New space/time interaction tests for spatiotemporal point processes

Seth R. Flaxman (sflaxman@cs.cmu.edu)
Event and Pattern Detection Laboratory, H.J. Heinz III College, Carnegie Mellon University, Machine Learning Department, School of Computer Science, Carnegie Mellon University

Daniel B. Neill (neill@cs.cmu.edu)
Event and Pattern Detection Laboratory, H.J. Heinz III College, Carnegie Mellon University

Alexander J. Smola (alex@smola.org)
Machine Learning Department, School of Computer Science, Carnegie Mellon University[a)]

(Dated: August 4, 2014)

We propose a new test for space-time interaction, using a Mercer kernel-based statistic for measuring the distance between probability distributions. As a concrete example, using a geocoded, time-stamped dataset from Chicago of almost 9 million calls to 911 between 2007 and 2010, we ask whether any of these call types are associated with shootings or homicides nearby in space and time. Standard correlation techniques do not produce meaningful results in the spatiotemporal setting because underlying spatial effects (e.g., "bad" neighborhoods) and temporal effects (e.g., more crimes in the summer) could introduce spurious correlations. To address this issue, a handful of statistical tests for space-time interaction have been proposed, which explicitly control for separable spatial and temporal dependencies. Yet these classical tests each have limitations. Our analysis sheds new light on the limitations of the existing tests, especially the Mantel test, and suggests a simple, theoretically grounded fix for this popular test (which has recently come under much scrutiny). We demonstrate how our new test can be extended to the bivariate and forward in time cases, enabling discovery of spatiotemporal leading indicators that are predictive of a target point process while controlling for purely spatial and purely temporal dependence. We compare our new test to existing tests on simulated and real data, where it outperforms the classical tests. We demonstrate its applicability to questions in criminology and predictive policing by using it to test for forward-in-time bivariate space-time interactions between disorder and crime and between various 911 call types and shootings/homicides.

PACS numbers: Valid PACS appear here

I. INTRODUCTION

What is the relationship between home foreclosures and violent crime[8,11]? It might be surprising at first to consider that simply estimating a meaningful measure of association, given two spatiotemporal point patterns, is non-trivial. We illustrate this point in Figure 1: in the left panel, we have plotted the time series of weekly violent crime and calls to Chicago's 311 non-emergency services number about vacant/abandoned buildings (vacant buildings were highlighted as a possible mediator of the link between foreclosures and crime in[8]). Pearson's correlation is 0.78. In the right panel, we display a scatterplot of the number of incidents of each event, with observations aggregated to the census tract level. Pearson's correlation is 0.73.

Can we thus conclude that there is a meaningful association between vacant/abandoned buildings and violent crime? Not necessarily. Based on a new statistical test for space-time interaction described in detail below, we conclude that the correlation structure between violent crime and vacant/abandoned buildings is explained by separable spatial and temporal factors (the p-value for the null hypothesis of no space-time interaction is 0.154, i.e. non-significant). Put another way, while the two types of events co-occur in space and co-occur in time, they do not co-occur in space and time more than is explained by spatial (e.g. neighborhood) and temporal (e.g. seasonal) effects. Of course, this is only an illustration: more careful evidence would be needed to evaluate the relationship between foreclosures and crime.

A key feature of spatial and temporal data is the non-independent nature of the observations. Valid statistical inference requires accounting for this dependence structure in the data. In the application area we consider, predictive policing, underlying spatial patterns (e.g., "bad neighborhoods" and varying density of offenders and targets) and temporal patterns (e.g., higher crime in the summer) are well known, and failing to take these factors into account will lead to spurious correlations between different types of crimes. Thus we wish to identify "leading indicators" for which occurrence in a particular place at a particular time is predictive of violent crime nearby in space and time, after controlling for purely spatial and purely temporal dependencies.

Knox[20], Mantel[25], and Diggle et al.[9] all developed important and widely used tests for space-time interaction with spatiotemporal point processes. While each of these statistical tests has different features, a fundamental limitation of each is the requirement that the user pre-specify a range of critical spatial and temporal distances of interest, i.e. a priori knowledge must be used to decide what distances are considered "close" versus "far", in space and time. One of the motivating goals of our work is to relax this assumption while not sacrificing statistical power. We take a new look at the assumptions underlying these tests, showing how each can be understood as testing a particular null hypothesis, namely that

[a)] Also at Strategic Technologies, Google

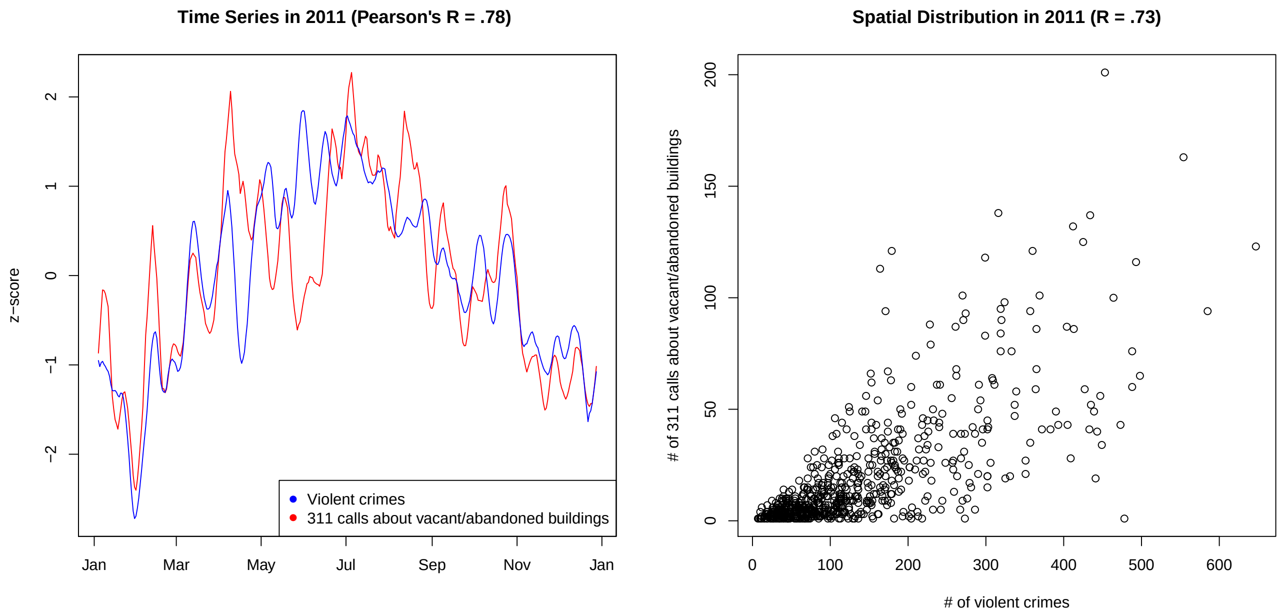


FIG. 1: Left: there is a strong correlation between the time series of violent crimes and calls to 311, Chicago's non-emergency services number, about vacant/abandoned buildings, due at least in part to common seasonal trends. Right: calls about vacant/abandoned buildings and violent crimes were aggregated to the census tract level (approximately 4,000 people) and plotted against each other. There is again a strong correlation.

the probability distributions over interpoint distances and interpoint time intervals are independent.

In this framework, we focus on the development of a set of new space-time interaction tests, based on the Hilbert Schmidt Independence Criterion (HSIC)[15], a kernel-based test statistic, for testing for independence between probability distributions. While HSIC was originally proposed for independent, identically distributed (iid) data, we motivate its use with spatiotemporal point processes, considering various alternative specifications. There has been limited development of space-time interaction tests for the bivariate case of measuring the space-time interaction between two spatiotemporal point processes. We extend the HSIC-based criterion to the bivariate case, and show interesting connections with the Mantel test along with a possible fix for some of the known issues with this test, using kernels.

We assess the power of our new test experimentally in a simulated dataset, where it compares favorably to existing methods for testing for space-time interaction, without requiring precise specification of various parameters. We also apply our new test to the application domain of predictive policing, using a data-driven approach to discover types of 911 calls which have significant forward-in-time space-time interaction with shootings, using geocoded, date-stamped crime data from the City of Chicago. This statistical formalization of leading indicators is a novel contribution to the criminology literature. As a final example, we address the well-known "broken windows" theory, looking for bivariate, forward-in-time associations between types of calls to Chicago's 311 (non-emergency services) number, indicative of low-level disorder, and violent and non-violent crime types.

A. Related Work

A number of different criminology methods and theories touch on the central focus of this paper, that of finding significant space-time interactions between crime incidents. From a spatial point of view, environmental criminology focuses on the criminal characteristics of places[6]. A related theory with an important spatial component is that of crime attractors, typified by the "broken windows" theory[36], that low-level disorder and crime act as signals which attract more, and more serious criminal behavior. The literature on crime hot-spots focuses on the fact that crimes tend to cluster spatially[5]. From a time series perspective, there has been much work on crime trends[4]. Recently, various advanced spatiotemporal models have been fit to crime data:[27] used self-exciting point process models, developed for earthquake modeling, for burglaries.[33] modeled violent crime using sophisticated semiparametric Bayesian techniques.[10] used nonparametric Bayesian techniques to segment crime data spatially. Explicitly addressing space-time interaction,[17] compared the "spatiotemporal signatures" of robbery, burglary, and assault. Previous work has evaluated the use of "leading indicator" crimes as predictors of violent crimes and property crimes using regression-based analyses with crimes aggregated into discrete bins in space and time[7] and spatial scan statistics[28]. There has been some work with univariate space-time tests to investigate the spatiotemporal dynamics of crime, including hot spots[17] and "near repeats"[35].

Space-time interaction tests are most widely used in the epidemiological literature, but most examples are univariate, focusing on the question of, e.g. the etiology of childhood leukemia[1]. The most similar work to this one appeared in ecology:[24] investigated the space-time interaction between spruce budworms and forest fires using point process methods.

B. Contributions

The space-time interaction test and our bivariate and forward-in-time extensions are novel tests based on the Hilbert-Schmidt Independence Criterion. The framework we propose for space-time interaction testing gives a new perspective on the classical Mantel test, provides an alternative to classical tests for space-time interaction, and shows how kernel-embedding techniques can be used with spatiotemporal point processes. In terms of applications, we are not aware of any previous work in the criminology literature that has focused on identifying leading indicators of crime through bivariate space-time interaction tests.

II. THEORETICAL DEVELOPMENT

A. Background: Classical Tests for Space-Time Interaction

Let $\mathcal{P} = \{(s_i, t_i), i = 1, \ldots, n\}$ be a realization of a spatiotemporal point process with two spatial dimensions ($s_i \in \mathcal{R}^2$) and a time dimension. We can think of $s_i \in A$ for a spatial region A and $t_i \in T$ for a time window T. An illustration is shown in Figure 2.

We start by stating the **Knox test**[20]. Given $\mathcal{P}$, we create a two-by-two contingency table as follows: pick a threshold distance for "near in space" s_0 and a threshold time interval for "near in time" t_0. Now, consider every pair of distinct points $s, s' \in \mathcal{P}$. Let $d_s(p, p')$ measure the Euclidean distance between p and p': $\sqrt{(x - x')^2 + (y - y')^2}$ and $d_t(p, p')$ measure the time interval: $|t - t'|$. Then, we can fill in the table by asking for each pair of points whether $d_s(p, p') \leq s_0$ and whether $d_t(p, p') \leq t_0$:

	near in space	far in space
near in time	X	n_1
far in time	n_2	$N - (X + n_1 + n_2)$

If there are $N = n(n-1)/2$ pairs of points, the test statistic is given by the difference between the number of pairs that we observe to be near in both time and space, X, and the number of pairs that we would expect to be near in both time and space if time and space are independent: $N \frac{X+n_1}{N} \frac{X+n_2}{N}$. Together this is: $X - \frac{1}{N}(X + n_1)(X + n_2)$. Since the null hypothesis is that space and time are independent, we can empirically find the distribution of X under the null by randomly permuting the time labels and recomputing the test statistic. Notice that $X + n_1$, the number of points that are close in time, is unchanged if the time labels are permuted. The same is true of $X + n_2$. This simplifies our calculations, and we need only consider the distribution of the test statistic X under the null. Various asymptotic approximations to the null distribution are discussed in[22].

The Knox test is very straightforward, but it clearly has limitations. Correctly specifying the spatial and temporal ranges is not always easy, and considering a range of values leads to problems of multiple hypothesis testing. As a toy example, Figure 3 shows how the power of the Knox test depends critically on the choice of cutoffs. We generated synthetic data from a point process with space-time clustering, using the setup discussed in Section III.A, and varied the spatial cutoff for the Knox test from 0 to 0.5. When the spatial cutoff is equal to about 0.1, the test correctly rejects the null in almost every case (α is fixed at 0.05). But for smaller and larger values of s_0, the power decreases. For further intuition, consider the illustrative dataset in Table I. If the cutoff is set such that the close and intermediate categories (in space and time) are collapsed, the Knox test will fail to reject the null hypothesis of no space-time interaction. Similarly, if the cutoff is set such that intermediate and far categories (in space and time) the Knox test will fail to reject. But there is evidently space-time interaction, at an intermediate distance.

	close in space	intermediate	far in space
close in time	1	1	1
intermediate	1	5	1
far in time	1	1	1

TABLE I: In this simple example, space-time interaction occurs at an intermediate distance in space and time.

Another concern is that the Knox test is based solely on distances between points, ignoring any other relevant features, like location in space and time. When Knox proposed his test, he was quite explicit, stating that *all* of the information required for a test of space-interaction is found in the interpoint time and space distances[20]. But his claim ignores the possibility of other types of inhomogeneities, as was pointed out at the time[3].

Next, we describe the **Mantel test**[25]. Given $\mathcal{P}$, create an $n \times n$ spatial distance matrix D_S with entries given by $d_s(p_i, p_j)$ for row i and column j and an $n \times n$ temporal distance matrix D_T with entries given by $d_t(p_i, p_j)$. As with the Knox test, we wish to ask whether space and time, now represented by two matrices, are independent. We string out the entries above the diagonal of each matrix as a vector with $n(n-1)/2$ entries, and calculate the Pearson correlation between these vectors. Notice, however, that the usual significance test for Pearson's correlation is not valid, because the observations are not independent. To derive the null distribution, we again turn to randomization testing, this time applying a given permutation to the rows and columns of one of the matrices, so as to preserve the dependence structure among the entries. We are mostly concerned about shorter time and spatial distances, but as described above, the Mantel test could be significant due to (spurious) longer range features. In[25], Mantel proposed the reciprocal transformation for both spatial and temporal distances x, forming the matrices of $f_s(d_s(p_i, p_j))$ and $f_t(d_t(p_i, p_j))$ where $f_s(x) = \frac{1}{x+\epsilon_s}$ and $f_t(x) = \frac{1}{x+\epsilon_t}$. The Mantel test is essentially a linear test of dependence, so we expect it to fail under the same conditions in which Pearson correlation and linear regression fails.

The **Diggle et al. test**[9] has a similar flavor to the Knox test, but rather than a single threshold value, it requires the specification of a range of values. First, we define Ripley's K function (also called the reduced second moment measure)

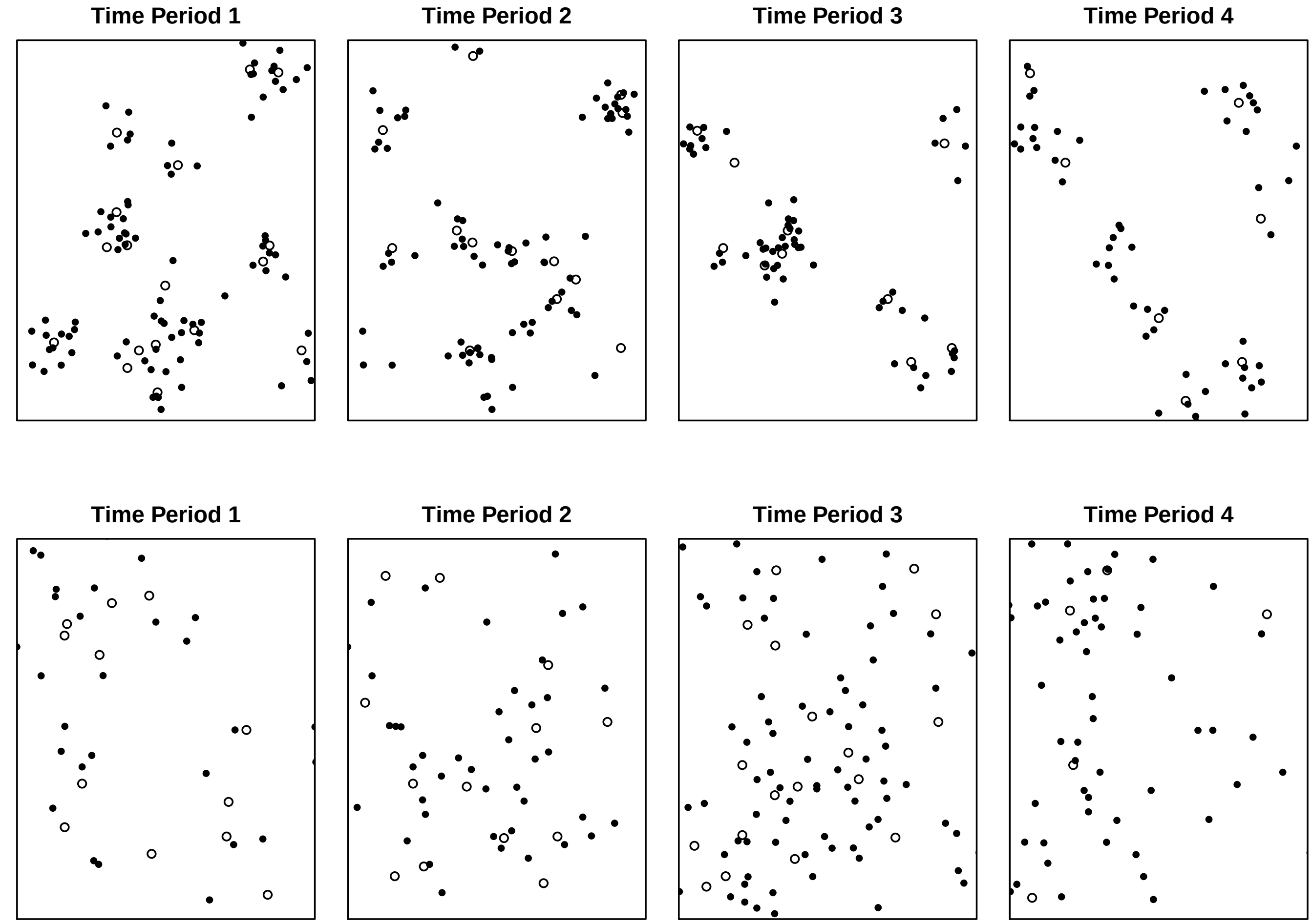


FIG. 2: Two different "infectious" Poisson cluster processes with parents shown as open circles and children shown as filled circles. Children are displaced from parents in space and time by iid draws from $N(0, \sigma)$. The top row displays the case for $\sigma = 0.05$, the bottom row for $\sigma = 0.2$. Visual inspection reveals space-time interaction in the first row while the second row is more ambiguous. Tests for space-time interaction correctly reject the null hypothesis (of no space-time interaction) in both cases, with $p \leq 0.01$.

for a single spatial point process as the following: $K(s) = \frac{1}{\lambda_S} E[\#$ of events occurring within a distance s of an arbitrary event where λ_S is the intensity of the point process. An estimate is given by $\widehat{\lambda_S} = N/A$ for N points in a spatial region with area A.

Given spatial point locations $S \in \mathbb{R}^{n \times 2}$ in a region with area A, the simplest way of estimating $\widehat{K}(s)$ is by averaging:

$$\widehat{K}(s) = \frac{1}{\widehat{\lambda_S}} \sum_{i=1}^{n} \frac{1}{n-1} \sum_{i \neq j} I(d_s(p_i, p_j) \leq s) \quad (1)$$

$$= \frac{A}{n(n-1)} \sum_{i} \sum_{i \neq j} I(d_s(p_i, p_j) \leq s) \quad (2)$$

This estimator assumes a constant first order intensity and since we are taking a ratio of expectations, we also assume the following regularity conditions: [TO BE FIXED]. This test also ignores the issue of edge corrections: at the boundary of the spatial or temporal region, "missing" observations bias the estimate. This becomes an issue for small n or s large compared to A. Corrections are given in[29]. The new test that we propose will not have this shortcoming.

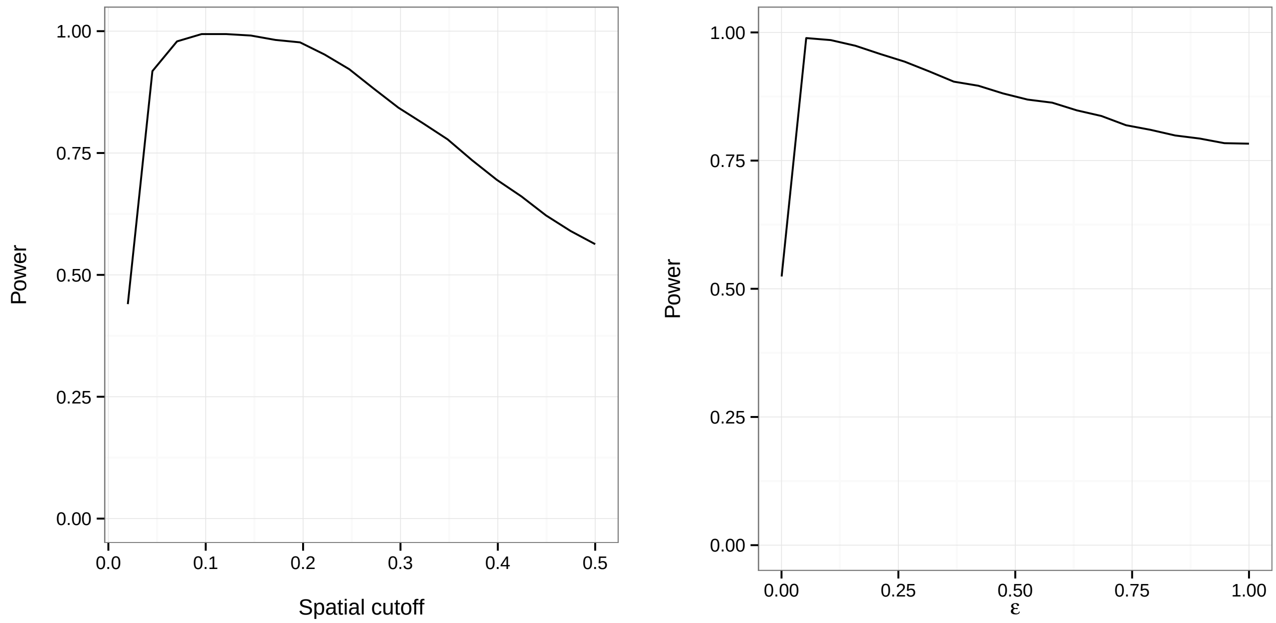
Ripley's K function has natural extensions to the purely temporal $K(t)$ and space-time $K(s, t)$ cases, with similar estimators to the above. We remark that $\widehat{\lambda_{ST}}\widehat{K}(s_0, t_0)$ is equal to the entry in the upper-left hand corner of the contingency table used in Knox's test, and similarly $\widehat{\lambda_S}\widehat{K}(s_0)$ and $\widehat{\lambda_T}\widehat{K}(t_0)$ are equal to the top row and left column, respectively.

Diggle et al. define residual space-time interaction at spatial scale s and time t as:

$$D(s, t) = K(s, t) - K(s)K(t)$$

Using this function, Diggle et al. define a test statistic calculated over a grid of pre-specified spatial distances $s_1, \ldots, s_k$ and time intervals $t_1, \ldots, t_l$:

$$R = \sum_{s_i} \sum_{t_j} D(s_i, t_j)$$



(a) Knox

(b) Mantel

FIG. 3: We generated synthetic data from a cluster point process with children points displaced from their parents a distance $\sim N(0, \sigma = .05)$ in space and time. For $\alpha = 0.05$, the Knox test correctly rejects the null when the spatial cutoff is well chosen, but as the cutoff decreases or increases, the power decreases. The temporal cutoff is fixed at 0.1 in every case. This demonstrates that the power of the Knox test depends on correctly specifying cutoffs for "close in space" and "close in time." Similarly, the Mantel test's power depends on correctly specifying a transformation from distance to "similarity." In Mantel's original formulation, distances x were transformed as $f(x) = \frac{1}{x+\epsilon}$ for some ϵ. On the right, the Mantel test correctly rejects the null almost all the time when ϵ is well chosen, but as ϵ increases or decreases the power decreases. The same transformation was used for space and time.

Under the null hypothesis of no space-time interaction, the expectation of R should be 0. The intuition is the same as for the previous tests: $K(s, t)$ tells us how many points we expect to see within a distance s and time t of an arbitrary point. As in the previous tests, permutation testing by shuffling the time labels is used to obtain the null distribution of R. The Diggle et al. test is meant to address the issue of multiple hypothesis testing that arises when the Mantel or Knox test are applied repeatedly. However, it may lose power due to the fact that it is measuring a statistic of interest over multiple thresholds: this statistic may be positive or negative at different thresholds, and thus may cancel out, or it may be zero at many thresholds and thus go undetected.

This completes our presentation of classical space-time interaction tests. Note that we have not provided an exhaustive review. Other tests for point processes include Jacquez's nearest neighbor based method[18]. Various improvements to the Knox test have been proposed in[2,21]. There is also a parallel literature in geostatistics and Gaussian processes on tests for the "separability" of space-time covariance functions[13,14].

Notice the commonalities among the tests: each is a hypothesis test with the same null hypothesis, that the interpoint spatial and temporal distributions are independent. To see this, note that the contingency table in the Knox test is used to ask whether binary indicator variables for pairs of points (near in space, near in time) are independent. The Mantel test uses Pearson correlation to test whether the interpoint space and interpoint time distributions are independent. Diggle et al.'s test asks whether there is a difference between the cumulative distribution of how many points are near in space and near in time and the product of the marginal distributions of how many points are near in space and how many points are near in time.

B. Extending the Classical Tests to Bivariate Space-Time Interaction

Given $\mathcal{P}^1 = \{(s_i^1, t_i^1), i = 1, \ldots, n_1\}$ and $\mathcal{P}^2 = \{(s_i^2, t_i^2), i = 1, \ldots, n_2\}$, we wish to know whether there is significant space-time interaction between $\mathcal{P}^1$ and $\mathcal{P}^2$. The null hypothesis is that there is no space-time interaction between the two processes. Notice that we are not interested in whether there is purely spatial dependence between $\mathcal{P}^1$ and $\mathcal{P}^2$: any two processes associated with, for example, an underlying population density will be spatially correlated. Similarly, we are not interested in purely temporal dependence between the two processes, e.g. due to seasonal trends. Instead, we wish to test whether seeing points of type 1 at a certain location in space and time makes it more or less likely that we will see points of type 2 nearby in space and time, once we have controlled for separable spatial and temporal correlations between $\mathcal{P}^1$ and

$\mathcal{P}^2$.

The Mantel, Knox, and Diggle et al. tests each focus on pairs of points. For the bivariate extension for each, we simply consider all $n_1 \cdot n_2$ cross-pairs of points. For the Knox test we create the same contingency table where each entry counts the number of cross-pairs that are near in time and near in space, the number of cross-pairs that are near in time and far in space, etc. For randomization testing, it is sufficient to permute the time labels of only one of the point processes. For the Mantel test, we create an $n_1 \times n_2$ spatial cross-distance matrix and an $n_1 \times n_2$ temporal cross-distance matrix, and the test statistic is the same. The bivariate version of the Mantel test was explored in[19]. The Diggle et al. extension is straightforward as well[24].

C. Background: Kernel Embedding of Distributions

We start with a presentation of the use of kernel embeddings of probability distributions for measuring distances between samples, focusing on the Hilbert-Schmidt Independence Criterion (HSIC)[15]. The motivation behind HSIC is as follows: given a joint distribution $(x, y) \sim (P, Q)$, we test the null hypothesis that $P \perp\!\!\!\perp Q$. If P and Q happened to be jointly Gaussian, then Pearson correlation could be used to test this hypothesis, or equivalently one can ask whether $\mathrm{Cov}(P, Q) = 0$. The obvious limitation of this approach is that it can only capture linear dependencies. HSIC can capture non-linear dependencies, while still using a linear statistic (covariance), by first embedding P and Q into a Reproducing Kernel Hilbert Space (RKHS). After representing $x \sim P$ and $y \sim Q$ as infinite dimensional vectors in the feature space representation of P and Q, we can define and measure the length of a covariance operator Σ_{PQ} in Hilbert space. At a purely mechanical level, this is an application of the "kernel trick," where dot products in the original space (the sample estimate of covariance is the dot product between two centered vectors of observations) are replaced by Mercer kernels. But RKHS theory takes this beyond merely a trick to give us statistical guarantees, and as we shall see, the use of kernels sheds new light on classical statistical tests for space-time interaction.

We now state HSIC. Assume we have an RKHS $\mathcal{H}_\mathcal{P}$ with a reproducing kernel k, where $\mathcal{H}_\mathcal{P}$ is the space of functions $f : \mathcal{R}^d \to \mathcal{R}$ and k is a positive semidefinite function over pairs of elements in $\mathcal{R}^d$. For concreteness, consider the Gaussian Radial Basis Function (RBF) kernel: $k(x, y) = e^{-\frac{\|x-y\|_2^2}{2\sigma^2}}$. Given a function $f \in \mathcal{H}_\mathcal{P}$ and an element $x \in \mathcal{R}^d$, the Riesz representation theorem states that we can evaluate $f(x)$ using the feature mapping: $f(x) = \langle \phi(x), f \rangle_{\mathcal{H}_\mathcal{P}}$. Now, we extend this to the case of expectations over elements $x \sim P$, defining a "mean-embedding" element $\mu_p \in \mathcal{H}_\mathcal{P}$ as follows:

$$E_x f(x) = E_x \langle \phi(x), f \rangle_{\mathcal{H}_\mathcal{P}} = \langle E_x \phi(x), f \rangle_{\mathcal{H}_\mathcal{P}} = \langle \mu_p, f \rangle_{\mathcal{H}_\mathcal{P}}$$

We need to establish under what conditions we can move the expectation inside the inner product; under these conditions, μ_p exists. Lemma 3 in[16] states:

Lemma 1. *If k is measurable and $E_x \sqrt{k(x, x)} < \infty$ then μ_p exists.*

Analogously, we define μ_q for a Hilbert space $\mathcal{H}_\mathcal{Q}$, and μ_{pq} for the Hilbert space formed by the tensor product of $\mathcal{H}_\mathcal{P} \otimes \mathcal{H}_\mathcal{Q}$. Just as $\mathrm{Cov}(P, Q) = E[XY] - E[X]E[Y]$ we can define an operator $\Sigma_{PQ} = \mu_{pq} - \mu_p \mu_q$ where:

$$\langle f, \Sigma_{PQ} g \rangle = E_{xy}[f(x)g(y)] - E_x[f(x)]E_y[g(y)] = \mathrm{Cov}(f(P), g(Q))$$

Recall that our goal is to test whether $X \perp\!\!\!\perp Y$. For f and g drawn from appropriately rich function classes, it is easy to see that: $X \perp\!\!\!\perp Y$ if and only if $\sup_{f,g} \mathrm{Cov}(f(X), g(Y)) = 0$

As a stylized example, consider $X \sim N(0, 1)$ and $Y = X^2$. X and Y are clearly dependent, but a linear test for independence would fail to discover this, since $\mathrm{Cov}(X, Y) = 0$. A simple transformation of variables with $g(y) = \sqrt{y}$ fixes the problem.

If we require that $f \in \mathcal{H}_\mathcal{X}, g \in \mathcal{H}_\mathcal{Y}, \|f\| \le 1, \|g\| \le 1$, then the supremum of this statistic can be found in closed form (i.e. without any explicit optimization):

$$\begin{aligned} &\sup_{f,g} \mathrm{Cov}(f(X), g(Y)) && (3)\\ &= \sup_{f,g} \langle f, \Sigma_{PQ} g \rangle && (4)\\ &= \sup_{f,g} \langle f \otimes g, \mu_{pq} - \mu_p \mu_q \rangle && (5)\\ &= \langle \frac{\mu_{pq} - \mu_p \mu_q}{\|\mu_{pq} - \mu_p \mu_q\|}, \mu_{pq} - \mu_p \mu_q \rangle && (6)\\ &= \|\Sigma_{PQ}\|_{HS} && (7) \end{aligned}$$

Step 6 follows because in a Hilbert Space (as in Euclidean space), given a vector u, the unit vector v with largest inner product $\langle v, u \rangle$ is $v = \frac{u}{\|u\|}$. If we further choose "universal" or "characteristic" kernels for $\mathcal{H}_\mathcal{P}$ and $\mathcal{H}_\mathcal{Q}$ than we can guarantee that $P \perp\!\!\!\perp Q$ if and only $\|\Sigma_{PQ}\|_{HS} = 0$[32]. The derivation presented above follows that of the closely related Maximum Mean Discrepancy test statistic[15]. An alternative derivation is to simply consider the Hilbert-Schmidt norm of Σ_{PQ}, and prove that this norm is 0 if and only if $P \perp\!\!\!\perp Q$ (for appropriate choice of kernels). This explains why it is called the Hilbert-Schmidt Independence Criterion[16]. We switch to the squared norm so that we can state a (biased) estimator:

$$\begin{aligned} \|\Sigma_{PQ}\|_{HS}^2 &= \langle \mu_{pq} - \mu_p \mu_q, \mu_{pq} - \mu_p \mu_q \rangle && (8)\\ &= \|\mu_{pq}\|^2 - 2\langle \mu_{pq}, \mu_p \mu_q \rangle + \|\mu_p \mu_q\|^2 && (9)\\ &= E_{x,y} E_{x',y'} k(x, x') \ell(y, y') && (10)\\ &- 2 E_{x,y} E_{x'} E_{y'} k(x, x') \ell(y, y') && (11)\\ &+ E_x E_y E_{x'} E_{y'} k(x, x') \ell(y, y') && (12)\\ &= \frac{1}{n^2} \sum_{i,j} k(x_i, x_j) \ell(y_i, y_j) && (13)\\ &- \frac{2}{n^3} \sum_{i,j,r} k(x_i, x_j) \ell(y_i, y_r) && (14)\\ &+ \frac{1}{n^4} \sum_{i,j,r,s} k(x_i, x_j) \ell(y_r, y_s) && (15)\\ &= \widehat{\mathrm{HSIC}} && (16) \end{aligned}$$

A particularly compact expression for this estimator, using "centered" matrices, is available[15]: $\frac{1}{n^2} \mathrm{tr}(\tilde{K}\tilde{L})$. We present the details here because they will be useful later in understanding connections with the Mantel test. If K is the Gram matrix with entries $k_{ij} = k(x_i, x_j)$ and L is the Gram matrix with entries $\ell_{ij} = \ell(y_i, y_j)$, let $\tilde{K}$ be the centered Gram matrix where $\tilde{K}_{ij} = \langle \phi(x_i) - \mu_p, \phi(x_j) - \mu_p \rangle$ so $\tilde{K} = K - \frac{1}{n}11^T K - \frac{1}{n}K11^T + \frac{1}{n^2}11^T K11^T = HKH$ where $H = I - \frac{1}{n}11^T$. Now, we can expand (using trace rotation and the fact that $HH = H$):

$$\frac{1}{n^2}\operatorname{tr}(\tilde{K}\tilde{L}) = \frac{1}{n^2}\operatorname{tr}(HKH\,HLH) \tag{17}$$
$$= \operatorname{tr}(HKH\,L) \tag{18}$$
$$= \frac{1}{n^2}\operatorname{tr}(KL) - \frac{1}{n^3}\operatorname{tr}(11^T KL) - \frac{1}{n^3}\operatorname{tr}(K11^T L) \tag{19}$$
$$+ \frac{1}{n^4}\operatorname{tr}(11^T K11^T L) \tag{20}$$
$$= \frac{1}{n^2}\operatorname{tr}(KL) - \frac{2}{n^3}\operatorname{tr}(11^T KL) + \frac{1}{n^4}\operatorname{tr}(11^T K11^T L) \tag{21}$$
$$= \frac{1}{n^2}\sum_{i,j} k(x_i, x_j)\ell(y_i, y_j) - \frac{2}{n^3}\sum_{i,j,r} k(x_i, x_j)\ell(y_i, y_r) \tag{22}$$
$$+ \frac{1}{n^4}\sum_{i,j,r,s} k(x_i, x_j)\ell(y_r, y_s) \tag{23}$$
$$= \widehat{\text{HSIC}} \tag{24}$$

The simplest way to derive the distribution of HSIC under the null hypothesis that $X \perp\!\!\!\perp Y$ is by randomization testing: given pairs (x_i, y_i) we shuffle the y's and recompute $\widehat{\text{HSIC}}$. An asymptotic test based on the Gamma distribution is stated in[15], and another test based on the eigenvalues of the kernel matrices is derived in[37]. We also note that[15] gives an unbiased test statistic, based on a U-statistics. As previously stated, for a universal kernel, $P \perp\!\!\!\perp Q$ if and only if $HSIC = 0$. This means that if P and Q have any kind of dependence, then HSIC will not be 0, but if HSIC is 0, then $P \perp\!\!\!\perp Q$.

D. Tests for Space-Time Interaction with Kernel Embeddings

As an intermediate step towards using kernel embeddings to test for space-time interaction, and because it sheds light on the classical version of the Mantel test, we define a kernelized version of the Mantel test. The Mantel test was described in Section II.A. We briefly restate it in a more general form, following[23]. The Mantel test measures the correlation between a pair of dissimilarity (distance) matrices. Given a set of objects P, and two different ways of measuring the dissimilarity between these objects, the null hypothesis is that the two different types of measurements are independent. Given, e.g. two $n \times n$ matrices of distances K and L where $k(i, j)$ gives the Euclidean distance between objects i and j and $\ell(i, j)$ gives some other dissimilarity measure, the Mantel test statistic is $\sum_{i \neq j} k(i, j)\ell(i, j)$. Interestingly, this is the first term in the estimator for HSIC, as shown in Equation 15. While the Mantel test is usually presented in terms of distance matrices, it is valid for similarity matrices as well. We propose considering a kernelized version of the Mantel test. Given objects $P = (p_1, \ldots, p_n)$ and two kernels k and ℓ, we construct the Gram matrices K and L and ask, as in the Mantel test, whether the two kernels are measuring independent properties of the objects of P.

Once we have Gram matrices, we proceed exactly as with the Mantel test, defining the test statistic $T = \sum_{i \neq j} k(i, j)\ell(i, j)$, and obtaining significance levels by randomization testing. We call this test the "kernelized Mantel test" and to our knowledge it has not been explicitly considered in the literature, but in fact, the reciprocal transformation considered by Mantel[25] ($f(s) = \frac{1}{s+\epsilon}$) is an example of a Mercer kernel: $k(x, x') = \frac{1}{\|x - x'\|^2 + \epsilon}$[26] (as cited in[31]).

With this as background, we are ready to define a new test for space-time interaction based on kernel embeddings.

In the spirit of the classical tests described in Section II.A, an alteriatve approach to using HSIC would be to define new distributions $P = \{d_s(i, j) : i \neq j\}$ for the Euclidean distances between pairs of points and $Q = \{d_t(i, j) : i \neq j\}$ for the interpoint time intervals, and apply HSIC as a black box to test whether the distributions P and Q are independent. This is not an attractive option computationally, as it leads to $O(n^4)$ computations because HSIC considers pairs of observations, and in this case observations are themselves pairs of points.

We present two alternative derivations of the same test. First, let us more closely inspect our kernelized Mantel test, and in particular how it differs from the HSIC test statistic. The last term of the HSIC estimator, which estimates $\|\mu_p \mu_q\|^2$, is unchanged by randomization, so the key difference is the cross-term, $\langle \mu_p \mu_q, \mu_{pq} \rangle$. Although it was not presented this way, we could have defined the covariance operator as:

$$\Sigma_{PQ} = E_{xy}[(\phi(x) - \mu_p) \otimes (\psi(y) - \mu_q)]$$

Thus, we see that the cross-term in $\|\Sigma_{PQ}\|^2_{HS}$ arises because the feature vectors $\phi(x)$ and $\psi(y)$ are centered before being multiplied together (by analogy, we can write: $\text{Cov}(P, Q) = E[(P - E[P])(Q - E[Q])]$). Returning to the Mantel test, this is the critical difference—the Mantel test measures dependence by calculating the inner product between two matrices treated as vectors, where these vectors are centered by subtracting the mean of their entries, that is, subtracting the mean of empirical distribution over distances But this is *not* equivalent to the centering done by HSIC: $\tilde{\phi}(x) = \phi(x) - \mu_p$ centers the feature embedding so that it has mean 0. From this, the centered Gram matrices $\tilde{K} = HKH$ and $\tilde{L} = HLH$ where $H = I - \frac{1}{n}11^T$ is calculated, and then the covariance is measured as $\frac{1}{n^2}\operatorname{tr}\tilde{K}\tilde{L}$, as proved in the previous section.

This suggests a simple fix for the Mantel test, which can even be applied to the classic version. Given similarity, dissimilarity, or Gram matrices K and L, calculate $\tilde{K}$ and $\tilde{L}$ and then apply the Mantel test: $\sum_{i,j} \tilde{K}_{ij}\tilde{L}_{ij}$. Since this is proportional to $\frac{1}{n^2}\operatorname{tr}(\tilde{K}\tilde{L})$, our final "Kernelized Space-Time" (KST) test takes the same form as HSIC.

We now show an alternative derivation of this test, starting with Hilbert spaces. Given a probability distribution over points in space $A = \{(s, t)\}$, with $s \in \mathcal{R}^2$ and $t \in \mathcal{R}$ and kernels k (for RKHS $\mathcal{H}_\mathcal{K}$) and ℓ (for RKHS $\mathcal{H}_\mathcal{L}$). Let $k(a, \cdot) = k(s, \cdot) = \phi(s)$ and $\ell(a, \cdot) = \ell(t, \cdot) = \psi(t)$, so that k embeds the spatial coordinates of A with feature map $\phi(s)$, ignoring the temporal coordinates, and ℓ embeds the temporal coordinates of A with feature map $\psi(t)$, ignoring the spatial coordinates. The null hypothesis we wish to test is that for a random $a \sim A$, $\phi(a) \perp\!\!\!\perp \psi(a)$, i.e. space and time are independent.

We proceed in feature space to test whether $P(\phi(p), \psi(p)) = P(\phi(p))P(\psi(p))$. For arbitrary functions $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$, we wish to test whether $\sup_{f,g} \mathrm{Cov}_a(\langle f, \phi(a)\rangle, \langle g, \psi(q)\rangle) = 0$. This expression is equal to Equation 8, so we reach the same conclusion, that we need to check whether $\|\Sigma_{PQ}\|_{HS} = 0$, which can be done using $\widehat{\mathrm{HSIC}}$ as a test statistic. Note that if we want to stay as close as possible to classical tests for space-time interaction, we could insist that k and ℓ be stationary so that $k(s, s') = k(\|s - s'\|)$ and $\ell(t, t') = \ell(|t - t'|)$, but this is not necessary.

Let us recap. For intuition, let k and ℓ be RBF kernels (with the same bandwidth for convenience): $k(a, a') = e^{-\|s-s'\|_2^2}$ and $\ell(a, a') = e^{-|t-t'|^2}$. Given the space-time coordinates of a set of points, we wish to test whether there is space-time interaction. Using kernels, we represent these points through their similarity to every other point, i.e. we represent these points using $\phi(a) = k(a, \cdot)$—a measure of the spatial distance between a and any other point and by $\psi(a) = \ell(a, \cdot)$—a measure of the time interval between a and every other point. Given these representations, we proceed just as in the classical tests, asking whether the distribution over spatial distances $P(\phi(a))$ is independent of the distribution over time intervals $P(\psi(a))$. Unlike the standard HSIC setting, we are already working in feature space, but the test is still valid, and it still provides the same guarantees, namely space and time (as represented by the kernels we picked, and provided the kernels are universal or characteristic) are independent if and only if $HSIC(\phi(A), \psi(A)) = 0$.

E. Extensions for Bivariate Space/Time Interactions

For the application domain we consider, we are interested in determining whether there is space-time interaction between two types of point processes.

As previously discussed in Section II.B, the extension of the classical tests to this bivariate case is straightforward. Let $\mathcal{P}^1$ and $\mathcal{P}^2$ denote two different point processes, with the same number of observations n for convenience. Call pairs of points (p^1, p^2) cross-pairs of points. For the Knox test, the contingency table simply counts the number of cross-pairs of points that are close vs. far. For the Mantel test, we consider only the submatrices of the original distances matrices corresponding to these cross-pairs. (See below.)

In Hilbert space, we can develop a similar extension, and again we can motivate it in a few ways.

Motivation 1. If we take as our starting point the Mantel test, then we have (distance, similarity, or Gram) matrices K and L. Let the submatrix K^{12} denote the matrix with entries $K^{12}_{ij} = k(x^1_i, x^2_j)$ where x^1_i is a point of type 1 and x^2_j is a point of type 2. We define L^{12} in the same way. Now, the Mantel test statistic would be $\sum_{ij} K^{12}_{ij} L^{12}_{ij}$. As we argued previously, this test statistic is not appropriately centered, and it is thus equivalent to taking an uncentered covariance. Moving to feature space, we can define a centered version of $\tilde{K}^{12}_{ij} = \langle \phi(x^1_i) - \mu^1_p, \phi(x^2_j) - \mu^2_p\rangle$ where $\mu^1_p = E_{x^1}\phi(x^1)$ and $\mu^2_p = E_{x^2}\phi(x^2)$. Notice that these entries are not the same as the equivalent cross-pair entries in the $\tilde{K}$ matrix previously defined because the centering is by μ^1_p for points of type 1 and μ^2_p for points of type 2. We define L^{12} analogously, and thus propose as our kernel-based test statistic: $\widehat{\mathrm{KST}}_{12} = \sum_{ij} \tilde{K}^{12}\tilde{L}^{12}$. This can be written compactly in matrix form as $\mathrm{tr}(K^{12}HL^{12}H)$ where $H = I - \frac{1}{n}11^T$.

Motivation 2. If we take as our starting point HSIC, we can define $\Sigma^1_{XY} = \mu^1_{pq} - \mu^1_p\mu^1_q$ and $\Sigma^2_{XY} = \mu^2_{pq} - \mu^2_p\mu^2_q$. Before, our test statistic was the Hilbert-Schmidt norm of the operator Σ_{XY}. Now, we propose taking the trace of the product $\Sigma^1_{XY}(\Sigma^2_{XY})^T = \langle \mu^1_{pq} - \mu^1_p\mu^1_q, \mu^2_{pq} - \mu^2_p\mu^2_q\rangle$. The idea is that Σ^1_{XY} is the Hilbert space operator representing the space-time interaction of points of type 1. When we evaluate $\langle f, \Sigma^1_{XY} g\rangle$ for some choice $f \otimes g$ we are calculating $Cov(f(X^1), g(Y^1))$. Now, rather than finding the supremum of this covariance (equivalently, calculating the Hilbert-Schmidt norm), we check it at a particular choice of $f \otimes g$: if $f \otimes g = \Sigma^2_{XY}$, then

$$\langle f, \Sigma^1_{XY} g\rangle = \Sigma^1_{XY}(\Sigma^2_{XY})^T$$

. In this way, we are evaluating the covariance in feature space by only considering cross-pairs of points. To see that this is the case, and that it gives the same test statistic as above, we expand using expectations:

$$\Sigma^1_{XY}(\Sigma^2_{XY})^T = \langle \mu^1_{pq} - \mu^1_p\mu^1_q, \mu^2_{pq} - \mu^2_p\mu^2_q\rangle \tag{25}$$

$$= E_{x^1,y^1}E_{x^2,y^2}[k(x^1, x^2)\ell(y^1, y^2)] \tag{26}$$

$$- E_{x^1,y^1}[E_{x^2}k(x^1, x^2)E_{y^2}\ell(y^1, y^2)] - E_{x^2y^2}[E_{x^1}k(x^1, x^2)E_{y^1}\ell(y^1, y^2 \tag{27}$$

$$+ E_{x^1}E_{x^2}k(x^1, x^2)E_{y^1}E_{y^2}\ell(y^1, y^2) \tag{28}$$

Notice that throughout the kernels k and ℓ are only applied to cross-pairs of points. An estimator is:

$$\widehat{\mathrm{KST}}_{12} = \frac{1}{n^2}\sum_{i,j} k(x^1_i, x^2_j)\ell(y^1_i, y^2_j) - \frac{1}{n^3}\sum_{i,j,r} k(x^1_i, x^2_j)\ell(y^1_i, y^2_r) - \frac{1}{n^3}\sum_{i,j,q} k(x^1_i, x^2_j$$

As with the original formulation of HSIC, some algebra remains to verify that this is indeed equal to $\frac{1}{n^2}\mathrm{tr}(\tilde{K}^{12}\tilde{L}^{12})$, the KST_{12} test statistic we proposed earlier.

F. Forward in Time Tests for Space-Time Interaction

In the previous section we showed how to measure space-time interaction between two types of points. We propose one further extension: we wish to measure space-time interaction in the bivariate case, where we insist that points of type 2 occur after points of type 1. We refer to this as bivariate, forward in time case $\mathrm{KST}_{1\to 2}$. As in the previous section, the extension follows by restricting the test statistic to consider only pairs of points (t^1_i, t^2_j) for which $t^1_j < t^2_j$. We can immediately modify our test statistic:

$$\widehat{\mathrm{KST}}_{1\to 2} = \frac{1}{N}\sum_{i,j:y^1_i<y^2_j} k(x^1_i, x^2_j)\ell(y^1_i, y^2_j)$$

$$- \frac{1}{nN}\sum_{i,j,r:y^1_i<y^2_r} k(x^1_i, x^2_j)\ell(y^1_i, y^2_r) - \frac{1}{nN}\sum_{i,j,q:y_q<y_j} k(x^1_i, x^2_j)\ell(y^1_q, y^2_j)$$

$$+ \frac{1}{n^2N}\sum_{i,j,q,r:y^1_q<y^2_r} k(x^1_i, x^2_j)\ell(y^1_q, y^2_r)$$

where $N = \sum_{i,j} I(y^1_i < y^2_j)$.

III. EXPERIMENTAL EVALUATION

Below, we describe three experimental evaluations of our new space-time test. First, using synthetic data, we compare the performance of our test as compared to the classical tests. Second, we show the applicability of our methods to the so-called "broken windows" hypothesis: using publicly available data on calls for service and crime incidents, we ask whether and which types of citizen complaints are correlated with violent crime. Finally, we turn our attention to the problem of predictive policing and evaluate whether our test can be used as a feature selection method to determine which leading indicators of 911 call types are predictive of shootings and homicides. We argue that the correlations uncovered by our model are more useful and interpretable than those discovered by standard sparse regression methods, while still giving comparable performance on the difficult task of predicting incidents of shootings and homicides across Chicago neighborhoods.

A. Synthetic Data

Our power analysis, inspired by the one in[9] uses the following setup for a Poisson cluster process: parent locations (x, y, t) are sampled on the unit cube. The number of children for each parent is drawn iid $\sim$ Poisson(5). The location of each child is generated as a random displacement from the parent's location, in space and time, where each coordinate's offset is independently sampled from $N(0, \sigma)$. This induces space-time interaction, and as σ increases, the signal of this interaction becomes swamped by noise. Figure 2 shows two examples, one with $\sigma = 0.025$ and the other with $\sigma = 0.2$. We consider $0 < \sigma \leq .4$.

We will consider the same set of tuning parameters $\Delta = \{0.05, 0.10, \ldots, 0.25\}$ for each test. For the Knox test, the spatial cutoff varies over Δ while the temporal cutoff is fixed at 0.1. For the Mantel test, we use the transformation considered earlier: $\frac{1}{x+\epsilon}$ for $\epsilon \in \Delta$. For the Diggle et al. test, we follow[9] and use a grid of side length varying over Δ for the points at which the K function is evaluated. The grid always has the same coarseness 0.01. For KST, the bandwidth σ of the RBF kernel varies over Δ. For each method, each value of Δ, and each value of σ, we draw 500 random point patterns and obtain p-values using randomization testing. The power is shown in Figure 4 as the fraction of simulations which correctly rejected the null hypothesis of independence between space and time at $\alpha = 5\%$. The four methods are compared in Figure 5. For each method, the relevant parameter that was chosen was the parameter with the highest power for $\sigma = 0.15$. When σ is small, all methods have equally high power, but as σ increases, the power decreases at different rates. The KST method we proposed has the highest power for $\sigma > 0.1$.

B. The "Broken Windows" Theory

As introduced in an influential magazine article by Wilson and Kelling[36], the "broken windows" theory suggests that low-level disorder, such as broken windows, leads to more low-level disorder (e.g. property crime), and eventually more serious, violent crime. There has been much research and debate about this theory. See, for example[30]. As an exploratory example, we demonstrate the use of our space-time test towards addressing this controversial hypothesis. We obtained geocoded, date-stamped non-emergency calls for service to Chicago's 311 number and geocoded, date-stamped criminal offense reports from January 2010 through December 2012. Continuously updated versions of both datasets are publicly available at Chicago's data portal, data.cityofchicago.org.

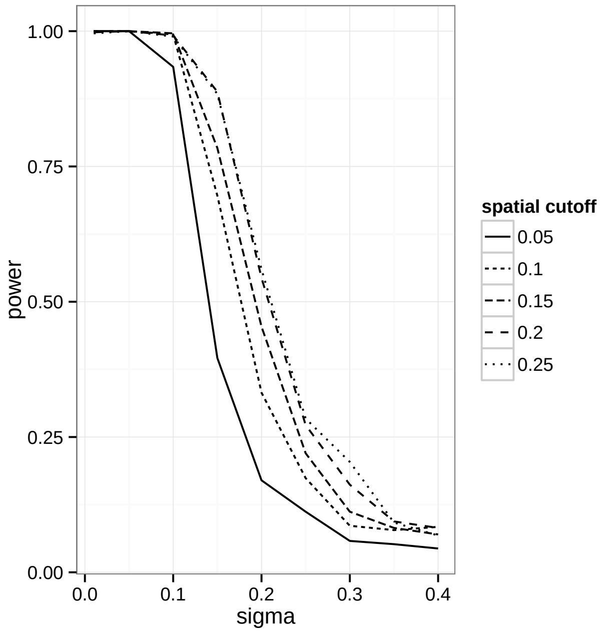
We used $\text{KST}_{1\to2}$, the bivariate, forward in time version of our space-time test to check whether any types of calls to 311 exhibit space-time interaction with crime in the future. The types of 311 calls considered were: abandoned vehicle, alley light out, garbage cart–related, graffiti, pot hole, rodent complaint, sanitation, street lights out, tree trim / tree debris, and vacant/abandoned building. We aggregated the following types of crimes into a single category, violent crime: homicide, criminal sexual assault, robbery, and aggravated assault / battery. We aggregated all other crime types into a category we call non-violent crime. With three years of data, our sample sizes range from about 30,000 (vacant/abandoned building reports) to more than 750,000 (non-violent crimes). There were almost 300,000 violent crimes reported. Our implementation of KST requires time $O(n^3)$ in the number of points n per permutation, so we used independent random thinning (randomly deleting points with a fixed probability, equivalent to taking a random subset of the points without replacement) to speed up computation times

Most statistics on spatiotemporal point patterns are invariant to independent random thinning. We performed sensitivity analyses to ensure that our estimates were robust to the size of the subsample chosen.

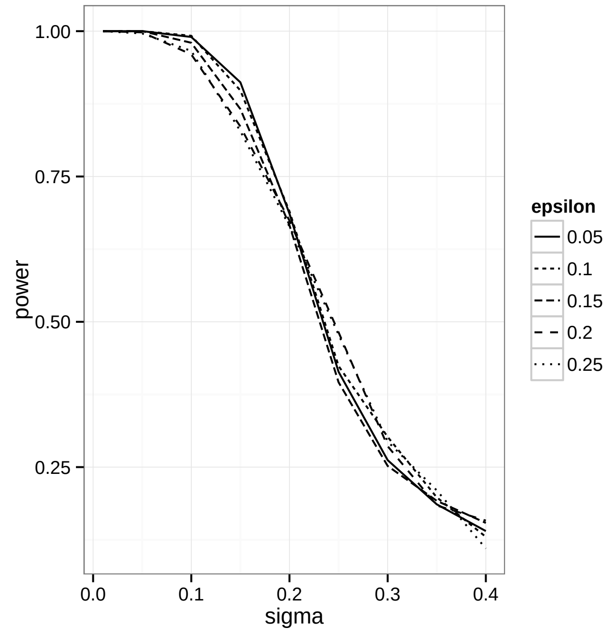
We used sample sizes $n = 50{,}000$ for each point pattern and randomization testing with 1000 replications, to obtain p-values for each pair of 311 call type and violent or non-violent crime (20 hypotheses in total), as shown in Table II. No types of call to 311 had a significant p-value, meaning that after controlling for spatial and temporal variation there is no (statistically significant) correlation between the incidence of 311 calls and either non-violent or violent crime in Chicago. A much deeper analysis would of course be necessary to provide a rigorous of assessment of the broken windows theory using these data. One immediate objection to this formulation is that 311 data is evidence of disorder as perceived by engaged citizens, rather than an objective measure of disorder, or disorder as perceived by criminals. And of course, correlation does not imply causation (although statistical independence usually rules out causation!)

C. Crime Data from the City of Chicago

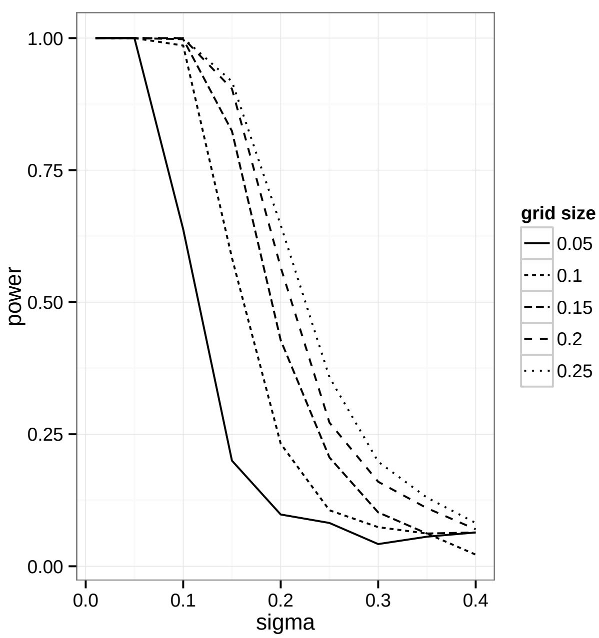
Based on conversations with City of Chicago officials, we focused on the following question: which types of calls to 911 exhibit space-time interaction with homicides and aggravated battery with a handgun (hereinafter, "shootings"). The ultimate goal is to enable the Chicago Police Department to respond more pro-actively to these leading indicators so as to prevent shootings. We used a dataset provided by the Chicago Police Department of geocoded, date and time-stamped calls to 911 and another dataset of crime incident reports. Data



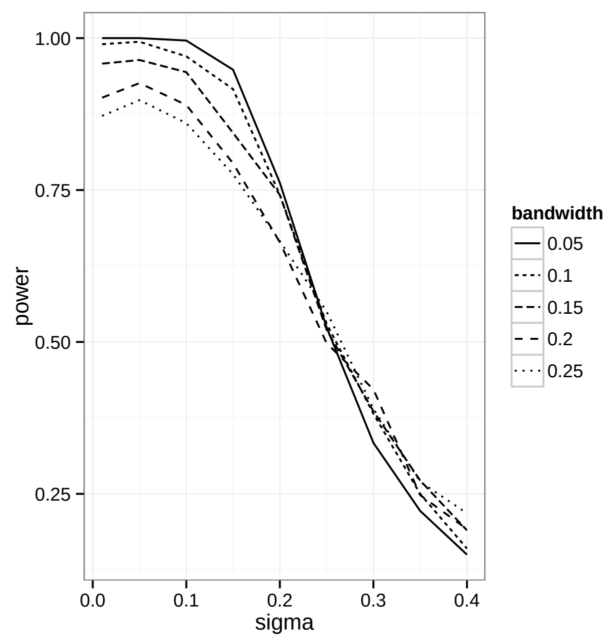
(a) Knox



(b) Mantel



(c) Diggle



(d) KST

FIG. 4: We compare the Knox, Mantel, Diggle et al., and newly-proposed KST tests on synthetic data. On the y-axis, we show the power as the fraction of simulations in which the test correctly rejected the null hypothesis of independence between space and time for $\alpha = 5\%$. In the simulations, a cluster point process is generated where children points are offset from a parent a distance $\sim N(0, \sigma)$ in each dimension. As σ grows, the problem becomes harder and each method's power decreases. For each method, we vary a tuning parameter: for Knox we vary the definition of "near" in space, for Mantel we vary the ϵ in the reciprocal transformation, for Diggle et al. we vary the grid size, and for KST we vary the bandwidth of the RBF kernel.

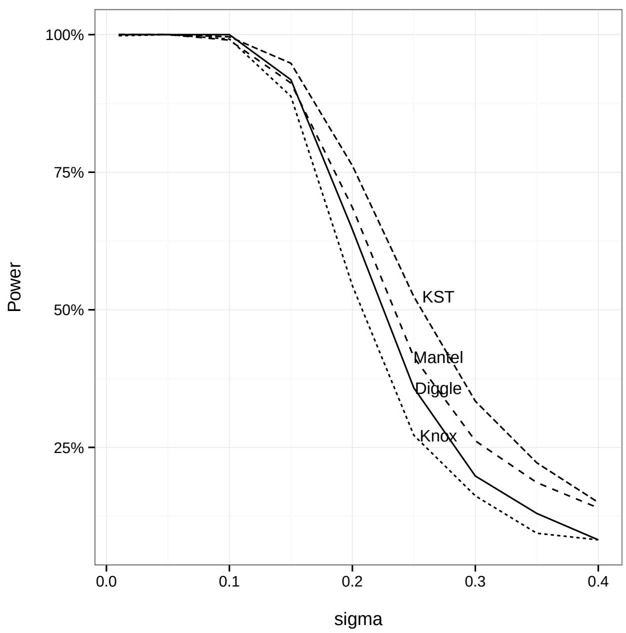


FIG. 5: The four methods from Figure 4 are shown here for comparison. For each method, the relevant parameter that was chosen was the parameter with the highest power for $\sigma = 0.15$. When σ is small, all methods perform equally well, but as σ increases, we see differences. The method we propose has the highest power for $\sigma \geq 0.1$.

were from 2007 through the first half of 2010. There were just under 9 million calls for service (911 calls, plus calls to the dispatcher made by police for a variety of reasons) and 9,087 homicides and shootings reported in this time period. There were about 275different types of calls to 911.

As in the previous section, we considered each separate type of 911 call as one point pattern and the point pattern of shootings as the other, and we $\text{KST}_{1\to 2}$ to look for significant space-time interactions. We used randomization testing with 1000 repetitions. We report call types which were significant for $p \leq 0.01$ in Table III.

From a predictive policing point of view, space-time tests can serve as a feature selection method. As an evaluation, we focused on the problem of predicting shootings at the police beat level. Chicago has 25 police districts, divided into 280 beats. The median area of a beat is 0.57 square miles, equivalent in area to a square with side length 6 city blocks (in Chicago, 8 city blocks are about a mile.) Chicago's community policing efforts occur at the beat-level.

We built a simple logistic regression model, meant to model the heat maps currently employed by many police departments. As features, for each type of 911 call j, for each day of data t, at the centroid of each police beat i, we used the previous 90 days of data to create a predictor X_{it}^j capturing an estimate of the intensity of each type of call, using (unnormalized) kernel density estimation with RBF kernels with $\sigma = \frac{1}{4}$ mile for space and 14 days for time. We used the same method to create a smoothed predictor for shootings S_{it}, using only the last week of data. The dependent variable Y_{it} was coded as binary: did a shooting occur on a given day in a given police beat? Our model is as follows:

$$\text{logit}(Y_{it}) \sim \beta_0 + \alpha S_{it} + \beta_1 X_{it}^1 + \ldots + \beta_p X_{it}^p + \epsilon$$

We split our data into a training set (January 2007-December 2008) and a test set (January 2009 - May 2010).

We could include every possible type of call to 911 as a predictor—with 3.5 years of data and 280 beats, our data is plentiful—and fit the full model using maximum likelihood. However, our model is quite simplistic as it does not account for spatial and temporal autocorrelation, so our effective sample size is certainly lower than the number of rows in the dataset we generated. Thus, we prefer a more parsimonious model; simpler models also have a better chance of actually being used by a police department.

For these reasons, we compare two scenarios. In the first, we use the Lasso to fit an L_1-regularization path to the full training dataset[34] as implemented in the R package glmnet[12]. In the second, we preprocess our dataset, selecting only the features that we found to be significant using our space-time test as shown in Table III. We used RBF kernels with bandwidths equivalent to the bandwidth of the smoothing kernel, $\frac{1}{4}$ mile for space and 14 days for time. The setup is the same as in the previous section, this time with random thinning to obtain subsamples of size $n = 50,000$. Finally, we use the Lasso on this smaller dataset. In both cases, we calculate the true positive rate (TPR) of our model for a fixed false positive rate of 10%. (This corresponds to the probability that the model will correctly predict a shooting on a certain day in a certain police beat while incorrectly predicting that there will be shootings in 10% of the beat-days.) We show the TPR across the regularization path, that is, as the number of features in the model increases. As shown in Figure 6, the two models' performance is pretty much indistinguishable on this metric (as a sensitivity analysis, we tried a variety of bandwidths for the smoothing kernel and the Mercer kernel, without discovering any notable gains in performance.) We can also compare the list of predictors in the order in which they enter the model. For the preprocessed features selected by KST, the first ten features to enter the model were: shots fired, shooting (the lagged version of the dependent variable), officer pursuing someone on foot, office heard shots fired, narcotics loitering, officer station assignment, person shot, meeting of the police beat unit, support unit request, gang loitering. The first ten features to enter the full model were: shots fired, domestic disturbance, person with a gun, shootings (the lagged version of the dependent variable), officer eating lunch, vicious animal, parking violation, gang disturbance, gambling, battery in progress. Both models seem to be picking up on important predictors, but Lasso without preprocessing seems to also be selecting some spurious predictors, e.g. police officers breaking for lunch and parking violations.

IV. DISCUSSION

A. Conclusion

In this paper we developed new statistical tests for space-time interaction using kernel-based statistics for measuring the distance between probability distributions, and compared their performance to classical tests due to Knox, Mantel, and

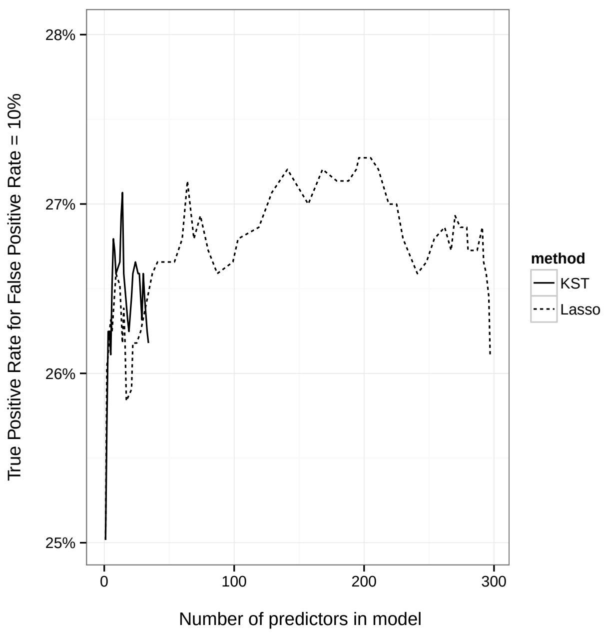


FIG. 6: We compared two models for predicting shootings and homicides based on kernel-smoothed 911 calls. There were about 275 different 911 call types (predictors), and both models used L_1 penalized-logistic regression on a training data set. In the "Lasso" case, no preprocessing on the predictors was done. In the "KST" case, the bivariate forward-in-time KST_{12} test was used to select features with a significant space-time interaction with the dependent variable (incidents of shootings and homicides). Subsequently, Lasso was used. The true positive rate is shown at a fixed false positive rate of 10% as the number of predictors increases.

p-value	311 call type
0.962	Abandoned Vehicle
0.317	Alley Light Out
0.193	Garbage Cart
0.042	Graffiti
0.974	Pot Hole
0.179	Rat
0.714	Sanitation
0.369	Street Lights
0.909	Tree
0.154	Vacant

(a) Space-time interaction tests with violent crime

p-value	311 call type
0.039	Abandoned Vehicle
0.757	Alley Light Out
0.652	Garbage Cart
0.111	Graffiti
0.108	Pot Hole
0.668	Rat
0.314	Sanitation
0.806	Street Lights
0.589	Tree
0.091	Vacant

(b) Space-time interaction tests with non-violent crime

TABLE II: Controlling for spatial and temporal variation, we used KST_{12} to test whether any types of 311 calls are predictive of violent crime nearby in space and time. Randomization testing was performed with 1000 repetitions. For $\alpha = .01$, no p-values were significant, meaning that none of the types of calls to 311 showed significant correlation with crime that could not be explained by underlying spatial and temporal trends.

Diggle et al. Our new test outperformed the existing tests. We illustrated the use of our test to the "broken windows" theory, finding that none of the call types placed to Chicago's non-emergency services number were significant predictors of either violent or non-violent crime. We also demonstrated the use of our methods in a predictive policing application, searching for leading indicators for shootings and homicides among 911 call types. Our method served as a feature selection, in which its performance in selecting features for a prediction problem was comparable to the Lasso.

B. Acknowledgments

Thanks to Joseph Candella, David Choi, Brendan O'Connor, Wilpen Gorr, Edward McFowland, Aaditya Ramdas, Sriram Somanchi, Skyler Speakman, and Nathan VanHoudnos for helpful discussions. Thank you to the Chicago Police Department for sharing data. This work was partially supported by the National Science Foundation, grants IIS-0916345, IIS-0911032, and IIS-0953330. Points of view or opinions contained within this paper are those of the authors and do not necessarily represent the official position or policies of the Chicago Police Department or the NSF.

[1] FE Alexander, P Boyle, PM Carli, JW Coebergh, GJ Draper, A Ekbom, F Levi, PA McKinney, W McWhirter, C Magnani, et al. Spatial temporal patterns in childhood leukaemia: further evidence for an infectious origin. euroclus project. *British journal of cancer*, 77(5):812, 1998.

[2] Rose D Baker. Testing for space-time clusters of unknown size. *Journal of Applied Statistics*, 23(5):543–554, 1996.

[3] M. S. Bartlett. The detection of space-time interactions (commentary). *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 13(1):pp. 30, 1964. ISSN 00359254. URL http://www.jstor.org/stable/2985220.

[4] Alfred Blumstein and Joel Wallman. *The crime drop in America*. Cambridge University Press, 2005.

[5] Anthony Allan Braga and David L Weisburd. *Policing problem places: Crime hot spots and effective prevention*. Oxford University Press, USA, 2010.

[6] P.J. Brantingham and P.L. Brantingham. *Environmental criminology*. Sage focus editions. Waveland Press, Incorporated, 1991. ISBN 9780881335392. URL http://books.google.com/books?id=gNQMAAAACAAJ.

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

[8] Lin Cui. Foreclosure, vacancy and crime. *Available at SSRN 1773706*, 2010.

[9] P.J. Diggle, A.G. Chetwynd, R. Häggkvist, and SE Morris. Second-order analysis of space-time clustering. *Statistical methods in medical research*, 4(2):124–136, 1995.

[10] Mingtao Ding, Lihan He, David Dunson, and Lawrence Carin. Nonparametric bayesian segmentation of a multivariate inhomogeneous space-time poisson process. *Bayesian Analysis*, 7(4):813–840, 2012.

[11] Ingrid Gould Ellen, Johanna Lacoe, and Claudia Ayanna Sharygin. Do foreclosures cause crime? *Journal of Urban Economics*, 2012.

p-value	911 call type	p-value	911 call type
0.002	STREETS & SAN PINK CARD	0.002	MENTAL UNAUTH ABSENCE
0.002	PERSON SHOT	0.002	DEATH REMOVAL
0.002	WALK DOWN	0.002	SHOTS FIRED (OV)
0.002	ASSAULT IP	0.002	CRIMINAL TRES. (OV)
0.002	EVIDENCE TECHNICIAN (PRI. 1)	0.002	ARSON REPORT
0.002	AUTO THEFT IP	0.002	TASTE OF CHICAGO
0.002	EVIDENCE TECHNICIAN (PRI. 3)	0.002	AMBER ALERT
0.002	PERSON WITH A GUN	0.004	DETAIL
0.002	MISSION	0.004	GANG DISTURBANCE
0.002	PERSON WANTED	0.004	PURSUIT FOOT (OV)
0.002	PERSON STABBED	0.006	BATTERY IP
0.002	SHOTS FIRED	0.006	NOTIFY
0.002	EVIDENCE TECHNICIAN (PRI. 2)	0.008	ON VIEW
0.002	PLAN 1-5	0.008	CRIM DAM. TO PROP IP
0.002	K9 REQUEST	0.008	RECOVERED STOLEN AUTO
0.002	OUTDOOR ROLL CALL	0.010	THEFT IP
0.002	CRIM DAM. TO PROP RPT	0.010	CRIM DAM. TO PROP (OV)
0.002	HOLDING OFFENDER (CITZ.)	0.010	MUNICIPAL ORD. VIOLATION

TABLE III: Which 911 call types predict shootings / homicide? Using data from 2007 through the first half of 2010, we used $\text{KST}_{1\to 2}$ to test for forward in time space-time interaction of each pair: 911 call type as points of type 1, shootings/homicides as points of type 2. We used a Gaussian RBF kernel with bandwidth $\frac{1}{4}$ mile and 14 days. P-values were obtained by randomization testing with 500 permutations. Results which were significant for $p \leq .01$ are shown.

[12] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[13] Montserrat Fuentes. Testing for separability of spatial–temporal covariance functions. *Journal of Statistical Planning and Inference*, 136(2):447–466, 2006.

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

[15] A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schoelkopf, and A. Smola. A kernel statistical test of independence. 2008. URL http://books.nips.cc/papers/files/nips20/NIPS2007_0730.pdf.

[16] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

[17] Tony H. Grubesic and Elizabeth A. Mack. Spatio-temporal interaction of urban crime. *Journal of Quantitative Criminology*, 24(3):285–306, 2008. ISSN 0748-4518. doi: 10.1007/s10940-008-9047-5. URL http://dx.doi.org/10.1007/s10940-008-9047-5.

[18] GEOFFREY M Jacquez. A k nearest neighbour test for space–time interaction. *Statistics in medicine*, 15(18):1935–1949, 1996.

[19] Melville R Klauber. Two-sample randomization tests for space-time clustering. *Biometrics*, pages 129–142, 1971.

[20] E. G. Knox. The detection of space-time interactions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 13(1): pp. 25–29, 1964. ISSN 00359254. URL http://www.jstor.org/stable/2985220.

[21] M. Kulldorff. A spatial scan statistic. *Communications in statistics-theory and methods*, 26(6):1481–1496, 1997.

[22] Martin Kulldorff and Ulf Hjalmars. The knox method and other tests for space-time interaction. *Biometrics*, 55(2):544–552, 1999.

[23] Pierre Legendre and Louis Legendre. *Numerical ecology*. Elsevier, 2012.

[24] H.J. Lynch and P.R. Moorcroft. A spatiotemporal ripley's k-function to analyze interactions between spruce budworm and fire in british columbia, canada. *Canadian Journal of Forest Research*, 38(12):3112–3119, 2008.

[25] Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220, 1967.

[26] CharlesA. Micchelli. Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2(1):11–22, 1986. ISSN 0176-4276. doi: 10.1007/BF01893414. URL http://dx.doi.org/10.1007/BF01893414.

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

[28] D.B. Neill and W.L. Gorr. Detecting and preventing emerging epidemics of crime. *Advances in Disease*, 2007.

[29] B.D. Ripley. Edge effects in spatial stochastic processes. *Statistics in theory and practice*, pages 247–262, 1982.

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

[31] César R Souza. Kernel functions for machine learning applications, 2010.

[32] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 99:1517–1561, 2010.

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

[34] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[35] Michael Townsley, Ross Homel, and Janet Chaseling. Infectious

burglaries. a test of the near repeat hypothesis. *British Journal of Criminology*, 43(3):615–633, 2003.

[36] James Q Wilson and George L Kelling. Broken windows. *Atlantic monthly*, 249(3):29–38, 1982.

[37] K. Zhang, J. Peters, D. Janzing, and B. Schoelkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.