient health, simply by providing regular reminders of safe practices around prescription opioid use. Similarly, referrals to voluntary, medication-assisted treatment programs can significantly reduce barriers to access for patients with opioid use disorder (Scott et al., 2019), and these outreach efforts may be targeted using prediction models trained on PDMP data. While selection into medication-assisted treatment programs or voluntary mental health treatment programs should never be made based *solely* on the outputs of a prediction model, these outputs nonetheless provide a way for human decision makers to distill the vast amount of information collected in the PDMP and leverage that information for better interventions.

A limitation of this study is that project partners were not able to provide access to linked data sets indicating patient outcomes related to opioid use, such as drug overdose deaths. We therefore relied on signals present in the primary PDMP data set to simulate settings where a small pool of patients with poor outcomes is known, even though these labels can be generated for any MME time series in the data set. A worthwhile research direction would be to replicate this study using separate data sets on hand at public health agencies to identify known, high-risk patients in the MME data set for the pool of labeled examples provided to the semi-supervised method proposed.

2.5 Conclusions

The results of this study have laid important groundwork indicating that machine learning models can provide useful outputs to help humans target interventions that provide assistance to individuals at risk of opioid overdose or addiction. Still, research gaps remain to be filled before the methods proposed in this study are deployed in an operational setting for targeting intervention. Critically, the next step in the research pipeline is to evaluate an intervention program in which predictions from machine learning models are integrated with existing models for providing public assistance and outreach. Future work would involve evaluating a human-in-the-loop system for targeting intervention, in which outputs from machine learning models provide an additional source of information and insight to public health specialists or clinicians, ideally with a randomized, controlled field trial.

The findings discussed in this study provide promising evidence that data available to public health agencies from prescription drug monitoring programs can be useful for detecting patients at risk of opioid misuse and directly predicting unsafe practices related to opioid prescription. In settings where patient outcomes are unknown, unsupervised approaches can help identify early warnings for patients with only a small number of records in a PDMP database. When patient outcomes are known for a small set of patients, semi-supervised approaches such as the SS-RGAN can leverage the large pool of PDMP data records in order to make predictions about a patient's future risk. We look forward to continued collaboration with project partners to understand what outputs from a risk assessment tool based on PDMP data would be most useful for public health workers and clinicians. We hope that continued research in this area will result in new techniques for prescription monitoring to mitigate the harms caused by the opioid epidemic while enabling safe provision of pain relief for those dealing with chronic pain.

Chapter 3

Policing Chronic and Temporary Hot Spots of Violent Crime: A Controlled Field Experiment[1]

3.1 Introduction

In recent years, police agencies have demonstrated an increased willingness to devote resources toward proactive policing strategies that target underlying causes of crime. In contrast to reactive policing, which prioritizes responding to 911 emergency requests for police response and investigating crimes that have already occurred, proactive policing instead aims to identify predictable patterns in which crimes typically occur or the underlying causal factors that lead to crimes of various types. Most recently, in the United States, discussion around the potential benefits and harms of proactive policing has become part of a wider debate about the role of police and their impacts on the communities they serve, sparked by multiple instances of police brutality against minorities and the resulting widespread protests against systemic racism in policing.

Here we present results from a controlled experiment of proactive policing which contributes several pieces of evidence to this debate, quantifying the crime prevention benefits of proactive patrols targeting predicted "hot spots" of serious violent crime, and confirming that certain undesirable outcomes (over-policing arrests of racial minorities) did not occur in the study period. We emphasize that these impacts are critically dependent on the community-oriented approach taken by police in response to predicted hot spots, while aggressive policing practices may substantially harm individuals and communities. In the

[1] This chapter is based on the research paper of the same title, co-authored by Wilpen Gorr and Daniel B. Neill.

discussion below, we note other critical issues to be addressed while evaluating the potential impacts of proactive policing, as well as describing a set of best practices intended to maximize benefits and minimize harms.

A well-established finding in the criminology literature is that crimes do not occur uniformly across time and space, but instead concentrate in micro-places, or "hot spots." Characterizing crime patterns at a fine resolution across time and space allows police agencies to identify areas that are most in need of targeted intervention (National Research Council, 2004). Hot spots range from single intersections or storefronts to areas encompassing a few city blocks and represent optimal locations for committing crimes (Block and Block, 1995; Brantingham and Brantingham, 1999; Eck and Weisburd, 1995; Sherman et al., 1989; Weisburd, 2015). Crime hot spots make up a tiny fraction of the total area or street networks in a city, yet tend to produce disproportionately many calls for police response and subsequent reports relating to a range of violent, disorder, and property crimes. Routine activity theory provides a theoretical characterization of hot spots as locations where likely offenders and suitable targets (individuals or their property) converge in time and space in the absence of capable guardianship (Cohen and Felson, 1979). For example, a crime hot spot might occur near an automated teller machine (ATM) in an economically depressed commercial area, where potential offenders can remain without appearing suspicious to police or shop owners and where pedestrians are known to be carrying cash after visiting an ATM.

Police cannot effectively patrol all parts of the city at risk of experiencing crime. Police command staff must weigh resource and manpower constraints when making decisions about how to distribute patrols throughout a city and which proactive patrol activities are appropriate in a given time and place. The day-to-day tactical decisions relating to patrol management have observable impacts on the prevalence of crime and disorder in patrolled neighborhoods (National Academies of Sciences, Engineering, and Medicine, 2018). A substantial body of evidence indicates that sending proactive patrols to crime hot spots can successfully lead to reductions in crime in those areas, but the amount of crime reduction exhibited substantial variation across studies to date (Braga et al., 2019).

Enforcement actions such as stops, searches, and arrests can prevent crime by directly incapacitating potential offenders (Weisburd and Eck, 2004). Incapacitation may have immediate effects on crime (Wyant et al., 2012) or longer-term effects if prolific offenders are removed from a community. However, it has been shown that aggressive policing practices, including frequent stops, summonses, and arrests for low-level crimes, have adverse impacts on community health (Geller et al., 2014), police-community relations (National Academies of Sciences, Engineering, and Medicine, 2018), and racial equity. Minorities can suffer from "over-policing" with aggressive enforcement actions. An example is the "broken windows" approach to policing that advocates zero tolerance by police to disorder behavior in the belief that it would reduce crime. Strict enforcement of disorder crimes leads to disproportionate harassment and arrest of minority persons and erodes public trust in police (Kamalu and Onyeozili, 2018).

Guardian actions, such as increased police presence through patrols and community-policing methods, deter crime by removing opportunities to commit crime and increasing the perceived risk of crime commission, even though patrols generally do not involve direct contact with potential offenders (Loughran et al., 2011; Sherman and Weisburd, 1995). Koper (1995) presents evidence of short-term residual benefits from patrols after police leave an area, finding that patrols to high-crime locations meeting a threshold dosage of about 10 minutes achieve a general deterrence effect which persists for several hours. Spatial diffusion of patrol benefits is also supported by the literature, with several studies finding that the general deterrent effect from police presence diffuses into nearby areas where police were not concentrating efforts (Clarke and Weisburd, 1994; Telep et al., 2014; Weisburd et al., 2006). Piza (2018) recently conducted a study of foot patrols, examining crime reduction of enforcement actions versus guardian actions, and found that only guardian actions have statistically significant effects. Similarly, a meta study on guardian actions in the form of community-oriented policing and problem solving versus enforcement actions in hot spots found only guardianship to have statistically significant effect sizes (Braga and Schnell, 2015).

Hot spots represent locations where police interventions have the greatest potential for crime deterrence, but how to best identify and characterize these areas remains an open question. Despite a large number of studies examining the effectiveness of hot spot-based policing programs, most focus on the impact of patrolling a fixed set of hot spots pre-selected using crime density maps or expert knowledge from police. Growing evidence shows that hot spots can be dynamic in nature, with some crime clusters exhibiting changes in shape, location, or magnitude over the course of days or weeks (Gorr and Lee, 2015, 2017; Herrmann, 2015; Mohler et al., 2015). Although human crime analysts can typically pinpoint areas with chronically high levels of crime with relative ease, subtle or emerging changes in crime patterns can be harder to detect. Data-driven predictive analytics provide an alternative strategy for hot spot selection. Crime forecasting models based on up-to-date crime data can respond extremely quickly to the rapidly changing landscape of crime patterns across time and space, and therefore provide a promising direction for dynamic selection of hot spots.

There have been relatively few studies which evaluate responsive selection of hot spots through predictive modeling, but two evaluative studies of predictive policing programs bear discussion. Hunt et al. (2014) present an evaluation of a predictive policing program implemented by the Shreveport, Louisiana Police Department in 2012. The program used multivariate logistic regression models to predict small areas with increased risk of property crimes, then deployed police patrols to conduct interventions at areas with highest predicted risk over a seven-month period. The predictive policing program was evaluated by comparing property crime rates in three treatment districts to three control districts where hot spots were chosen using conventional crime mapping approaches. Hot spots were selected each month and discussed daily with patrol officers during roll call. The evaluation found no statistically significant difference in crime rates between treatment and control districts, which the authors attribute to low statistical power of tests and significant variations in proactive patrol implementation across districts and over the course of the experiment.

In contrast, Mohler et al. (2015) do find a statistically significant reduction in crime volume within patrolled hot spots in an evaluation of crime forecasting using a self-exciting point process (SEPP) for predictions (Mohler et al., 2011) in Los Angeles, CA. Predictions were evaluated by comparing hot spots selected by the SEPP model with hot spots selected by trained crime analysts. Police patrols were deployed to predicted hot spots for both treatment and control hot spots. The Los Angeles field experiment indicated that relative to hot spots selected by crime analysts, hot spots generated by the prediction model experienced an average 7.4% reduction in crime volume per week at mean patrol levels. Within treatment areas, the authors also observed a statistically significant negative relationship between patrol time and crime volume, indicating that increased patrol time is more beneficial within hot spots selected by the SEPP model.

Little work has been done to rigorously compare the effectiveness of multiple forecasting methods on the task of hot spot selection for proactive patrols. In this study, we present findings from (1) an empirical comparison of crime prediction methods on the task of one-week-ahead crime prediction, and a (2) controlled field study evaluating a hot-spot-based predictive policing program in Pittsburgh, PA. Section 3.2 compares the performance of several place-based forecasting models on predicting historical crime data for the purpose of selecting prediction models for the Pittsburgh field study. Section 3.3 describes the design, implementation, and results of a controlled field study conducted in partnership with the Pittsburgh Bureau of Police (PBP). General implications for predictive policing are discussed Section 3.4. We end with concluding remarks and ideas for future research directions in Section 3.5.

3.2 Empirical Comparison of Prediction Methods for Crime Forecasting

A hot-spot-based policing program inherently relies on the ability of police to identify crime hot spots correctly and in a timely manner, so that proactive patrols can be dispatched to areas most in need of intervention. The forecasting literature provides a wide array of options for predictive methods which may be applied in the criminology setting to varying degrees of success. Small tweaks to model specifications and parameters can have significant implications on the performance of the method for the crime prediction task, even a within a single class of predictive methods. For application in a predictive policing program, hot spot selection models can have competing goals, such as high prediction accuracy and spatial dispersion of predicted areas. To understand the tradeoffs of various prediction methods and model specifications and design an appropriate process for weekly hot spot selection, we evaluated the performance of a set of prediction models on five years of historical crime data obtained from the PBP. Results from this analysis directly informed the choice and design of prediction models used in the field study of a hot-spot-based predictive policing program in Pittsburgh.

3.2.1 Data

To simulate a setting in which police command staff and crime analysts make weekly decisions about where to distribute proactive patrols, we relied only on data sources readily available to PBP for one-week ahead crime forecasts. Specifically, we obtained two data sets spanning five years of historical data from June 1, 2011 through June 1, 2016. The first data set is compiled from the City of Pittsburgh's Automated Police Reporting System (APRS), and contains data on all 206,150 crime incidents recorded by the PBP within the five-year period of analysis. For each incident, the data set contains an associated crime code corresponding to the Uniform Crime Reports (UCR) hierarchy employed by the Federal Bureau of Investigation (FBI). Criminal offenses are divided into two primary groups in the UCR hierarchy; Part I offenses represent the most serious categories of crime that are likely to be reported to police, and Part II crimes include less serious offenses. Part I crimes are further divided into two categories: violent crimes (P1V) and property crimes (P1P). Table 3.1 reports crime counts for the seven component crime types which make up all Part 1 offenses in the APRS data. Though arson is also categorized by the FBI as an eighth Part 1 offense, data on arson incidents were not provided by the PBP. Analysis of crime incidents

Table 3.1 Counts of APRS crime incidents for all Part 1 offenses from June 1, 2011 through June 1, 2016 in Pittsburgh, PA.

	UCR Code	Crime Type	Frequency
Part 1 Violent (P1V) Offenses	1	Criminal Homicide	296
	2	Forcible Rape	991
	3	Robbery	5822
	4	Aggravated Assault	5856
Part 1 Property (P1P) Offenses	5	Burglary	11665
	6	Larceny/Theft	37230
	7	Vehicle Theft	3870

focused on the seven crime types reported in Table 3.1, with P1V crimes ultimately selected as the primary target for prediction.

The second data set provided by PBP includes information on geotagged 911 calls for assistance to the Pittsburgh police, totaling approximately one million calls over the five-year period of analysis. This computer-aided dispatch (CAD) data includes the time and place that a call was made, as well as a code indicating the reason the call was made and a descriptive field indicating the outcome of the call. Taken together, the APRS data and CAD data provide a comprehensive picture of when and where crimes are being reported in Pittsburgh. An important caveat is that crimes that go unreported to police are not represented in these data sources. This analysis therefore provides an evaluation of prediction methods at forecasting crimes *as they are reported to police*, with the purpose of simulating one-week-ahead forecasts of reported crime made by police analysts.

3.2.2 Methodology

To evaluate the effectiveness of various prediction models for hot spot selection, we divided the five years of historical data into a two-year model calibration period (June 1, 2011 through June 2, 2013) and a three-year evaluation period (June 3, 2013 through June 1, 2016). P1V crimes comprise the most serious, violent offenses occurring throughout the city and are therefore considered a top priority for crime deterrence efforts by police. We conducted rolling one-week-ahead forecasts of P1V crime counts over the entire study period to simulate crime analysts making weekly forecasts for hot spot selection. Input features provided to the models varied by model, but can include lagged counts of the target variable, and lagged counts of various "leading indicator" variables. Leading indicators represent events which may be predictive of the outcome variable, such as other crime types or categories of 911 calls. The set of leading indicator variables provided as inputs for multivariate models are

Table 3.2 Leading indicator variables provided as inputs to multivariate prediction models.

Data Source	Leading Indicator Variable
Automated Police Reporting System (APRS)	Criminal Homicide
	Forcible Rape
	Robbery
	Aggravated Assault
	Burglary
	Larceny/Theft
	Vehicle Theft
	Simple Assault
	Vandalism
	Liquor Law Violations
	Public Drunkenness
	Disorderly Conduct
	Criminal Mischief
	Trespass
Computer-Aided Dispatch (CAD)	Assault
	Burglary
	Criminal Mischief
	Disorderly Person
	Disturbance
	Drug-related Complaint
	Harassment
	Larceny
	Suspicious Activity
	Vehicle Theft
	Weapons or Gunshot Complaint

reported in 3.2. Models were retrained on a rolling basis throughout the evaluation period using two years of training data, allowing models to adapt to emerging changes in crime patterns to make optimal predictions.

Many of the methods under analysis require a division of the study area into a number of small spatial units representing the units of analysis for prediction. These units can be defined using an arbitrary grid overlay in order to ensure uniform areas among all spatial units, or may alternatively correspond to real-world geographies such as street segments or police beats. As units of analysis get smaller in area, the spatial resolution of prediction increases, but localized crime clusters may also be more likely to be split across multiple spatial units and thus to become more difficult to detect. For this comparison of methods, we divided the city into an arbitrary grid of 500 foot by 500 foot square cells. Exploratory work indicated that steps for grid optimization improved predictions fairly uniformly across prediction methods, but did not result in significant changes to the *relative* performance of methods with respect to each other. We therefore do not report results across grid optimizations but instead for a single arbitrary grid overlay.

We compared 10 prediction models that have demonstrated good performance on forecasting crime counts or other types of spatio-temporal count data in the literature. The models under comparison are:

- **Within-Cell Moving Average (MAVG)**. A window length of fixed size was selected to optimize prediction accuracy on the calibration period. At prediction time, target crimes are tabulated for each spatial unit over the entire time window ending at the present time period.

- **Kernel Density Estimation (KDE)**. A density surface of target crimes was estimated from spatial occurrences of the target variable over a fixed time window using a Gaussian kernel. Kernel bandwidth and time window size were chosen to optimize prediction accuracy over the calibration period.

- **Logistic regression with lagged count features and $L1$-regularization (LASSO-LC)**. A logistic regression model with $L1$-regularization was trained using lagged count features from leading indicator variables made up of individual crime types and 911 call categories. Counts were converted to binary class labels indicating presence/absence of crime, and the regularization parameter was selected to optimize prediction performance on the calibration period.

- **Logistic regression with crime cluster features and $L1$-regularization (LASSO-CC)**. Spatiotemporal clusters of leading indicator crimes were detected using the Fast Subset Scan approach presented by Neill (2012b). Cluster characteristics (size, duration, intensity) were used as features for a sparse logistic regression model trained and tuned identically to LASSO-LC.

- **Gaussian Process Regression (GP)**. A Gaussian process regression model was trained assuming separable covariance across time and the two spatial dimensions. The isotropic squared exponential covariance function was applied for all three dimensions, and counts for leading indicators were included as linear terms in the mean function (Rasmussen and Williams, 2005; Saatçi, 2011). GP hyperparameters were tuned over the calibration period.

- **Univariate Self-Exciting Point Process (SEPP-UNI)**. As presented by Mohler et al. (2011), a univariate self-exciting point process model was trained in which a set of background crimes of the target type are assumed to occur independently across time and space according to a stationary Poisson process, and subsequently result in elevated predicted risk for offspring events in the spatial vicinity. Background and offspring event rates are estimated iteratively through variable-bandwidth kernel density estimation.

- **Self-Exciting Point Process with spatial covariates (SEPP-MULTI)**. An extension of the SEPP-UNI model was trained to allow counts from leading indicator crimes to contribute to the overall intensity function of the target crime, following the training procedure described by Reinhart and Greenhouse (2017).

- **Multilayer Perceptron (MLP) with cell-specific lagged count features (MLP-LC)**. A densely-connected feedforward network was trained to predict target crime counts from 52 weeks of lagged crime data for a set of leading indicator crime types. Model architecture was selected based on prediction accuracy in the calibration period, with a single hidden layer and 10 hidden units outperforming other variants under consideration.

- **MLP with lagged count and local neighborhood features (MLP-NH)** Additional neighborhood features were added to the MLP-LC model, consisting of lagged crime counts for leading indicators within the eight cells adjacent to the target cell for each observation.

- **Convolutional Neural Network (CNN)**. A convolutional neural network was adapted to predict target crime counts from a rectangular grid overlay of the city and lagged target crime counts within each cell. The CNN is designed to identify spatial patterns in target crimes which appear throughout the city and are predictive of future crime, rather than learning spatial structure individually within neighborhoods or other micro-areas.

The training and prediction framework was standardized as much as possible across the 10 models. Results from these models over the three-year evaluation period are reported in the following section.

3.2.3 Evaluation of prediction methods

A variety of evaluation metrics are available for measuring the relative and absolute performance of different prediction methods. For this analysis, the evaluation framework was designed to align closely with police goals for a hot-spot-based predictive policing program. In the place-based forecasting setting, a natural approach to assessing performance of crime prediction methods is to measure the prediction accuracy as the percent of crime volume captured within a fixed amount of forecasted area. Because crime tends to concentrate in areas making up a small fraction of the city, smaller and smaller proportions of crime are captured per unit of area patrolled as additional hot spots are added to a proactive policing program. A *prediction tradeoff curve* plots the crime volume captured versus area forecasted over a range of area proportions from zero up to an upper limit determined by patrol resource constraints. The curve provides a visual depiction of the benefit/cost tradeoff faced by police when allocating resources to proactive patrols. If a single model outperforms others everywhere along the tradeoff curve (i.e., is the highest curve), then that model is likely to perform best regardless of how much area is ultimately selected for treatment. When curves from two models cross within feasible ranges of patrol, the preferred model will depend on the specific amount of area being forecasted and patrolled in a hot spot program.

Consideration of outcomes other than crime prediction accuracy is also necessary when designing a predictive policing program for field implementation. Equity in crime deterrence benefits across the city is an important consideration for police, as measured by the spatial dispersion of the areas selected as hot spots across time. As discussed in Culyer (2001), one concept of equity is defined as the provision of services to all areas in a city that are predicted to be in need of crime prevention services, provided it is feasible for police to administer those services. A hot spot program that selects the same subset of locations every week for patrols results in a highly unbalanced distribution of proactive police effort throughout the city, leaving most of the city without the benefits of crime deterrence even in areas where there is a predicted need. Gorr and Lee (2017) use the annual footprint of predicted hot spots, as measured by the area of hot spots with prevention services in one or more weeks, as one measure of equity in the allocation of police resources.

For our empirical comparison of forecasting methods, we report *hot spot prediction entropy* as a measure of spatial dispersion of top-predicted areas. For given prediction model, let n_i be the number of times grid cell i appears in the top-predicted 1% of all N cells across all forecast periods, and define p_i as the proportion of total hot spot selections occupied by cell i:

$$p_i = \frac{n_i}{\sum_j^N n_j}$$

Then hot spot prediction entropy H is calculated as the entropy of this hot spot distribution across grid cells:

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i$$

Hot spot selection entropy is highest for predictions that lead to a uniform distribution of hot spots across the city, and lowest for prediction methods that choose the same hot spots in every time period. This statistic therefore provides a useful point of comparison relating to the geographic dispersion of hot spots selected by crime forecasting models.

Prediction tradeoff curves for all 10 forecasting models on predicting P1V crimes are presented in Figure 3.1. All models under consideration demonstrate some effectiveness at predicting crime, with all curves showing similar performance at the smallest levels of area forecasted within the city. As more area is forecasted, the relative performance of the models becomes clearer, with all models experiencing diminishing marginal rates of crime captured as more area is forecasted. The MAVG model shows the best prediction performance in the top-predicted areas, from zero up to 3% of the city area forecasted. Police departments are typically only able to effectively patrol 1-3% of a city in a hot spot policing program, thus these results indicate the relatively simple MAVG method is likely the best candidate for selecting hot spots that capture the maximum amount of crime on average. With the exception of the relative poor performance of the SEPP-UNI and LASSO-CC models, most of the prediction models performed fairly similarly within the 0-3% range of feasible areas shown in Figure 3.1. Although some of the models evaluated here are considerably more difficult to design, tune, and train than the MAVG model, it seems that the additional complexity of these models does not improve prediction performance within these feasible levels of patrol.

Table 3.3 reports summary statistics from the comparison of prediction methods. The partial area under the tradeoff curve (pAUC) reports the area under the tradeoff curve up to 1% of the city area forecasted. The pAUC results are a measure of predictive accuracy which relate directly to feasible levels of patrols that may be dispatched to the top 1% of hot spots selected by these models. As depicted visually by the prediction tradeoff curves, predictions from the MAVG model give the highest pAUC at the 1% level compared with all other models. This high level of prediction accuracy comes at the expense of decreased hot spot selection entropy (H) values. The MAVG model has among the lowest entropy values across models, indicating that the same cells are selected repeatedly in the top 1% of predictions for this model. This lack of spatial dispersion of predictions is unsurprising given the design of the MAVG model, which leverages only long-term trends in crime to make predictions. The MLP-LC model provided the greatest variance in top-predicted cells, with hot spot selection entropy of 8.42 over the evaluation period. These results indicate that for a hot spot program where prediction accuracy and equity of proactive patrols are both key objectives, an approach that leverages multiple models is necessary to achieve good performance on both metrics simultaneously.

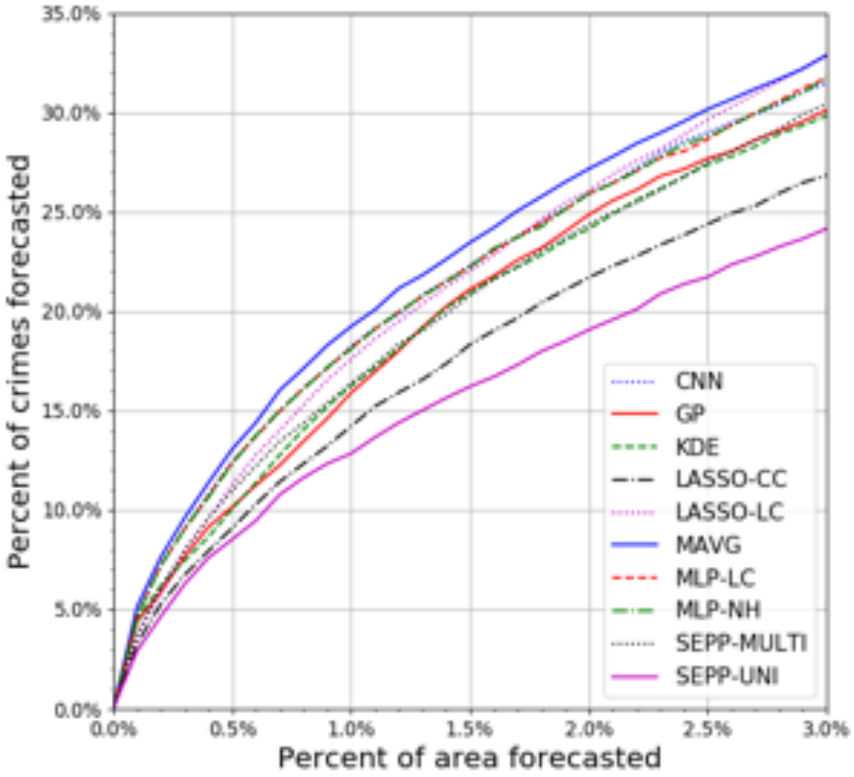


Fig. 3.1 Prediction tradeoff curves for 10 forecasting models on one-week-ahead predictions for Part 1 violent (P1V) crimes across a three year evaluation window.

3.3 Hot Spot Field Experiment in Pittsburgh, PA

Pittsburgh, PA is a city of 300,268 population with 64.9% white (non-Hispanic), 22.8% African American, 5.7% Asian, and 2.3% Hispanic racial composition. The Pittsburgh Bureau of Police (PBP) has approximately 900 sworn officers distributed across six police zones, each with a police station and commander. Results from the comparison of forecasting methods discussed in the previous section directly informed design of a field study in Pittsburgh, aimed at evaluating the effectiveness of hot spot-based proactive patrols. The field study was implemented in close partnership with the Pittsburgh Bureau of Police (PBP). The goal of the field study was to assess how well a small-scale hot spot policing program based on crime forecasts could deter serious crime offenses and equitably distribute police effort to locations in need throughout the city.

Table 3.3 Partial area under the tradeoff curve (*pAUC*) and hot spot selection entropy (H) for a 1% target area from 10 forecasting models across a three year evaluation window.

	Part 1 violent crimes	
	$pAUC \times 10^2$	H
MAVG	**1.31**	6.38
MLP-LC	1.23	**8.42**
MLP-NH	1.23	8.21
CNN	1.23	7.43
LASSO-LC	1.14	7.06
SEPP-MULTI	1.10	7.58
KDE	1.06	6.83
GP	1.05	6.30
LASSO-CC	0.94	8.31
SEPP-UNI	0.87	6.34

Best-performing model for each column denoted in bold face.

Prior to our field study, PBP maintained a policy that uniformed patrol officers conduct proactive patrols to prevent crimes and protect citizens. For such patrols, officers used discretionary time when not responding to 911 calls for service or other official duties. PBP policy for proactive patrols included community-oriented policing methods such as engaging with local citizens, avoiding enhanced enforcement actions such as zero-tolerance arrests and field interrogations. Command staff and individual officers determined patrol locations using judgment informed by experience and recent crime events. Officers were free to choose between car and foot patrol, although there was a preference by police leaders for foot patrol. Foot patrol used “park and walk” so that parked marked police cars provided additional police presence.

3.3.1 Experimental design

For the hot spot program under consideration, we relied on the same APRS and CAD data sets described in Section 3.2.1, refreshed on a weekly basis in city-maintained databases available to PBP crime analysts. Based on the outcomes of the empirical comparison of prediction methods discussed in Section 3.2, we determined that using a single prediction model for hot spot selection would not perform well on the multiple competing objectives of (1) prediction accuracy and (2) equitable distribution of hot spots throughout the city. A hot spot program with the sole aim of capturing the most crime within hot spots would likely be very static, with the same small set of *chronic hot spots* being selected for patrols each week.

Table 3.4 reports the percent of hot spot selections occupied by frequently-selected locations for a hot spot program targeting 1% of a city's total area, which provides an intuitive basis for understanding how frequently the top-selected hot spots change over time across models. Even the MLP-LC model, which had a high hot spot selection entropy relative to other methods, results in predicted hot spots that persist for long periods of time, with 43.6% of hot spot selections occupied by locations that remain hot spots for greater than 75% of the study period.

Through discussion with PBP command staff, we identified an objective to predict *temporary hot spots*, which represent short-term flare-ups over baseline levels of crime. In order to predict temporary flare-ups in target crimes throughout the city, we modified the MLP-LC model by tweaking the target variable for prediction. MLP-DIFF is identical to MLP-LC in structure and inputs, but the outcome variable for prediction is changed from total observed crime counts to the difference between observed counts and a one-year moving average of target counts within each cell. The target variable is clipped at zero to prevent negative observed outcomes and predictions. The predictions from MLP-DIFF no longer represent the predicted number target crime counts at each location, but instead represent the predicted positive difference from baseline levels of crime. Although the MLP-DIFF model results in lower prediction accuracy with respect to total crime captured, the hot spot selection entropy for this model is 10.53, much higher than the highest-entropy model considered in the previous section. Further, Table 3.4 indicates that the MLP-DIFF model results in top-predicted areas persisting for shorter periods of time than the MLP-LC and MAVG models.

Based on this evidence that the within-cell moving average (MAVG) and MLP-DIFF model address different police objectives for a proactive patrol program, we selected these two models for separate identification of chronic and temporary hot spots. Despite the MLP-DIFF predictions no longer representing a predicted total number of crimes occurring in a predicted area, we can still examine the prediction tradeoff curve and hot spot selection entropy from the temporary hot spot model to compare it with models predicting raw counts. In Figure 3.2, we show the tradeoff curves for the MAVG and MLP-DIFF models, as well as a composite curve that represents a combined chronic-temporary hot spots program with equal numbers of chronic (MAVG) and temporary (MLP-DIFF) hot spots.

Chronic and temporary hot spots exhibit different temporal patterns and relationships to minor crimes or other leading indicators, motivating our decision to select separate prediction models for selecting weekly chronic and temporary hot spots for patrol. The temporary hot

Table 3.4 Percent of hot spot selections occupied by frequently-selected locations (1% target area).

	% of Total Hot Spot Selections Occupied by Locations Persisting For:			
	> 25% of All Weeks	> 50% of All Weeks	> 75% of All Weeks	100% of All Weeks
MAVG	96.2%	91.0%	82.6%	69.6%
MLP-LC	74.2%	55.5%	43.6%	0.0%
MLP-DIFF	3.5%	1.7%	0.0%	0.0%

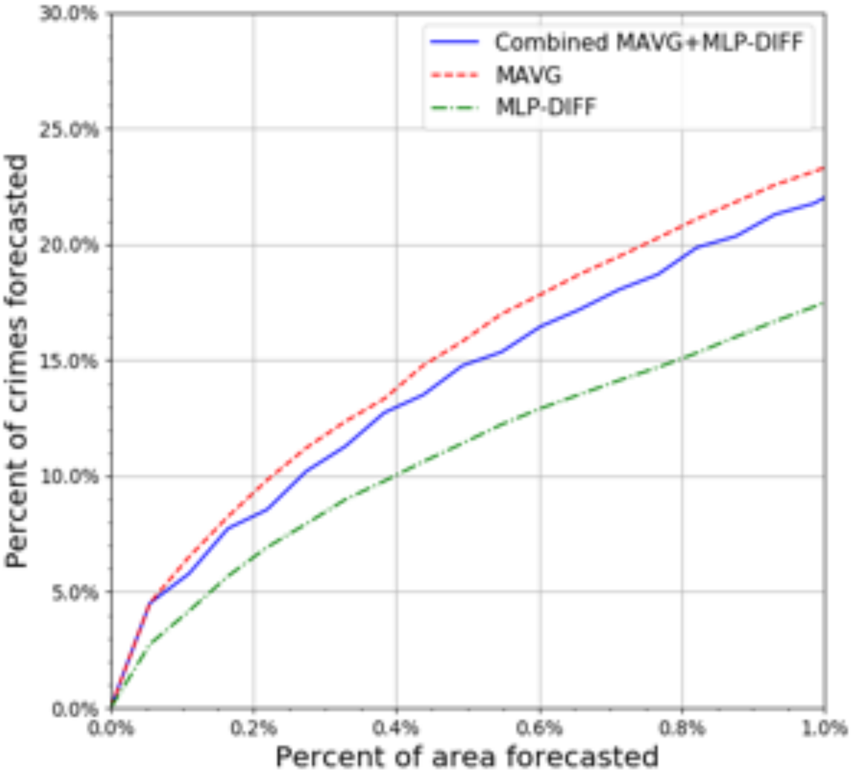


Fig. 3.2 Composite tradeoff curve representing a combined chronic-temporary hot spot program.

spot model includes weekly time lags of target crimes, as well as lags for leading indicator crimes and 911 call types. Hot spots from both prediction models were selected every week for each of the six police zones in Pittsburgh, and hot spots were provided to officers conducting patrols (on foot and in patrol cars) as part of their usual shifts throughout the day and night. P1V crimes were selected as the target crime type for prediction, as reducing the frequency of these violent offenses is considered a top priority for PBP command staff.

We evaluated the hot spot-based predictive policing program through a controlled field study motivated by crossover trials in the field of public health. Using historical data on P1V crime frequency, we divided all six police zones in Pittsburgh into two halves of equal area and roughly equal counts of historical violent crimes (Figure 3.3). For the experiment, the areas of Pittsburgh exposed to treatment (increased patrols in the predicted hot spots) were randomized then alternated on a weekly basis, so that no grid cell was selected for treatment two weeks in a row. In partitioning the city, we also avoided separating chronic hot spots across divisions in order to minimize spatial spillover effects from patrolling hot spots directly on the boundary between treatment and control areas. Residual crime deterrence effects after patrols leave an area are typically short (Koper, 1995; Telep et al., 2014), thus temporal spillover effects from treatment in previous weeks were expected to be negligible. Predicted hot spots in the "control" partitions for a given week were not provided to PBP; thus, control hot spots had policing as usual, including reactive policing in response to 911 emergency calls for service and a limited number of police-designed proactive patrols that existed before and during the experiment, but were not specifically targeted for patrols.

Beginning on February 20, 2017, we began a pilot period in which we initiated hot spot selection for one of the six police zones in Pittsburgh. By May 1, 2017, we had expanded the program to all six police zones, and upper level command staff from all zones were involved in directing proactive patrols to hot spots selected by the forecasting models. We initially identified 12 (six chronic and six temporary) hot spots per zone per week, totaling 72 hot spots targeted for additional patrols. The target treatment dosage for each 500 ft. by 500 ft. hot spot cell was three 15-minute foot patrols or nine 5-minute car patrols per hot spot per day. Hot spots were presented to patrol officers at the beginning of each week in a roll call meeting. A hot spot dashboard was also provided to officers through the computers in their vehicles, so that officers could easily locate hot spots while out on patrol. Through discussion with police and analysis of data on proactive patrols, we determined that individual hot spots were not receiving adequate levels of patrol dosage; we therefore dropped to six hot spots (three chronic and three temporary) per zone per week for the experimental phase. The result was a small-scale program with 36 hot spots for the city consisting of three chronic and three temporary hot spots for each of Pittsburgh's six police zones with 0.5 percent of the area of the city under treatment at any given time. During the study year treatment duration averaged slightly less than 20 minutes per day for both chronic and temporary hot spots, less than the targeted 45 minutes per day.

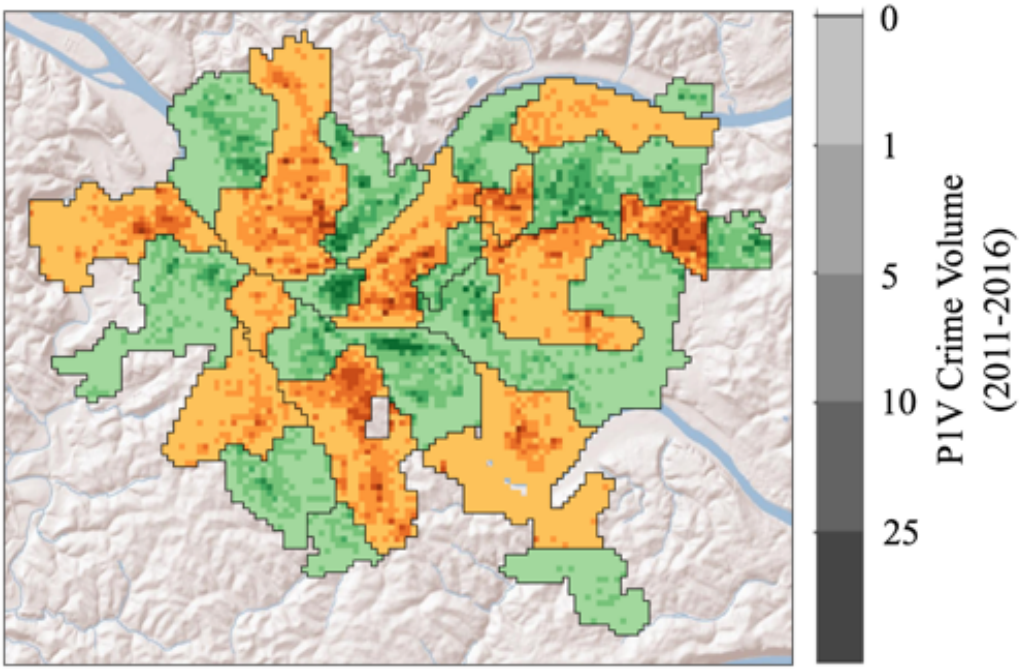


Fig. 3.3 Two equal-area partitions of Pittsburgh, Pennsylvania used to separate treatment from controls each week of the field trial. Partitions were selected to roughly balance historical P1V crime counts within each of six police zones. Darker areas indicate higher concentration of P1V crimes from 2011 through 2016.

The experimental phase of the field study ran for 12 months (November 6, 2017 through November 4, 2018). During the study year, there were a total of 37 homicides, 56 rapes, 686 robberies, 682 aggravated assaults, 1230 burglaries, 6394 larcenies, and 772 motor vehicle thefts in the city of Pittsburgh (exclusive of domestic violent crimes and retail crimes). Policing-as-usual for patrol officers in Pittsburgh, while mainly consisting of responding to 911 calls for emergency services, also includes a number of police-generated proactive patrols. We therefore expected that control hot spots, known to the researchers but not provided to PBP, would have some police-generated proactive patrols. Overall, control hot spots had proactive patrols in numbers equal to 14.8 percent of patrols conducted in treatment hot spots (1.9 vs. 12.9 patrols per hot spot per week), with 18.8 percent for chronic hot spots (2.6 vs. 13.8 patrols per hot spot per week) and 10.1 percent (1.2 vs. 12.1 patrols per hot spot per week) for temporary hot spots. Thus, the observed differences between P1V crimes in treatment and control hot spots represent the impact of targeting predicted (chronic or temporary) hot spots on a given week of the experiment, resulting in an additional 11 proactive patrols to targeted cells on average.

We employ a fixed effects regression model to evaluate statistical significance of per-cell P1V crime reductions from the hot spot-based proactive patrol program in Pittsburgh. The units of analysis are cell-weeks, and outcomes Y are Part 1 Violent (P1V) crime or Part 1 Property (P1P) crime counts per cell. We defined a regression model (Equation 3.1) with a binary variable $T_{p,w}$ indicating whether partition p was selected for treatment in week w, $HS_{c,w}$ indicating whether cell c was selected as a hot spot in week w, and $D_{c,w}$ indicating the number of proactive patrols in cell c during week w. $HS_{c,w} \times D_{c,w} \times T_{p,w}$ is then defined as an interaction term between hot spot selection and treatment dose. Fixed effect terms for partitions and weeks are included as α_p and δ_w respectively, and $\varepsilon_{c,p,w}$ terms are cell-specific errors. Model coefficients are estimated through ordinary least squares (OLS).

To minimize spillover of deterrence effects across treatment and control boundaries, cells in the control partition but directly adjacent to treatment hot spots were dropped from the analysis. Similarly, cells in the treatment partition that were directly adjacent to control hot spots were dropped. The primary coefficient of interest is β_1, as this represents the per-cell reduction in target crimes within treatment hot spots associated with an additional proactive patrol.

3.3.2 Results

We separately estimate and report dose-dependent and non-dose-dependent estimates of treatment effect on Part 1 Violent (P1V) crimes and Part 1 Property (P1P) crimes during the study year.

Dose-dependent per-cell crime reductions

We report per-cell dose-dependent regression results for four model variants: Model 1 (Equation 3.1) examines only the dose-dependent treatment effect of patrols to hot spot cells and includes week and zone-partition fixed effects. Model 2 (Equation 3.2) additionally examines possible changes in crime volume in cells adjacent to hot spots and in all other cells (cells not selected as a hot spot and not adjacent to a hot spot). Model 3 (Equation 3.3) estimates separate dose-dependent treatment effects for chronic and temporary hot spots. Finally, Model 4 (Equation 3.4) estimates separate dose-dependent treatment effects for car patrols and foot patrols on chronic and temporary hot spots. Previous studies have found that car patrols are approximately 1/3 as effective as foot patrols for preventing crime; therefore we compute combined patrol dose as (number of foot patrols) + (1/3)*(number of car patrols) for Models (1-3). See Tables 3.5 and 3.7 for the full table of regression results with P1V outcomes. We also provide results in Tables 3.6 and 3.8 on a second outcome variable, Part 1 Property crimes, to demonstrate that crime prevention was not limited to the targeted P1V crimes.

$$Y_{c,p,w} = \beta_0 + \beta_1[HS_{c,w} \times DALL_{c,w} \times T_{p,w}] + \beta_2 HS_{c,w} + \alpha_p + \delta_w + \varepsilon_{c,p,w} \tag{3.1}$$

$$\begin{aligned} Y_{c,p,w} = {} & \beta_0 + \beta_1[HS_{c,w} \times DALL_{c,w} \times T_{p,w}] + \beta_2 HS_{c,w} \\ & + \beta_3[ADJ_{c,w} \times T_{p,w}] + \beta_4 ADJ_{c,w} + \beta_5 NADJ_{c,w} + \alpha_p + \delta_w + \varepsilon_{c,p,w} \end{aligned} \tag{3.2}$$

$$\begin{aligned} Y_{c,p,w} = {} & \beta_0 + \beta_1[HSCHRONIC_{c,w} \times DALL_{c,w} \times T_{p,w}] + \beta_2 HSCHRONIC_{c,w} \\ & + \beta_3[HSTEMP_{c,w} \times DALL_{c,w} \times T_{p,w}] + \beta_4 HSTEMP_{c,w} \\ & + \beta_5[ADJ_{c,w} \times T_{p,w}] + \beta_6 ADJ_{c,w} + \beta_7 NADJ_{c,w} + \alpha_p + \delta_w + \varepsilon_{c,p,w} \end{aligned} \tag{3.3}$$

$$\begin{aligned} Y_{c,p,w} = {} & \beta_0 + \beta_1[HSCHRONIC_{c,w} \times DCAR_{c,w} \times T_{p,w}] \\ & + \beta_2[HSCHRONIC_{c,w} \times DFOOT_{c,w} \times T_{p,w}] + \beta_3 HSCHRONIC_{c,w} \\ & + \beta_4[HSTEMP_{c,w} \times DCAR_{c,w} \times T_{p,w}] \\ & + \beta_5[HSTEMP_{c,w} \times DFOOT_{c,w} \times T_{p,w}] + \beta_6 HSTEMP_{c,w} \\ & + \beta_7[ADJ_{c,w} \times T_{p,w}] + \beta_8 ADJ_{c,w} + \beta_9 NADJ_{c,w} + \alpha_p + \delta_w + \varepsilon_{c,p,w} \end{aligned} \tag{3.4}$$

A concern in estimating dose-dependent treatment effects is that dose was not randomized across hot spots. It is possible that officers may direct proactive patrols either towards or away from cells with high incidence of P1V, thereby affecting the dose-dependent treatment effect coefficient. We refute this possibility by computing the correlation coefficient between average number of proactive patrols when treated and average P1V when control for the 464 cells that are selected as both treatment and control at least once in the study period, and find null correlations for all patrol types (foot patrols vs. P1V: 0.017; car patrols vs. P1V: -0.009; foot patrols + car patrols vs. P1V: 0.000).

Model 1. The coefficient on Hotspot × Treatment Dose is negative and statistically significant (β=-0.0009, 95% CI=[-0.001, -0.000], P<0.001), while the coefficient on the Hotspot indicator variable is positive and statistically significant (β= 0.0431, 95% CI=[0.040, 0.046], P<0.001). Taken together, these results indicate that cells selected as a hot spot experience higher crime volume on average, but treating these areas with additional proactive patrols tends to decrease crime relative to hot spots that do not receive patrols.

Model 2. We find a weakly statistically significant reduction in P1V crimes in cells adjacent to hot spots (β=-0.0014, 95% CI=[-0.003, 0.000], P=0.082), providing some evidence for a small degree of spatial diffusion of crime deterrence benefits to areas surrounding predicted hot spots. Note that this result represents the effect of being adjacent to a treatment hot spot regardless of treatment dose at that location. We find no statistically significant change in P1V crime in cells not adjacent to hot spots (β=-0.0001, 95% CI=[-0.001, 0.000], P=0.895).

Model 3. The coefficients on Hotspot × Treatment Dose for both hot spot types are negative (indicating a reduction in crime volume from treatment) and statistically significant (chronic hot spots: β=-0.0015, 95% CI=[-0.002, -0.001], P<0.001; temporary hot spots: β=-0.0008, 95% CI=[-0.001, 0.000], P=0.024). Differences in the estimated treatment effects in chronic and temporary hot spots provide evidence that equivalent patrol protocols to chronic versus temporary hot spots may result in a differing magnitude of crime deterrence benefits.

Model 4. For foot patrols, the coefficients on Hotspot × Treatment Dose for both hot spot types are negative (indicating a reduction in crime volume from treatment) and statistically significant (chronic hot spots: β=-0.0031, 95% CI=[-0.005, -0.002], P<0.001; temporary hot spots: β=-0.0025, 95% CI=[-0.004, -0.001], P=0.001). In contrast, car patrols do not appear to provide P1V crime prevention. The dose-dependent treatment effect coefficient is not significant for car patrols in chronic hot spots, while additional car patrols have a positive and weakly statistically significant treatment effect in temporary hot spots (chronic hot spots:β=0.0002, 95% CI=[0.000, 0.001], P=0.47); temporary hot spots: β=0.0006, 95% CI=[0.000, 0.001], P=0.094).

Non-dose-dependent per-cell crime reductions

We additionally provide non-dose-dependent regression results in Table 3.7 estimating treatment effect of a cell being displayed to patrol officers as a hot spot, regardless of the observed number of proactive patrols to the location while selected as treatment. The interaction terms with the Treatment variable represents the treatment effect of being displayed as a hot spot for hot spot cells, and the effect of being in a treatment partition for cells adjacent to or not adjacent to a hot spot. While P1V crimes represent the primary target for proactive patrols, we additionally provide P1P results in Table 3.8 as evidence that crime prevention benefits are not limited to the targeted P1V crimes.

$$Y_{c,p,w} = \beta_0 + \beta_1[HS_{c,w} \times T_{p,w}] + \beta_2 HS_{c,w} + \alpha_p + \delta_w + \varepsilon_{c,p,w} \tag{3.5}$$

$$\begin{aligned} Y_{c,p,w} = {} & \beta_0 + \beta_1[HS_{c,w} \times T_{p,w}] + \beta_2 HS_{c,w} \\ & + \beta_3[ADJ_{c,w} \times T_{p,w}] + \beta_4 ADJ_{c,w} + \beta_5 NADJ_{c,w} + \alpha_p + \delta_w + \varepsilon_{c,p,w} \end{aligned} \tag{3.6}$$

$$\begin{aligned} Y_{c,p,w} = {} & \beta_0 + \beta_1[HSCHRONIC_{c,w} \times T_{p,w}] + \beta_2 HSCHRONIC_{c,w} \\ & + \beta_3[HSTEMP_{c,w} \times T_{p,w}] + \beta_4 HSTEMP_{c,w} \\ & + \beta_5[ADJ_{c,w} \times T_{p,w}] + \beta_6 ADJ_{c,w} + \beta_7 NADJ_{c,w} + \alpha_p + \delta_w + \varepsilon_{c,p,w} \end{aligned} \tag{3.7}$$

Model 1. The first model (Equation 3.5) examines only the treatment effect of patrols to hot spot cells, and includes week and zone-partition fixed effects as discussed in Section 3.3.1. For both crime types, the coefficient on Hotspot × Treatment is negative and statistically significant (P1V: β=-0.0128, 95% CI=[-0.017, -0.009], P<0.001; P1P: β=-0.0093, 95% CI=[-0.020, 0.001], P=0.091), while the coefficient on the Hotspot indicator variable is positive and statistically significant (P1V: β=0.0465, 95% CI=[0.044, 0.049], P<0.001; P1P: β=0.1501, 95% CI=[0.142, 0.158], P<0.001). Taken together, these results indicate that cells selected as a hot spot experience higher crime volume on average, but treating these areas with proactive patrols tends to decrease crime relative to hot spots that do not receive patrols.

Model 2. The second model (Equation 3.6) additionally examines possible changes in crime volume in cells adjacent to hot spots and in all other cells (cells not selected as a hot spot and not adjacent to a hot spot). We find a statistically significant reduction in

Table 3.5 Ordinary least squares (OLS) regression results estimating dose-dependent treatment effect of proactive patrols on P1V crime counts. Units of analysis are cell-weeks. Reference group for Model (1) is non-hot spot cells. Reference group for Models (2-4) is control cells not selected as hot spots or adjacent to hot spots.

	(1)	(2)	(3)	(4)
HS	0.0431***	0.0437***		
	(0.040, 0.046)	(0.041,0.046)		
$HS \times D_{ALL}$	-0.0009***	-0.0009***		
	(-0.001, 0.000)	(-0.001, 0.000)		
$HS_{chronic}$			0.0772***	0.0768***
			(0.074, 0.081)	(0.073, 0.080)
$HS_{chronic} \times D_{all}$			-0.0015***	
			(-0.002, -0.001)	
$HS_{chronic} \times D_{car}$				0.0002
				(0.000, 0.001)
$HS_{chronic} \times D_{foot}$				-0.0031***
				(-0.005, -0.002)
HS_{temp}			0.0122***	0.0116***
			(0.009, 0.016)	(0.008, 0.015)
$HS_{temp} \times D_{all}$			-0.0008**	
			(-0.001, 0.000)	
$HS_{temp} \times D_{car}$				0.0006*
				(0.000, 0.001)
$HS_{temp} \times D_{foot}$				-0.0025***
				(-0.004, -0.001)
$Adj.$		0.0102***	0.0102***	0.0102***
		(0.009, 0.011)	(0.009, 0.011)	(0.009, 0.011)
$Adj. \times T$		-0.0014*	-0.0014*	-0.0014*
		(-0.003, 0.000)	(-0.003, 0.000)	(-0.003, 0.000)
$Not\ Adj. \times T$		-0.0001	-0.0001	-0.0001
		(-0.001, 0.000)	(-0.001, 0.000)	(-0.001, 0.000)
Constant	0.0041***	0.0035***	0.0035***	0.0035***
	(0.002, 0.006)	(0.002, 0.005)	(0.002, 0.005)	(0.002, 0.005)
Week FE	Included	Included	Included	Included
Zone-Partition FE	Included	Included	Included	Included
Adj. R^2	0.005	0.007	0.009	0.009
No. Observations	361719	361719	361719	361719

Notes: 95% confidence intervals shown in parentheses. Significance at the 1% level is denoted by ***; ** denotes significance at the 5% level; and * significance at the 10% level.

Table 3.6 Ordinary least squares (OLS) regression results estimating dose-dependent treatment effect of proactive patrols on P1P crime counts. Units of analysis are cell-weeks. Reference group for Model (1) is non-hot spot cells. Reference group for Models (2-4) is control cells not selected as hot spots or adjacent to hot spots.

	(1)	(2)	(3)	(4)
HS	0.1394***	0.1419***		
	(0.135, 0.149)	(0.041,0.046)		
$HS \times D_{ALL}$	0.0018***	0.0017***		
	(0.001, 0.003)	(0.001, 0.003)		
$HS_{chronic}$			0.2210***	0.2179***
			(0.211, 0.231)	(0.208, 0.228)
$HS_{chronic} \times D_{all}$			0.0019**	
			(0.000, 0.003)	
$HS_{chronic} \times D_{car}$				0.0063***
				(0.005, 0.008)
$HS_{chronic} \times D_{foot}$				-0.0103***
				(-0.014, -0.006)
HS_{temp}			0.0682***	0.0696***
			(0.059, 0.078)	(0.060, 0.079)
$HS_{temp} \times D_{all}$			-0.0001	
			(-0.002, 0.002)	
$HS_{temp} \times D_{car}$				-0.0020**
				(0.000, 0.008)
$HS_{temp} \times D_{foot}$				0.0039*
				(-0.004, -0.001)
$Adj.$		0.0341***	0.0341***	0.0342***
		(0.031, 0.037)	(0.031, 0.037)	(0.031, 0.037)
$Adj. \times T$		-0.0034	-0.0034	-0.0034
		(-0.008, 0.001)	(-0.008, 0.001)	(-0.008, 0.001)
$Not\ Adj. \times T$		0.0001	0.0001	0.0001
		(-0.001, 0.001)	(-0.001, 0.001)	(-0.001, 0.001)
Constant	0.0171***	0.0150***	0.0150***	0.0150***
	(0.013, 0.021)	(0.011, 0.019)	(0.011, 0.019)	(0.011, 0.019)
Week FE	Included	Included	Included	Included
Zone-Partition FE	Included	Included	Included	Included
Adj. R^2	0.012	0.014	0.016	0.016
No. Observations	361719	361719	361719	361719

Notes: 95% confidence intervals shown in parentheses. Significance at the 1% level is denoted by ***; ** denotes significance at the 5% level; and * significance at the 10% level.

P1V crimes in cells adjacent to hot spots, providing evidence for a small degree of spatial diffusion of crime deterrence effects to areas surrounding predicted hot spots (β=-0.0014, 95% CI=[-0.003, 0.000], P=0.083). No significant crime deterrence effect was observed for P1P crimes in cells adjacent to hot spots. For both crime types, we find no statistically significant change in crime in cells not adjacent to hot spots.

Model 3. Finally, we estimate separate treatment effects for chronic and temporary hot spots (Equation 3.7). For P1V target crimes, the coefficients for both hot spot types are negative (indicating a reduction in crime volume from treatment) and statistically significant (chronic: β=-0.0203, 95% CI=[-0.026, -0.014], P<0.001; temporary: β=-0.0053, 95% CI=[-0.011, 0.000], P=0.072). For P1P target crimes, the treatment coefficient for chronic hot spots was statistically significant, but the coefficient for temporary hot spots was not, indicating that the benefit from patrolling temporary hot spots may be limited to reductions in violent crime and not other crime types (chronic: β=-0.0157, 95% CI=[-0.031, 0.000], P=0.043; temporary: β=-0.0029, 95% CI=[-0.018, 0.012], P=0.710). Differences in the estimated treatment effects in chronic and temporary hot spots were observed for both outcome crime types, providing evidence that equivalent patrol protocols to chronic versus temporary hot spots may result in a differing magnitude of crime deterrence benefits.

Table 3.7 Ordinary least squares (OLS) regression estimating non-dose-dependent treatment effect on Part 1 Violent crime counts. Units of analysis are cell-weeks. Reference group for Model (1) is non-hot spot cells. Reference group for Models (2) and (3) is control cells not selected as hot spots or adjacent to hot spots.

	(1)	(2)	(3)
HS	0.0465***	0.0472***	
	(0.044, 0.049)	(0.044, 0.050)	
$HS \times T$	-0.0128***	-0.0128***	
	(-0.017, -0.009)	(-0.017, -0.009)	
$HS_{chronic}$			0.0818***
			(0.078, 0.086)
$HS_{chronic} \times T$			-0.0203***
			(-0.026, -0.014)
HS_{temp}			0.0125***
			(0.008, 0.017)
$HS_{temp} \times T$			-0.0053*
			(-0.011, 0.000)
$Adj.$		0.0102***	0.0102***
		(0.009, 0.011)	(0.009, 0.011)
$Adj. \times T$		-0.0014*	-0.0014*
		(-0.003, 0.000)	(-0.003, 0.000)
$Not\ Adj. \times T$		-0.0001	-0.0001
		(-0.001, 0.000)	(-0.001, 0.000)
Constant	0.0041***	0.0035***	0.0035***
	(0.002, 0.006)	(0.002, 0.005)	(0.002, 0.005)
Week FE	Included	Included	Included
Zone-Partition FE	Included	Included	Included
Adj. R^2	0.005	0.007	0.009
No. Observations	361719	361719	361719

Notes: 95% confidence intervals shown in parentheses. Significance at the 1% level is denoted by ***; ** denotes significance at the 5% level; and * significance at the 10% level.

Table 3.8 Ordinary least squares (OLS) regression estimating non-dose-dependent treatment effect on Part 1 Property crime counts Units of analysis are cell-weeks. Reference group for Model (1) is non-hot spot cells. Reference group for Models (2) and (3) is control cells not selected as hot spots or adjacent to hot spots.

	(1)	(2)	(3)
HS	0.1501***	0.1525***	
	(0.142, 0.158)	(0.145, 0.160)	
$HS \times T$	-0.0093*	-0.0093*	
	(-0.020, 0.001)	(-0.020, 0.001)	
$HS_{chronic}$			0.2356***
			(0.225, 0.246)
$HS_{chronic} \times T$			-0.0157**
			(-0.031, -0.000)
HS_{temp}			0.0693***
			(0.058, 0.080)
$HS_{temp} \times T$			-0.0029
			(-0.018, 0.012)
$Adj.$		0.0342***	0.0342***
		(0.031, 0.037)	(0.031, 0.037)
$Adj. \times T$		-0.0034	-0.0034
		(-0.008, 0.001)	(-0.008, 0.001)
$Not\ Adj. \times T$		0.0001	0.0001
		(-0.001, 0.001)	(-0.001, 0.001)
Constant	0.0171***	0.0150***	0.0150***
	(0.013, 0.021)	(0.011, 0.019)	(0.011, 0.019)
Week FE	Included	Included	Included
Zone-Partition FE	Included	Included	Included
Adj. R^2	0.012	0.014	0.016
No. Observations	361719	361719	361719

Notes: 95% confidence intervals shown in parentheses. Significance at the 1% level is denoted by ***; ** denotes significance at the 5% level; and * significance at the 10% level.

Table 3.9 Counts and standard deviations of P1V crimes in treatment and control hot spots across cell-weeks.

	Control P1V			Treatment P1V			% Reduction
	Sum	Std. Dev.	N	Sum	Std. Dev.	N	in P1V
All Hot Spots	95	0.22	1872	71	0.20	1872	25.3%
Chronic Cells	80	0.29	936	61	0.26	936	23.8%
Temp. Cells	15	0.13	936	10	0.10	936	25.3%
Adj. Cells	163	0.12	11869	149	0.11	12135	8.6%

Spatial dispersion of chronic and temporary hot spots

As discussed in 3.2, the temporary hot spot model was selected to encourage spatial dispersion of top-selected hot spots as measured by a selection entropy index. Figure 3.4 shows the hot spot footprints for chronic and temporary hot spots across the 12-month period of evaluation. Cells are flagged in red if they were included as a hot spot at least once through the study period. As expected, the chronic hot spots selected using a one-year moving average remained fairly static, only occurring in a small fraction of the city over 12 months. By contrast, temporary hot spots selected by the multilayer perceptron predicting short-term flare-up in crime were spread over almost five times the area of chronic spots over the same period. Most cells in the temporary hot spot footprint received foot patrols at least once, with 6.5% of the city area receiving at least one foot patrol while treated as a temporary hot spot.

Overall reduction in crime volume from proactive patrols

The observed differences in P1V crimes and P1P crimes between control and treatment hot spots are reported in Tables 3.10 and 3.11. Within all hot spots, we measured 24 fewer P1V crimes in treatment hot spots relative to control hot spots, or a 25.3% reduction in P1V crimes per hot spot cell ($n_{treatment} = n_{control} = 1872$ cell-weeks). We measured 18 fewer P1P crimes in treatment cells, or a 5.5% reduction in P1P crimes in treatment hot spots relative to control hot spots. We see a larger percentage reduction in temporary hot spots for P1V crimes relative to chronic hot spots, with a reduction of 33.3% in temporary hot spots ($n = 936$) and 23.8% in chronic hot spots ($n = 936$). In non-hot-spot cells in treatment partitions, we measured an 4.5% decrease in P1V crimes per cell in treatment areas, and a 1.6% decrease in P1P crimes per cell in treatment areas.

While the numbers of crimes reduced are small, their impact in terms of cost avoidance to society and victims is large compared to the cost of proactive police patrols. In Table 3.10, we compute that overall there were $3,411,328 in crime costs avoided, while the cost of

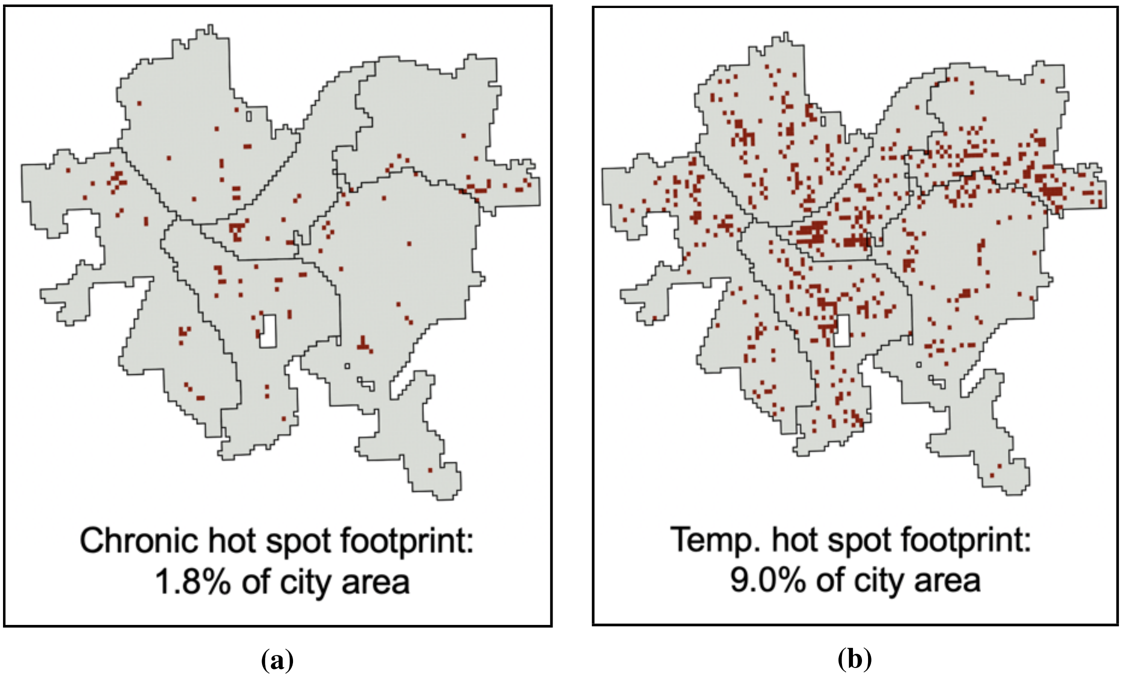


Fig. 3.4 Hot spot footprint in Pittsburgh over 12 months for two types of hot spots, (a) chronic hot spots predicted using a long-term moving average, and (b) temporary hot spots predicted using a multilayer perceptron. Cells are highlighted if they were selected as a hot spot at least once in the study period.

patrols was less than $1 million (see Appendix 3.A). Counts and standard deviations of P1V crimes in treatment and control cells are reported in Table 3.9. We observe a statistically significant reduction in the number of African American and other non-white victims of P1V crime, with 25 (39.7%) fewer victims in treatment vs. control cells over the study. Prevention of P1V crimes with multiple victims accounts for a higher proportional reduction in minority victims than the 25.3% reduction in overall P1V crime.

Proactive patrols increased relative to pre-study-year levels in control areas as well as treatment areas, likely due to patrol officers continuing to conduct proactive patrols to recent treatment areas during control weeks. We next estimate the overall impact of the hot spot program on prevention of serious violent crimes in both treatment and control areas, which requires several additional assumptions. First, we estimate the overall number of P1V crimes prevented by all proactive patrols (in both treatment and control hot spots) during the study year as 33.9 (27.8 in chronic hot spots + 6.1 in temporary hot spots), assuming that the crime reduction effect is linear in the number of patrols. Second, we assume that the counterfactual of "business as usual" in the absence of a hot spot program would have kept proactive patrols at pre-study-year levels (1.1 and 0.2 patrols per hot spot per week for chronic and temporary hot spots respectively), leading to 4.0 crimes prevented (3.8 in chronic hot spots + 0.2 in temporary hot spots) under the assumption of linear crime reduction effect. Finally, we note that the number of patrols outside hot spots was essentially unchanged from pre-study-year levels (204.7 vs. 197.5 average patrols per week), and (as discussed above) we observed no evidence of crime displacement, thus suggesting that the program had negligible impact on patrols or crime outside hot spots. These analyses suggest an estimate of 33.9 – 4.0 = 29.9 P1V crimes prevented in total by the hot spot program when considering crimes prevented in both treatment and control areas, with an associated $4,265,221 in crime costs avoided by the program.

Arrests during hot spot patrols

Over-policing is unnecessarily large amounts of police control and arrests, particularly for racial minorities or economically disadvantaged communities (Ben-Porat, 2008). Over-policing is especially a concern for "Broken Window" crimes that do not threaten public safety, such as consumption of alcohol on streets, possession of small amounts of marijuana, disorderly conduct, loitering, disturbing the peace, spitting, or jaywalking. Over-policing results from racial profiling, institutional biases ingrained in the culture or policies of a police department, or personal biases in officers. Pittsburgh police policy for proactive patrols is to engage with the local residents and employ best practices for community-oriented policing (for example, warning of possible violent crimes and gathering information). To assess the impacts of the hot spot program on potential harms from over-policing, PBP crime analysts (Johnson, 2019) measured the number of arrests occurring during hot spot patrols. In the time window during and 30 minutes after the 20,000 hot spot patrols conducted during the study year, only four arrests occurred while on patrol. Two of these were on-view arrests for minor drug offenses, and the remaining two occurred during 911 calls for service to domestic locations. PBP reported no arrests during hot spot patrol for other commonly over-policed crimes, such as trespassing, disorderly conduct, loitering, and public drunkenness.

3.4 Discussion

The results from this field study provide statistically significant evidence that a small-scale hot spot program based on proactive patrols, targeting only 0.5% of Pittsburgh, can lead to measurable reductions in serious violent crime in those areas, while avoiding over-policing arrests of racial minorities. These results contribute to the ongoing debate around proactive policing, showing both benefits and mitigation of certain potential harms when community-oriented policing practices are employed. Nonetheless, additional research is necessary to fully address the public policy question of whether conducting hot spot policing in a given

Table 3.10 Observed differences in Part 1 violent (P1V) crime counts between control areas and treatment areas over 12 months of proactive hot spot patrols (November 6, 2017 - November 4, 2018)

	Part 1 violent crimes		
	% Change	# of Crimes	Est. Costs Avoided*
All Hot Spots	-25.3%	-24	$3,411,328
Chronic Hot Spots	-23.8%	-19	$2,701,891
Temporary Hot Spots	-33.3%	-5	$709,437
Non-Hot-Spot Cells	-4.5%	-30	$5,553,501

*Costs per P1V crime computed based on costs to society from individual component crimes reported in (McCollister et al., 2010) and inflated to 2018 dollars. Note that the cost of the proactive police patrols was less than $1 million. See Table 3.12 for costs of individual component crimes.

Table 3.11 Observed differences in Part 1 property (P1P) crime counts between control areas and treatment areas over 12 months of proactive hot spot patrols (November 6, 2017 - November 4, 2018)

	Part 1 property crimes		
	% Reduction	# of Crimes	Est. Costs Avoided*
All Hot Spots	-5.5%	-18	$98,341
Chronic Hot Spots	-6.2%	-15	$90,960
Temporary Hot Spots	-3.5%	-3	$7,382
Non-Hot-Spot Cells	-1.6%	-61	$279,105

*Costs per P1P crime computed based on costs to society from individual component crimes reported in (McCollister et al., 2010) and inflated to 2018 dollars. Note that the cost of the proactive police patrols was less than $1 million. See Table 3.12 for costs of individual component crimes.

jurisdiction provides a net benefit to targeted communities. Predictive policing systems present non-trivial risks of harms resulting from over-policing or under-policing, or potential exacerbation of societal inequalities. We encourage future researchers and any users of predictive policing systems to employ a set of best practices for assessing and mitigating risks of bias and over-policing. These best practices include but are not limited to (1) using beneficial, community-oriented patrol protocols rather than aggressive policing interventions, (2) appropriately choosing minimally-biased target variables and predictor variables for prediction models, (3) regularly vetting prediction models across multiple evaluation criteria, (4) ensuring geographic dispersion of targeted patrols with a dynamic prediction model, (5) considering impacts of the scale and intensity of a targeted intervention, and (6) designing a predictive policing system around place-based rather than person-based predictions. Police leadership, crime analysts, and policy-makers should carefully consider not only the potential impacts of a program on crime reduction, but also the other consequences, intended and unintended, of such systems before implementing them at full operational scale in urban settings.

While this study focused on assessment of impacts on (1) crime volume, (2) spatial dispersion of predicted areas, and (3) arrests of commonly over-policed crimes, we recommend that future studies additionally evaluate how proactive patrols impact community sentiment surrounding increased presence of police and more frequent interactions with patrol officers. A survey-based approach, in which civilian researchers assess community sentiment and perception of potential harms related to over-policing in both treatment and control areas, would provide important evidence to inform how targeted patrols in treatment areas affects police-community relations and the perception of increased police presence. Additionally, observation of proactive patrols through review of body-mounted video camera footage would provide insight into the extent to which officers follow department-mandated protocols for community-oriented policing during proactive patrols. If policy-makers find that the benefits of a hot-spot-based predictive policing program outweigh the risks, adequate oversight procedures should be created to ensure that proactive patrols avoid aggressive and harmful policing practices and evaluate whether the program continues to achieve crime prevention benefits while avoiding unintended harms.

The potential for bias and feedback loops in the data generating process is a critical consideration within the experimental framework applied in this study. Historical biases in policing activity disproportionately affect minority communities, and thus relying on historical crime reports could result in a further entrenchment of these biases. Brantingham et al. (2018) examine whether biases in algorithmic place-based policing result in discriminatory consequences for minority groups, and find no significant difference in arrest rates across racial-ethnic groups between treatment and control areas. Still, feedback loops are possible when relying on reported crime data for hot spot selection. Areas may be initially selected for increased patrols due to high volume of historical crime reports; these areas subsequently generate additional reports of crime due to increased police presence, then the same areas are again selected as hot spots, and so on. To mitigate this issue, we selected hot spots based on predictions of P1V crimes, which consist of violent offenses such as homicide, rape, robbery, and aggravated assault. Reporting of these extremely serious and violent offenses is less likely than other crime types to depend on presence or absence of police in an area. Predictions of these crimes are therefore less prone to feedback loops resulting from increased patrols to hot spots.

A reduction of 24 P1V crimes represents 1.6% of observed citywide violent crime during this period, corresponding to an estimated $3.4 million in avoided crime costs to citizens and society. While this P1V reduction is small with respect to citywide crime volume, it results from a fairly minimal amount of effort per patrol shift (an average of 20 minutes of hot spot patrol per day). It is reasonable to expect that hot spot-based policing would experience diminishing marginal benefits as the number of patrolled hot spots grows, but the small dosage of hot spot patrolling evaluated in this study nonetheless leaves considerable capacity for scaling up the patrol program, as would eliminating car patrols and using foot patrols only. Additional studies are needed to evaluate the effects on crime volume of programs with different scales and to ensure that potential undesirable impacts related to over-policing are avoided as programs scale up. Nonetheless, this study provides evidence that hot-spot-based policing programs may be worthwhile even for police departments that cannot afford to invest heavily in proactive patrols, as even small-scale programs can lead to meaningful crime reductions.

Chronic and temporary crime hot spots exhibit fundamental differences affecting the mechanisms by which proactive patrols deter crime. Chronic hot spots may experience elevated rates of crime for years or decades at a time. These highest-crime areas are typically known to police, but crime volume can remain high even in the presence of regular police patrols. By contrast, residential areas tend to experience low baseline rates of crime, but may represent areas where the fear of crime is highest (Moore and Trojanowicz, 1988; Skogan and Maxfield, 1981). Patrolling of temporary hot spots results in a spreading out of police resources to residential areas away from commercial zones where crime is chronically high. Results from this study indicate that patrols to temporary hot spots provide a meaningful reduction in crime counts, despite having significantly less overall crime volume than chronic hot spots.

Research has shown that guardianship actions are more effective than enforcement actions in preventing crimes in hot spots (Braga and Schnell, 2015). This study supports recent findings (Piza, 2018; Ratcliffe et al., 2011) that foot patrol is an effective approach for crime deterrence, in contrast to proactive car patrols which do not provide crime prevention. Foot patrol represents an opportunity to design a proactive policing protocol that can be integrated with other community-oriented policing practices to improve police-community relations, reduce fear of crime, and make residents feel safe while simultaneously deterring serious violent crime.

One concern related to targeting high crime locations with proactive patrols is the possibility for displacement of crime to other locations outside of the area being patrolled. Does patrolling high-crime areas reduce overall crime volume, or do patrols simply push crime to other areas? Information about where police are located can travel rapidly through social communication networks that are not restricted by arbitrary cell boundaries used in this analysis. An examination of areas outside of hot spots is necessary to understand of overall impacts on crime volume from policing hot spots. In this field study, we separately examined the effect of patrols on per-cell crime volume in areas within hot spots, adjacent to hot spots, and away from hot spots. We find no statistically significant evidence of crime displacement from hot spots to other areas, whether they are adjacent to predicted areas or further away. We instead find some evidence of a diffusion of crime deterrence effects from patrolled areas, as cells adjacent to hot spots experience a small but statistically significant reduction in P1V crimes on average.

A limitation of this study is that we do not account for crimes displaced from treatment areas into control areas. Crimes that shift from treatment to control partitions as a result of patrols would lead to an overestimate of treatment effect size. We believe this is not a common occurrence, as a relatively small proportion of grid cells lie on a boundary between partitions and the above results suggest that the amount of crime displacement is low or nonexistent within partitions. We attempted to further mitigate this issue by drawing partition boundaries that avoid bisecting high-crime neighborhoods.

3.5 Conclusions

This study has examined the possible reductions in reported Part 1 offenses from a hot spot-based predictive program in Pittsburgh, PA, and provides evidence that even a small amount of effort and resources invested in such a program can lead to measurable and practically significant reductions in crime. We also examine the difference in crime deterrence benefits within *chronic* and *temporary* hot spots, and find that foot patrols to both hot spot types are effective at deterring crime.

Predictive analytics for policing is an emerging field, and more empirical studies are needed to understand the potential impacts on crime volume and other citizen outcomes from hot spot programs of different scales and across cities. This study focuses on hot spot models that rely on reported crime data, and additional work is needed to understand the impacts of specific areas and communities being over-represented or under-represented in reported crime data. Further research is also needed to identify what patrol activities and strategies are most effective at fostering goodwill among the communities being policed in addition to providing crime reduction benefits. Ultimately, designing a predictive policing system that is both transparent and equitable is essential for long-term support from the general public.

Appendix 3.A Estimates of Costs Avoided from Observed Crime Reductions

Table 3.12 shows estimates of per-offense costs to society as reported by McCollister et al. (2010). Costs were inflated from 2008 dollars to 2018 dollars using the CPI inflation rate reported by the U.S. Bureau of Labor Statistics. We estimate the cost of proactive patrols overall to be less than $1 million, as follows. The 2018 PBP budget was $105 million. We apply a conservative estimate of 66% as the proportion of this budget dedicated to patrol officers. 40 officers on-duty at any point in time over 3 shifts per day results in 43,800 officer-days per year. We apply a conservative estimate that each officer spends their full 8-hour shift on patrol, resulting in 350,400 hours of patrol per year and $197.77 spent per hour of patrol by PBP. The hot spot program in Pittsburgh resulted in 4,336 patrol hours to hot spots over the course of the study year, making $857,588 the estimated total cost of the program.

Table 3.12 Per-offense estimates of costs to society from Part 1 crime types.

		Cost* ($2008)	Cost** ($2018)	% of Crime Volume***
P1V	Murder/Manslaughter	$8,982,907	$10,510,001	2.5%
	Forcible Rape	$240,776	$281,708	4.5%
	Aggravated Assault	$107,020	$125,213	46.4%
	Robbery	$42,310	$49,503	46.6%
P1P	Burglary	$6,462	$7,561	15.5%
	Larceny	$3,532	$4,132	75.7%
	Vehicle Theft	$10,772	$12,603	8.8%

*Reported in McCollister et al. (2010).
**Costs in 2008 dollars were inflated to 2018 dollars using https://www.bls.gov/data/inflation_calculator.htm.
*** Proportion of crime volume within each Part 1 offense category (P1V and P1P) was calculated using five years of historical crime data from Pittsburgh's APRS system (June 1, 2011 through June 1, 2016).

References

Attorney general sessions announces opioid fraud and abuse detection unit. Technical report, U.S. Department of Justice Office of Public Affairs, 2017.

The pothole report: Can the bay area have better roads? Technical report, Metropolitan Transportation Commission, June 2011.

G. Ben-Porat. Policing multicultural states: lessons from the canadian model. *Policing and Society*, 18:411–425, 2008.

Richard L. Block and Carolyn R. Block. Space, place, and crime: Hot spot areas and hot places of liquor-related crime. 1995.

L. Bottou and Chih-Jen Lin. *Large-Scale Kernel Machines: Support vector machine solvers*. MIT Press, 2007.

Anthony A. Braga, Brandon Turchan, Andrew V. Papachristos, and David M. Hureau. Hot spots policing of small geographic areas effects on crime. *Campbell Systematic Reviews*, 15(3), 2019.

Welsh B.C. Braga, A.A. and C. Schnell. Can policing disorder reduce crime? a systematic review and meta-analysis. *Criminal Justice Policy Review*, 52, 2015.

Patricia L. Brantingham and P.J. Brantingham. Theoretical model on crime hot spot generation. *Studies on Crime and Crime Prevention*, 8:7–26, 1999.

P.J. Brantingham, M. Valasik, and G.O. Mohler. Does predictive policing lead to biased arrests? results from a randomized controlled trial. *Statistics and Public Policy*, 5(1):1–6, 2018.

A. L. Cancado, A. R. Duarte, L. H. Duczmal, S. J. Ferreira, C. M. Fonseca, and E. C. Gontijo. Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters. *International Journal of Health Geographics*, 9(55), 2010.

T.J. Cicero, M.S. Ellis, H.L. Surratt, and S.P. Kurtz. The changing face of heroin use in the united states: a retrospective analysis of the past 50 years. *JAMA Psychiatry*, 71(7), 2014.

R. Clarke and David Weisburd. Diffusion of crime control benefits: Observations on the reverse of displacement. *Crime Prevention Studies*, 2:165–184, 1994.

Lawrence E. Cohen and M. Felson. Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 44(4):588, 1979.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

M. A. Costa and M. Kulldorff. Maximum linkage space-time permutation scan statistics for disease outbreak detection. *International Journal of Health Geographics*, 13(20), 2014.

Anthony J. Culyer. Equity—some theory and its policy implications. *Journal of Medical Ethics*, 27:275–283, 2001.

M. Cuturi. Fast global alignment kernels. In *Proceedings of the 28th international conference on machine learning*, page 929–936, 2011.

T.G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

P.J. Diggle, J.A. Tawn, and R.A. Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society (Series C: Applied Statistics)*, 47(3), 2002.

D. Dowell, T. Haegerich, and R. Chou. Cdc guideline for prescribing opioids for chronic pain. Technical report, Centers for Disease Control and Prevention, 2016.

A. R. Duarte, L. Duczmal, and S. J. Ferreira. Internal cohesion and geometric shape of spatial clusters. *Environmental and Ecological Statistics*, 17(2):203–229, 2010.

James Ducharme and Sean Moore. Opioid use disorder assessment tools and drug screening. *Missouri Medicine*, 116(4):318–324, 2019.

L. Duczmal and R. Assuncao. A simulated annealing strategy for the detection of arbitrary shaped spatial clusters. *Computational Statistics and Data Analysis*, 45:269–286, 2004.

L. Duczmal, M. Kulldorff, and L. Huang. Evaluation of spatial scan statistics for irregularly shaped clusters. *Journal of Computational and Graphical Statistics*, 15(2):428–442, 2006.

L. Duczmal, A. L. Cancado, R. H. Takahashi, and L.F. Bessegato. A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics and Data Analysis*, 52: 43–52, 2007.

L. H. Duczmal, A.L.F. Cançado, and R. H. C. Takahashi. Delineation of irregularly shaped disease clusters through multiobjective optimization. *Journal of Computational and Graphical Statistics*, 2008(1):243–262, 2012.

John E. Eck and David Weisburd. Crime places in crime theory. In *Crime and Place*, volume 4, pages 1–33. Monsey: Criminal Justice Press, 1995.

M. Ester, H.P. Kriegal, J. Sander, and X. Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

L.M. Ferris, B. Saloner, N. Krawczyk, Schneider K.E., M.P. Jarman, K. Jackson, B.C. Lyons, M.D. Eisenberg, T.M. Richards, K.W. Lemke, and J.P. Weiner. Predicting opioid overdose deaths using prescription drug monitoring program data. *American Journal of Preventive Medicine*, 57(6):1917–1923, 2019.

Dylan J. Fitzpatrick, Wilpen L. Gorr, and Daniel B. Neill. Keeping score: predictive analytics in policing. *Annual Review of Criminology*, 2:473–491, 2019.

A. Geller, J. Fagan, T. Tyler, and B.G. Link. Aggressive policing and the mental health of young urban men. *American Journal of Public Health*, 104(12):2321–2327, 2014.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Proceedings of the 28th International Conference on Neural Information Processing*, 2014.

Wilpen Gorr and YongJei Lee. Early warning system for temporary crime hot spots. *Journal of Quantitative Criminology*, 31:25–47, 2015.

Wilpen Gorr and YongJei Lee. *Unraveling the Crime Time-Space Connection, New Directions in Theory and Policy*, volume 22, chapter Chronic and Temporary Hot Spots. Taylor and Francis Group, Abingdon, 2017.

J. Gorski, F. Pfeuffer, and K. Klamroth. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research*, 66(3): 373–407, 2007.

M. Haran. Gaussian random field models for spatial data. *Handbook of Markov Chain Monte Carlo*, pages 449–478, 2011.

Justine S. Hastings, Mark Howison, and Sarah E. Inman. Predicting high-risk opioid prescriptions before they are given. *PNAS*, 117(4):1917–1923, 2020.

Holly Hedegaard, Arialdi M. Miniño, and Margaret Warner. Drug overdose deaths in the united states, 1999–2017. Technical report, National Center for Health Statistics, 2018.

E. Hernandez, R. Torres, and A.L. Joyce. Environmental and sociological factors associated with the incidence of west nile virus cases in the northern san joaquin valley of california, 2011–2015. *Vector-Borne and Zoonotic Diseases*, 19(11):851–858, 2019.

Christopher R. Herrmann. The dynamics of robbery and violence hot spots. *Crime Science*, 4(33), 2015.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 2011.

Priscilla Hunt, Jessica Saunders, and John S. Hollywood. *Evaluation of the Shreveport Predictive Policing Experiment*. RAND Corporation, 2014.

Timothy R. Hylan, Michael Von Korff, Kathleen Saunders, Masters Elizabeth, Roy E. Palmer, David Carrell, David Cronkite, Jack Mardekian, and David Gross. Automated prediction of risk for problem opioid use in a primary care setting. *American Pain Society*, 16(4): 380–387, 2015.

Stephanie L. Hyland, Cristóbal Esteban, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *ArXiv e-prints*, 2017.

Heath Johnson. Personal communication (2019-06-01) with Pittsburgh Bureau of Police crime analyst, 2019.

N.C. Kamalu and E.C. Onyeozili. A critical analysis of the ‘broken windows’ policing in new york city and its impact: Implications for the criminal justice system and the african american community. *African Journal of Criminology and Justice Studies*, 11(1):71–94, 2018.

C.S. Koper. Just enough police presence: Reducing crime and disorderly behavior by optimizing patrol time in crime hot spots. *Justice Quarterly*, 12:649–672, 1995.

B.M. Kuehn. Opioid prescriptions soar: Increase in legitimate use as well as abuse. *JAMA*, 297(3):249–251, 2007.

M. Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26(6):1481–1496, 1997.

M. Kulldorff, L. Huang, L. Pickle, and L. Duczmal. An elliptical spatial scan statistic. *Statistics in Medicine*, 25:3929–3943, 2006.

R. L. Lampman, N. M. Krasavin, M. P. Ward, T. A. Beveroth, E. W. Lankau, B. W. Alto, E. Muturi, and R. J. Novak. West nile virus infection rates and avian serology in east-central illinois. *Journal of the American Mosquito Control Association*, 29(2):108–122, 2013.

Rachel N. Lipari and Eunice Park-Lee. Key substance use and mental health indicators in the united states: Results from the 2018 national survey on drug use and health. Technical report, Substance Abuse and Mental Health Services Administration, 2018.

T.A. Loughran, R. Paternoster, Alex R. Piquero, and G. Pogarsky. On ambiguity in perceptions of risk: Implications for criminal decision making and deterrence. *Criminology*, 49: 1029–1061, 2011.

K.E. McCollister, M.T. French, and H. Fang. The cost of crime to society: New crime-specific estimates for policy and program evaluation. *Drug and Alcohol Dependence*, 108(1-2): 98–109, 2010.

G.O. Mohler, M.B. Short, P.J. Brantingham, F.P. Schoenberg, and G.E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106 (493):100–108, 2011.

G.O. Mohler, M.B. Short, Sean Malinowski, Mark Johnson, G.E. Tita, Andrea L. Bertozzi, and P.J. Brantingham. Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, 110:1399–1411, 2015.

M.H. Moore and R.C. Trojanowicz. *Policing and the fear of crime.* National Institute of Justice, Washington, 1988.

G. J. P. Moreira, L. Paquete, L. H. Duczmal, D. Menotti, and R. H. C. Takahashi. Multi-objective dynamic programming for spatial cluster detection. *Environmental and Ecological Statistics*, 22(2):369–391, 2015.

National Academies of Sciences, Engineering, and Medicine. *Proactive Policing: Effects on Crime and Communities*. The National Academies Press, Washington, DC, 2018.

National Research Council. *Fairness and Effectiveness in Policing: The Evidence*. The National Academies Press, Washington, DC, 2004.

D. B. Neill. Expectation-based scan statistics for monitoring spatial time series data. *International Journal of Forecasting*, 25:498–517, 2009.

D. B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)*, 74(2):337–360, 2012a.

D. B. Neill and A. W. Moore. Rapid detection of significant spatial clusters. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 256–265, 2004.

D. B. Neill, A. W. Moore, M.R. Sabhnani, and K. Daniel. Detection of emerging space-time clusters. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 218–227, 2005.

Daniel B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)*, 74(2):337–360, 2012b.

J. Paparrizos and L. Gravano. k-shape: Efficient and accurate clustering of time series. *SIGMOD*, pages 1855–1870, 2015.

G.P. Patil and C. Taillie. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11:183–197, 2004.

E.L. Piza. The effect of various police enforcement actions on violent crime: Evidence from a saturation foot-patrol intervention. *Criminal Justice Policy Review*, 29, 2018.

T. Pollard, A. Johnson, J. Raffa, L. A. Celi, O. Badawi, and R. Mark. eicu collaborative research database (version 2.0), 2019.

Carl E. Rasmussen and Christopher K.L. Williams. *Gaussian Processes for Machine Learning*. MIT Press., 2005.

J.H. Ratcliffe, T. Taniguchi, E.R. Groff, and J.D. Wood. The philadelphia foot patrol experiment: A randomized controlled trial of police patrol effectiveness in violent crime hotspots. *Criminology*, 49(3):795–831, 2011.

A. Reinhart and J. Greenhouse. Self-exciting point processes with spatial covariates: Modelling the dynamics of crime. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2017.

K.K. Rigg, S. J. March, and J. A. Inciardi. Prescription drug abuse diversion: Role of the pain clinic. *Journal Drug Issues*, 40(3):681–702, 2010.

M.O. Ruiz, E.D. Walker, E.S. Foster, L.D. Haramis, and U.D. Kitron. Association of west nile virus illness and urban landscapes in chicago and detroit. *International Journal of Health Geographics*, 6(10), 2007.

Yunus Saatçi. *Scalable inference for structured Gaussian process models*. PhD thesis, University of Cambridge, 2011.

L. Sadeghi, Y. Zhang, Balmos A., J.V. Krogmeier, and J.E. Haddock. Algorithm and software for proactive pothole repair (joint transportation research program publication no. fhwa/in/jtrp-2016/14). Technical report, West Lafayette, IN: Purdue University, 2016.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and Xi. Chen. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing*, page 2234–2242, 2016.

B. Schölkopf, J.C. Platt, J.C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443 – 1471, 2001.

C.K. Scott, M.L. Dennis, C.E. Grella, R. Kurz, J. Sumpter, L. Nicholson, and R.R. Funk. A community outreach intervention to link individuals with opioid use disorders to medication-assisted treatment. *Journal of Substance Abuse Treatment*, 108:75–81, 2019.

Lawrence W. Sherman and David Weisburd. General deterrent effects of police patrol in crime "hot spots": a randomized, controlled trial. *Justice Quarterly*, 12(4):625–648, 1995.

Lawrence W. Sherman, Patrick R. Gartin, and Michael E. Buerger. Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology*, 27:27–56, 1989.

W. Skogan and M. Maxfield. *Coping with crime: individual and neighborhood reactions*. Sage Publications, Beverly Hills, CA, 1981.

S. Speakman, E McFowland, and D. B. Neill. Scalable detection of anomalous patterns with connectivity constraints. *Journal of Computational and Graphical Statistics*, 24(4): 1014–1033, 2015.

S. Speakman, S. Somanchi, E McFowland, and D. B. Neill. Penalized fast subset scanning. *Journal of Computational and Graphical Statistics*, 25(2):382–404, 2016.

K. Takahashi, M. Kulldorff, T. Tango, and K. Yih. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *International Journal of Health Geographics*, 7(14), 2008.

C.W. Telep, David Weisburd, C. Gill, Z. Vitter, and D. Teichman. Displacement of crime and diffusion of crime control benefits in large-scale geographic areas: A systematic review. *Journal of Experimental Criminology*, 10:515–548, 2014.

K.E. Vowles, M.L. McEntee, P.S. Julnes, T. Frohe, J.P. Ney, and D.N. van der Goes. Rates of opioid misuse, abuse, and addiction in chronic pain: A systematic review and data synthesis. *Pain*, 156:569–576, 2015.

David Weisburd. The law of crime concentration and the criminology of place. *Criminology*, 53(2):133–157, 2015.

David Weisburd and John E. Eck. What can police do to reduce crime, disorder, and fear? *The ANNALS of the American Academy of Political and Social Science*, 593:42–65, 2004.

David Weisburd, L.A. Wyckoff, J. Ready, John E. Eck, J.C. Hinkle, and F. Gajewski. Does crime just move around the corner? a controlled study of spatial displacement and diffusion of crime control benefits. *Criminology*, 44:549–592, 2006.

B.R. Wyant, R.B. Taylor, J. H. Ratcliffe, and J. Wood. Deterrence, firearm arrests, and subsequent shootings: A micro-level spatio-temporal analysis. *Justice Quarterly*, 29: 524–545, 2012.

N. Yiannakoulias, R. J. Rosychuk, and J. Hodgson. Adaptations for finding irregularly shaped disease clusters. *International Journal of Health Geographics*, 6(28), 2007.