Correlates of homicide: new space/time interaction tests for spatiotemporal point processes*

Seth R. Flaxman†‡, Daniel B. Neill†, Alex J. Smola‡
sflaxman@cs.cmu.edu, neill@cs.cmu.edu, alex@smola.org

April 23, 2013

Abstract

Statistical inference on spatiotemporal data often proceeds by focusing on the temporal aspect of the data, ignoring space, or the spatial aspect, ignoring time. In this paper, we explicitly focus on the interaction between space and time. Using a geocoded, time-stamped dataset from Chicago of almost 9 millions calls to 911 between 2007 and 2010, we ask whether any of these call types are associated with shootings or homicides. Standard correlation techniques do not produce meaningful results in the spatiotemporal setting because of two confounds: purely spatial effects (i.e. "bad" neighborhoods) and purely temporal effects (i.e. 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. We propose a new test for space-time interaction, using a Mercer kernel-based statistic for measuring the distance between probability distributions. We compare our new test to existing tests on simulated and real data, where it performs comparably to or better than the classical tests. For the application we consider, we find a number of interesting and significant space-time interactions between 911 call types and shootings/homicides.

1 Introduction

What is the relationship between home foreclosures and violent crime [8, 12]? It might be surprising at first to consider that simply estimating a meaningful measure of association correctly is non-trivial. We illustrate this point in Figure 1: in the left panel, we have plotted the time series of violent crimes and calls to Chicago's 311 non-emergency services number about vacant/abandoned buildings (which was 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 statistical test for space-time interaction, 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 0.68, 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 dataset consisting of location and time-coded events is referred to as a "spatiotemporal point process," as each observation is a point in space/time, and together these observations are stochastic. A key feature of spatial and temporal data is the non-independent nature of the observations. Valid statistical inference requires accounting

*This is a Heinz College working paper. See http://www.heinz.cmu.edu/faculty-and-research/research for updates.

†Event and Pattern Detection Laboratory, H.J. Heinz III College, Carnegie Mellon University
‡Machine Learning Department, School of Computer Science, Carnegie Mellon University

for this dependence structure in the data: the question of scientific interest is to detect space-time interaction after controlling for separable dependencies. In the application area we consider, predictive policing, underlying spatial patterns (e.g., "bad neighborhoods" and varying population density) 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, once purely spatial and purely temporal dependencies are removed.

Knox [17], Mantel [21], and Diggle et al. [9] all developed important and widely used tests for space-time interaction. While each 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 this work is to try 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 the distributions of 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) [13], 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 "pair" case of deciding whether two spatiotemporal point processes have space-time interaction, once purely spatial and purely temporal dependence is accounted for. We extend the HSIC-based criterion to the pair case, and show interesting connections with and a new twist on the Mantel 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 space-time interaction with shootings, using geocoded, date-stamped crime data from the City of Chicago. Finally, we take crime incident reports from Chicago's publicly available crimes dataset[1] and ask whether any of them could be individually considered "infectious." [2]

1.1 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" hypothesis [29], 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 fancy spatiotemporal models have been fit to crime data: [22] used self-exciting point process models, developed for earthquake modeling, for burglaries. [28] modeled violent crime using sophisticated semiparametric Bayesian techniques. [10] used nonparametric Bayesian techniques to segment crime data spatially. Explicitly addressing space-time interaction, [16] compared the "spatio-temporal 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 scan statistics [24].

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 mine appeared in ecology: [20] asked whether forest fires and spruce budworms have space-time interaction using point process methods.

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

[2] While this terminology is suggestive of a causal process, we are focused in this work on statistical tests for independence, not causal inference. When we find a lack of independence and conclude that there is evidence for space-time interaction in a point process or pair of processes, we are claiming correlation, not causation.

1.2 Contributions

First, no previous work in criminology has combined the focus on leading indicators with space-time interaction tests. We argue that this perspective is an important contribution to the literature on leading indicators. Second, the space-time interaction test that we develop is a novel extension of HSIC to the multivariate spatiotemporal point process domain. It gives a new perspective on the classical Mantel test, provides a new alternative to classical tests for space-time interaction, and shows how HSIC can be used to compare pairs of point processes.

2 Theoretical Development

2.1 Background: Classical Tests for Space-Time Interaction

Let $\mathcal{P} = \{(s_i, t_i), i = 1, \ldots, n\}$ be 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** [17]. 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 $d_s(s, s') \leq s_0$ and $d_t(s, s') \leq t_0$:

	near in space	far in space
near in time	X	n_1
far in time	n_2	n_3

If there are $N = n(n-1)/2$ pairs of points, the test statistic is given by the difference between the number of points that we observe to be near in both time and space, X, and the number of points 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 randomizing 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 $X + n_2$. Thus, we could simplify our calculations and consider the test statistic as simply X.

The Knox test is very straightforward, but it clearly has limitations. Correctly specifying the spatial and temporal ranges is not always easy, and testing a range of values leads to problems of multiple hypothesis tests. Another concern is that the 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 made this point explicit, stating that *all* of the information required for a test of space-interaction is found in the interpoint time and space distances[17]. 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** [21]. Given the set of points $\mathcal{P}$, 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. Similarly to the test statistic X for Knox, we compute $\sum_{i,j} d_s(p_i, p_j)d_t(p_i, p_j)$ (equivalently, we can view each matrix as a vector, and find the dot product between D_S and D_T). This is like a correlation, and versions of Pearson's and Spearman's correlation have been applied. Notice, however, that the usual significance test for Pearson's correlation is not valid, because the entries in the matrix are not independent. To derive the null distribution, we again turn to randomization testing, this time applying a given permutation to both the rows and columns of one of the matrices, so as to preserve the dependence structure among the entries.

Notice that 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 [21], Mantel proposed the reciprocal transformation for both spatial and temporal distances x. Mantel argued that by turning a distance measure into a measure of closeness he was focusing on shorter rather than longer ranges. To deal with the problem of taking the reciprocal of 0, Mantel proposed adding a small constant first, before taking the reciprocal. Notice further that the Mantel test reduces to the Knox test with thresholds s_0 and t_0 if the transformation function f is the indicator functoin: $f(d_s) = I(d_s \leq s_0)$ and similarly to d_t.

Diggle's test [9] has a similar flavor to the Knox test, but rather than a single threshold value, Diggle's test requires the specification of a range of values. First, we define Ripley's K function (also called the reduced second moment measure) for a single spatial point process as the following:

$$K(s) = \frac{1}{\lambda_S} E[\text{\# of events occurring within a distance } s \text{ of an arbitrary event}]$$

where λ_S is the intensity of the point process. λ_S can be estimated by 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[3]

$$\widehat{K}(s) = \frac{1}{\lambda_S} \sum_{i=1}^{n} \frac{1}{n-1} \sum_{i\neq j} I(d(i,j) \leq s) \tag{1}$$

$$= \frac{A}{n(n-1)} \sum_{i} \sum_{i\neq j} I(s_{ij} \leq s) \tag{2}$$

Ripley's K function has natural extensions to the purely temporal and space-time cases, with similar estimators to the above:

$$K(t) = \frac{1}{\lambda_T} E[\text{\# of events occurring within a time } t \text{ of an arbitrary event}]$$

$$K(s,t) = \frac{1}{\lambda} E[\text{\# of events occurring within a distance } s \text{ and a time } t \text{ of an arbitrary event}]$$

We remark that $\lambda K(s_0, t_0)$ is the entry in the upper-left hand corner of the contingency table used in Knox's test, and $K(s_0)$ and $K(t_0)$ are proportional to the top row and left column, respectively.

Diggle defines 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 defines a test statistic calculated over a grid of pre-specified spatial distances $s_1, \ldots, s_k$ and time separations $t_1, \ldots, t_l$:

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

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.

Notice the commonalities among the tests: each is really asking whether 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. Diggle'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. Finally, the Mantel test explicitly uses correlation to test whether the transformed interpoint space and interpoint time distributions are independent. By Bayes' rule, this is equivalent to testing (as per Diggle) whether a joint distribution factors into the product of marginals. If a random variable X is independent of a random variable Y then $P(X|Y) = P(Y)$ so by Bayes' rule: $P(X,Y) = P(X|Y)P(Y) = P(X)P(Y)$. (The implications work in the other direction as well.)

[3]This estimator assumes a constant first order intensity and ignores the issue of edge corrections. With n small or s large compared to A, edge corrections must be made, otherwise this estimate will be biased. See [25].

2.2 Extending the Classical Tests to Bivariate Space-Time Interaction

The Mantel, Knox, and Diggle tests can be readily extended to the bivariate case, where we are given $\mathcal{P}^1 = \{(s_i^1, t_i^1), i = 1, \ldots, n_1\}$ and $\mathcal{P}^2 = \{(s_i^2, t_i^2), i = 1, \ldots, n_2\}$, and 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 $\mathcal{P}^2$.

Mantel, Knox, and Diggle each focus on pairs of points. For the bivariate extension for each, we simply consider all $n_1 \cdot n_2$ cross-pairs[4] of points. For the Knox test we create the same contingency table where each entry counts the number of cross-pairs near in time, near in space, etc. For randomization testing, we 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. Randomization testing is by permuting the columns or rows of the time matrix.

Diggle's test in the bivariate case was explored in [20]. The cross-K-functions can be readily defined, i.e.: $K_{12}(s) = \frac{1}{\lambda s} E[\#\text{ of events of type 2 occurring within a distance } s \text{ of an event of type 1}]$, and estimators are available by considering the $n_1 \cdot n_2$ cross-pairs of points.

2.3 Background: MMD and HSIC

We start with a presentation of the Maximum Mean Discrepancy (MMD) test [14], which was proposed to solve the "two-sample problem" of testing whether two probability distributions p and q are equal. Let $x_1, \ldots, x_n \sim p$ be samples from p, and let $y_1, \ldots, y_n \sim q$ be a sample of equal size[5], with $x_i, y_i \in \mathcal{R}^d$.

There are already a variety of ways to test whether the probability distributions $P(p)$ and $P(q)$ are equal, including the Kolmogorov-Smirnoff (KS) statistic, which compares the empirical cumulative distribution functions (CDFs) of the two distributions, and Parzen window-based estimates [2]. KS is not easily extended beyond univariate distributions, and while Parzen window-based estimates are well-defined for multivariate dimensions, Parzen windows (i.e. density estimation) suffer from the curse of dimensionality.

On the other hand, MMD does not require density estimation, so it can be used in highly multivariate settings. Another important feature of MMD is that it is not limited to the space $\mathcal{R}^d$: since it is based on Mercer kernels, it could be applied to images, text, and even structured objects like graphical models—any setting where a Mercer kernel can be used to compare objects of interest.

The setup of MMD is as follows: we search for a function $f : \mathcal{R}^d \to \mathcal{R}$ which maximizes the mean discrepancy between the two functions evaluated on the two distributions:

$$\text{MMD} = \sup_f \left(E_{x \sim p}[f(x)] - E_{y \sim q}[f(y)] \right) \tag{3}$$

If $p = q$ then any choice of f will give MMD = 0. We would like this claim to be if and only if, and we have yet to say how we can arrive at a sample estimate of MMD. First, we need to constrain the function class f. Let f be a bounded continuous function in $\mathcal{R}^d$, which we denote as $f \in C(\mathcal{R}^d)$. By Lemma 9.3.2 of [11], MMD for bounded continuous functions is 0 if and only if $p = q$, that is, MMD is injective. Now, we turn to the question of estimating this quantity from a sample. An obvious estimator is:

$$\widehat{\text{MMD}} = \sup_f \left(\frac{1}{n} \sum_i f(x_i) - \frac{1}{n} \sum_j f(y_j) \right) \tag{4}$$

It is still not clear how we can find the function f that maximizes this statistic. For this, we need to turn to Mercer kernels and the theory of reproducing kernel Hilbert spaces (RKHS). For details, see [27]. Assume we have an

[4] In the spatial statistics literature, "cross" refers to functions calculated on pairs of points where the two points in the pair are of different types.

[5] The equal size requirement is for ease of presentation only.

RKHS $\mathcal{H}$, where $\mathcal{H}$ is the space of functions $f : \mathcal{R}^d \to \mathcal{R}$. This space needs an inner product, which is defined by a Mercer kernel k, which is is a positive semidefinite bilinear function over pairs of elements in $\mathcal{R}^d$. The range of k is $\mathcal{R}$. Intuitively, we can think of $k(x, y)$ as telling us how similar x is to y. For example, the commonly used Gaussian Radial Basis Function (RBF) kernel is given by $k(x, y) = e^{-.5\sigma^2\|x-y\|_2^2}$[6].

Kernels also provide us with a feature-space representation, which takes an element x and turns it into an infinite dimensional vector (equivalently, a function) which can be evaluated at any point y. In the Gaussian RBF case, this is notated $\phi(x) = k(x, \cdot)$, where $\phi(x)$ lives in $\mathcal{H}$, and it is a function from $\mathcal{R}^d$ to $\mathcal{R}$. Another way to think about this is to to take the RBF case and write $\phi(x)(y) = e^{-\sigma^2\|x-y\|_2^2}$ where x is fixed and y is the argument to the function $\phi(x)(\cdot)$. For x fixed, the function $\phi(x)$ is like an unnormalized Gaussian density centered at x: to find out how similar x is to y, we look at the distance between x and y, and then transform this distance with a Gaussian distribution. One important feature of RKHS's is the following: given two elements x and y, take their feature representation $\phi(x)$ and $\phi(y)$. Inner products in the infinite dimensional space $\mathcal{H}$, denoted $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$, can be evaluated with the so-called "kernel trick" as $k(x, y)$ without explicitly calculating the feature representation.

Finally, given a function f and an element x, the Riesz representation theorem states that we can evaluate $f(x)$ using the feature mapping: $f(x) = \langle \phi(x), f \rangle_{\mathcal{H}}$. Now, we extend this to the case of expectations over elements $x \sim p$, defining a "mean-embedding" element $\mu_p \in \mathcal{H}$ as follows:

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

We need to state under what conditions we can move the expectation inside the inner product; under these conditions, μ_p exists. We state the following lemma, which is Lemma 3 in [14].

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

We now return to MMD, restricting the functions f to be functions in an RKHS $\mathcal{H}$ for which $\|f\|_{\mathcal{H}} \leq 1$. This function class is still quite rich: we want MMD to be injective, and as long as we choose a kernel which gives functions that are dense in $C(\mathcal{R}^d)$ (so-called "universal" kernels), MMD will be injective, i.e. $p = q$ if and only if MMD = 0. The most attractive feature of this function class is that we do not need to do any explicit maximization, as follows:

$$\begin{aligned}
\text{MMD} &= \sup_{\|f\|_{\mathcal{H}} \leq 1} E_x f(x) - E_y f(y) && (5)\\
&= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mu_p f - \mu_q f && (6)\\
&= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle \mu_p - \mu_q, f \rangle_{\mathcal{H}} && (7)\\
&= \langle \mu_p - \mu_q, \frac{\mu_p - \mu_q}{\|\mu_p - \mu_q\|_{\mathcal{H}}} \rangle_{\mathcal{H}} && (8)\\
&= \|\mu_p - \mu_q\|_{\mathcal{H}} && (9)
\end{aligned}$$

Equation 9 is true for Hilbert spaces. For intuition as to why, think about a Euclidean space with the L_2 norm: if we have a vector v and we wish to find a unit vector w that maximizes $\langle v, w \rangle$, we choose w to point in the same direction as v and have unit length: $w^* = \frac{v}{\|v\|_2}$. The function f^* which maximizes MMD is called the "witness" function:

$$f^* = \frac{\mu_p - \mu_q}{\|\mu_p - \mu_q\|_{\mathcal{H}}}$$

With the above expression for MMD, it is now possible to state an estimator which does not include an optimization. Let $x, x' \sim p$ and $y, y' \sim q$. For a particular choice of kernel k, we have the following (where we have switched to MMD^2 for convenience):

$$\begin{aligned}
\text{MMD}^2 &= \|\mu_p - \mu_q\|_{\mathcal{H}}^2 && (10)\\
&= \langle \mu_p - \mu_q, \mu_p - \mu_q \rangle && (11)\\
&= \langle \mu_p, \mu_p \rangle - 2\langle \mu_p, \mu_q \rangle + \langle \mu_q, \mu_q \rangle && (12)\\
&= E_{x,x'} k(x, x') - 2E_{x,y} k(x, y) + E_{y,y'} k(y, y') && (13)
\end{aligned}$$

[6]Later, we refer to the bandwidth of this kernel, which is equal to $1/\sigma$

This suggests the following estimator:

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) - \frac{2}{n^2} \sum_{i,j} k(x_i, y_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) \tag{14}$$

Now we present the closely related Hilbert Schmidt Independence Criterion (HSIC) [13]. Although it was not originally presented as such, HSIC can be thought of as an extension of MMD: if we have a joint distribution $p \sim P(X, Y)$ for variables X and Y and the product of two marginal distributions $q \sim P(X)P(Y)$ and we wish to use MMD to test whether $p = q$, which is equivalent to asking whether $P(X, Y) = P(X)P(Y)$, i.e. whether $X \perp\!\!\!\perp Y$, where $\perp\!\!\!\perp$ means independent. Two variables X and Y are statistically independent if knowledge of the value of one does not give any information about the value of the other: $P(X|Y) = P(X)$. By Bayes' Theorem this is true if and only if $P(X, Y) = P(X)P(Y)$.

MMD can be readily applied to this new setting: we need to define a suitable kernel over pairs (x, y). A simple choice is to take a kernel $k(x, x')$ and another kernel $l(y, y')$ and construct the cross-product kernel: $v((x, y), (\cdot, \cdot)) = k(x, \cdot)l(y, \cdot)$. In terms of Hilbert spaces this is the cross-product Hilbert space: $\mathcal{V} = \mathcal{H}_k \times \mathcal{H}_l$. We use the notation μ_{XY} to denote the mean embedding of the joint distribution and $\mu_X \cdot \mu_Y$ as the product of the mean embeddings for the marginal distributions. (This decomposition does not hold in general, but for the kernel v we constructed it does.) Now we consider MMD^2, which we call HSIC:

$$\mathrm{HSIC} = \|\mu_{XY} - \mu_X \cdot \mu_Y\|^2_{\mathcal{H}} \tag{15}$$

$$= E_{x,y}E_{x',y'}k(x, x')l(y, y') - 2E_{x,y}E_{x'}E_{y'}k(x, x')l(y, y') + E_xE_yE_{x'}E_{y'}k(x, x')l(y, y') \tag{16}$$

A (biased) estimator is as follows:

$$\widehat{\mathrm{HSIC}} = \frac{1}{n^2} \sum_{i,j} k(x_i, x_j)l(y_i, y_j) - \frac{2}{n^3} \sum_{i,j,q} k(x_i, x_j)l(y_i, y_q) + \frac{1}{n^4} \sum_{i,j,q,r} k(x_i, x_j)l(y_q, y_r) \tag{17}$$

(The reason this has the name Hilbert-Schmidt Independence Criterion is explained in [13], where it is derived by considering the cross-covariance operator between X and Y in a Hilbert space. If this operator is 0 then X and Y are independent, and the Hilbert-Schmidt norm gives a test statistic for whether the cross-covariance operator is 0, which is equivalent to the expressions above.)

A particularly simple expression for this estimator is available. Let K be the Gram matrix with entries $k_{ij} = k(x_i, x_j)$ and L be the Gram matrix with entries $l_{ij} = l(y_i, y_j)$. Let $H = I - \frac{1}{n}11^T$ where 1 is an $nx1$ vector of ones. Then the following is equivalent to Equation 17: $\frac{1}{n^2}\mathrm{tr}HKHL$.

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 $\widehat{\mathrm{HSIC}}$. An asymptotic test based on the Gamma distribution is stated in [13], and another test based on the eigenvalues of the kernel matrices is derived in [30].

2.4 HSIC for Space-Time Interaction

As an intermediate step towards using HSIC for testing for space-time interaction, and as an interesting new test in its own right, applicable beyond the space-time case, we define a kernelized version of the Mantel test. The Mantel test was described in Section 2.1. We briefly restate it in a more general form, following [19]. The Mantel test compares a pair of distance matrices. Given a set of objects P, and two different ways of measuring the distances 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 distance between objects i and j, the Mantel test statistic is $\sum_{i \neq j} k(i, j)l(i, j)$, and its significance is usually tested by randomization. Interestingly, this is the first term in the estimator for HSIC. 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 could proceed exactly as with the Mantel test, defining the test statistic $T = \sum_{i \neq j} k(i, j)l(i, j)$, and obtaining significance levels by randomization testing where each time, a single permutation

is applied to both the rows and columns of L (or K, but the choice should be made once). We call this test the "kernelized Mantel test" and to our knowledge it has not been previously considered in the literature.

With this as background, we are ready to define a new test for space-time interaction based on HSIC.[7] Recall that HSIC tests whether a joint probability distribution $P(X, Y)$ is equal to the product of the marginal distributions $P(X)P(Y)$. Given a set of objects P, and kernels k and l, these kernels define feature representations for these objects in some feature space, i.e. for an object p the feature representation $\phi_k(p) := k(p, \cdot)$ and $\phi_l(p) := l(p, \cdot)$. For intuition, we return to the spatio-temporal case. Let the objects be points p and p' given by (s, t) and (s', t'). k only considers spatial coordinates: $k(s, s')$ while $l(t, t')$ only considers temporal coordinates, so $\phi_k(p) = k(s, \cdot)$ and $\phi_l(p) = l(t, \cdot)$. The null hypothesis is that k and l are independent. We can use HSIC to test this hypothesis, choosing a translation invariant kernel (like Gaussian RBF) to stay close to the classical tests. We can think of there being a latent probability measure over objects P. If we draw p and p' from P the kernels tell us how to embed them into feature space, and we can think of $\phi_k(p)$ and $\phi_l(p)$ as random variables. Now we use HSIC to ask whether the joint distribution over embeddings $P(\phi_k(p), \phi_l(p))$ factors into the product of the marginals $P(\phi_k(p))P(\phi_l(p))$.

Recall that originally in HSIC we had samples $(x, y) \sim \mathcal{X} \times \mathcal{Y}$, kernels k and l, feature embeddings $\phi_k(x) = k(x, \cdot)$ and $\phi_l(y) = l(y, \cdot)$, and mean embeddings $\mu_{\mathcal{XY}}$ and $\mu_{\mathcal{X}} \cdot \mu_{\mathcal{Y}}$. For draws $p \sim P$, we still have kernels k and l, feature embeddings, and mean embeddings. In some sense, we have skipped a step in HSIC, because we are already working in feature space, and it is only in feature space that we have a joint distribution and product of marginal distributions which we are interested in maximizing the difference between. (In the original space, we only had a distribution P from which we drew objects p.) We can write the criterion we wish to maximize as:

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \langle \mu_{\mathcal{XY}} - \mu_{\mathcal{X}}\mu_{\mathcal{Y}}, f \rangle_{\mathcal{H}}$$

This is exactly the same as HSIC, so we already derived the answer, $\|\mu_{\mathcal{XY}} - \mu_{\mathcal{X}}\mu_{\mathcal{Y}}\|^2$, with the same estimators as stated in Section 2.3.

As in the case of MMD, we can write out the witness function:

$$f^* = \frac{\mu_{\mathcal{XY}} - \mu_{\mathcal{X}}\mu_{\mathcal{Y}}}{\|\mu_{\mathcal{XY}} - \mu_{\mathcal{X}}\mu_{\mathcal{Y}}\|_{\mathcal{H}}}$$

This witness function can be evaluated at any point p in the original space. The estimator is:

$$\hat{f}^*(p) \propto \frac{1}{n-1} \sum_{i:p_i \neq p} k(p_i, p)l(p_i, p) - \frac{1}{(n-1)^2} \sum_{i:p_i \neq p} k(p_i, p) \sum_{j:p_j \neq p} l(p_j, p)$$

If we calculate $\hat{f}^*(p_i)$ for each point i (we ignore the proportionality constant—we can always rescale later) then we have an $n-$dimensional witness vector $\vec{w}$. If an entry w_i is large in magnitude, than the interpretation is that point p_i provides evidence against the null hypothesis, i.e. p_i is a point for which the two distance measures, k and l are not independent. [8]

While we have been highlighting commonalities between the Mantel test and HSIC, the test statistics do not have precisely the same form. The kernelized Mantel test statistic is:

$$\sum_{i,j} k(x_i, x_j)l(x_i, x_j) \tag{18}$$

[7] In the spirit of the classical tests described in Section 2.1, one approach to using HSIC would be to define new distributions $P = \{d_s(i, j) : i \neq j\}$ for the Euclidean distance between pairs of points and $Q = \{d_t(i, j) : i \neq j\}$ for the interpoint time interval, 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.

[8] Returning to the original Mantel test, we can use the original distance matrices K and L to calculate an analogous witness vector $\vec{w}$ for the Mantel test, where

$$w_q = \frac{1}{n-1} \sum_{i \neq q} k_{iq}l_{iq} - \frac{1}{(n-1)^2} \sum_{i \neq q} k_{iq} \sum_{j \neq q} l_{jq}$$

Notice that the original Mantel test statistic is proportional to the average of the first term of w_q over all q.

while for HSIC we have:

$$\frac{1}{n^2}\sum_{i,j} k(x_i, x_j)l(y_i, y_j) - \frac{2}{n^3}\sum_{i,j,r} k(x_i, x_j)l(y_i, y_r) + \frac{1}{n^4}\sum_{i,j,q,r} k(x_i, x_j)l(y_q, y_r) \tag{19}$$

The first term of Equation 19 is proportional to Equation 18, and the last term of Equation 19 remains unchanged under permutation, so leaving it out will not affect significance levels derived from HSIC. The key difference is the second term of Equation 19, which is missing from the Mantel test. This term is the cross-term from the maximization, so it is certainly necessary for the derivation of HSIC. Note that because it is asymmetric, it *does* change under permutation testing. It is not obvious what effect its omission will have on the kernelized Mantel test. We explore this experimentally below. It is possible that some of the empirical criticisms [18] of the Mantel test's power could be addressed by including this missing cross term, which would be an interesting line of future research. [18] recommends restricting the use of the Mantel test to hypotheses that can be "formulated only in terms of distances." It would also be interesting to investigate the use of a kernelized Mantel test for these cases.

Calculating HSIC and kernelized Mantel both take time $O(n^2)$. For HSIC, randomization or an asymptotic method is used to assess significance. [15] gives methods for optimally choosing the kernel for MMD which could be adapted to this case.

2.5 HSIC for Bivariate Space/Time Interactions

Extending HSIC to the bivariate case, which we call HSIC_{12}, follows by analogy to the extension of the classical tests to the bivariate case. Of particular interest is the interpretation of HSIC_{12} in terms of Hilbert spaces and kernels. In the classical cases, which only focused on spatial and temporal distances, we simply restricted our sample to cross-distances between two types of points. One approach is by analogy to the kernelized Mantel test. Recall that for the classical Mantel test, we used matrices of size $n_1 \times n_2$ with entries for each cross-pair of points. The kernelized Mantel test can be extended in the same way, with the test statistic: $\sum_{i,j} k(i, j)l(i, j)$.

Similarly, we propose an extension of the HSIC approach above, which we call HSIC_{12}. Consider point processes $P^1 = \{(s_i^1, t_i^1), i = 1, \ldots, n_1\}$ and $P^2 = \{(s_j^2, t_j^2), j = 1, \ldots, n_2\}$. In what follows, we stipulate that all variables s and t are of type 1 and all variables s' and t' are of type 2, and we consider the following test statistic:

$$\text{HSIC}_{12} = E_{st}E_{s't'}[k(s, s')l(t, t')] - E_{st}[E_{s'}k(s, s')E_{t'}l(t, t')] - E_{s't'}[E_s k(s, s')E_t l(t, t')] + E_s E_{s'}k(s, s')E_t E_{t'}l(t, t')$$

An estimator is:

$$\widehat{\text{HSIC}}_{12} = \frac{1}{n^2}\sum_{i,j} k(s_i^1, s_j^2)l(t_i^1, t_j^2) - \frac{1}{n^3}\sum_{i,j,r} k(s_i^1, s_j^2)l(t_i^1, t_r^2) - \frac{1}{n^3}\sum_{i,j,q} k(s_i^1, s_j^2)l(t_q^1, t_j^2) + \frac{1}{n^4}\sum_{i,j,q,r} k(s_i^1, s_j^2)l(t_q^1, t_r^2)$$

There are a few different ways to interpret and justify this proposed statistic. The first is to compare it to the kernelized (cross) Mantel test, and notice that just as HSIC adds the middle term to the Mantel test statistic, this new "pair HSIC" adds similar middle terms to the kernelized cross Mantel test (the last term remains invariant under permutation).

Another way to understand this statistic is by writing it using mean embeddings as follows:

$$\langle \mu_{XY} - \mu_X\mu_Y, \mu_{X'Y'} - \mu_{X'}\mu_{Y'} \rangle$$

In the standard MMD case, we would have arrived at a symmetric inner product, i.e. we would be measuring the squared length of the vector $\mu_{XY} - \mu_X\mu_Y$ in the Hilbert space. In this case, we are looking at two vectors (embeddings in feature space), the first is the discrepancy between the joint and product of the marginals for $\mathcal{P}^1$, i.e. it is an infinite dimensional vector (a function) that characterizes the space-time interaction throughout the support of the distribution. The second is the same, but for $\mathcal{P}^2$. The inner product between these two vectors in the Hilbert space measures the similarity between the space-time interaction present in each of the separate processes. If the statistic is zero, it means they have independent space-time interactions. If the statistic is far from zero, it means the two vectors are close together in the feature space, i.e. the space-time interaction of each is dependent.

3 Experimental Evaluation

3.1 Synthetic Data

In [9], the authors conducted a small power analysis of their new method, using simulated data. The setup is as follows: parent locations are sampled on the unit cube. The number of children for each parent is drawn $\sim$ Poisson(μ) where $\mu = 1$ (Figure 3) or 5 (Figure 4). 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. [9] allowed the cutoffs for the points at which the K function was evaluated to vary, revealing the importance of correctly pre-selecting these values: for a 25×25 grid extending from $[0, .25]$ in each axis, the test had the highest power. We reproduced these simulations, drawing 100 point processes for each value of σ and using [26] to obtain p-values. The power is shown as the fraction of simulations which rejected at $\alpha = .05$. For comparison, we ran HSIC with a Gaussian RBF kernel with bandwidth .01 for time and .02 for space. The power was comparable to the best set of thresholds for Diggle's test. In Figure 3, with only (on average) one child per parent, the power of each test falls off quickly as σ increases, while this trend is slower in Figure 4 where there are 5 children (on average) per parent. Throughout, HSIC outperforms Diggle's test.

In the spirit of these simulations, we made a simple model to capture the bivariate case of two types of point processes: after drawing random cluster centers in space and time, where each coordinate is $\sim$ Uniform$(0, 1)$, we drew points displaced as in the previous setup, by a separate draw from Normal$(0, \sigma)$ for a range of σ's. We then randomly assigned each of these points to be one of two types, and threw away the original cluster centers. We compared the kernelized Mantel test, Knox test, and HSIC. See Figure 5. HSIC does slightly better than both kernelized Mantel and Knox. This provides some evidence suggesting that the cross-term in Equation 19 is a meaningful difference between HSIC and kernelized Mantel.

Finally, we used the log-Gaussian Cox Process model [23] with a separable covariance function. We randomly assigned each point to be one of two types. This simulation was meant to capture some of the real situations we focus on below: there is both spatial and temporal dependence in these datasets, but it is separable, so a test for space-time interaction should only reject a fraction α of the time for significance level α. With the significance level set to $\alpha = .05$, we assessed whether the kernelized Knox test, HSIC_{12}, and the Mantel test correctly rejected the null 5% of the time. Each was close to the correct value, the the distribution of the p-values looks uniform, as it should under the null hypothesis. See Figure 6.

3.2 Crime Data from the City of Chicago

Based on conversations with city officials, we focused on a very specific and relevant 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 respond more pro-actively to these leading indicators so as to prevent homicides. We used a dataset provided by the Chicago Police Department of geocoded, date and timestamped calls to 911 and another dataset of crime incident reports. Data 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 271 different types of calls to 911.

We considered each separate type of 911 call as one point process and homicides / shootings as the other point process, and we used HSIC_{12} and the bivariate Knox test (both in the forward-in-time setting) to look for significant space-time interactions. We used randomization testing with 1000 repetitions, where each time we shuffled the time labels of the types. Call types which were significant for $p \leq 0.01$ for either test are in Table 1. We also used Chicago's publicly available geocoded, date and timestamped crime incident reports for 2012. Considering each type of crime as a single point process, I used HSIC and Knox to test for space-time interaction. Results are in Table 2.

4 Discussion

As a first step towards evaluating the results in Table 1, we observe that the method finds obvious call types such as "shots fired" and "person with a gun" have significant space-time interaction with homicides and shootings. We

clarified some of the more obscure call type descriptions with our collaborator at the Chicago Police Department, who explained that, for example, "mission" is generated by police officers when they tell the dispatcher they are staying within a specific area for some reason, and are thus not available for other assignments. "Plan 1-5" and "10-1" are codes for major events. While these are not citizen calls to 911, they are further confirmations that the method is finding reasonable correlations.

Next, we investigated rows in Table 1 where HSIC found a significant interaction but the Knox test did not. Following [9], we plotted the residual space-time interaction for a range of spatial and temporal separations in Figures 7 and 8. The formula is:

$$D(s,t) = \frac{F_{S,T}(s,t) - F_S(s)F_T(t)}{F_S(s)F_T(t)}$$

Figure 7 shows the space-time interaction between calls to 911 about hearing "shots fired" and shootings, revealing the spatial scale over which this interaction operates: 0 to 200 feet and 0 to 2 days. Both the Knox test and HSIC reject the null with $p \leq 0.001$. Knox does not perform as well in Figure 8 where "person with a gun" is considered as a predictor of shootings: HSIC has $p \leq 0.001$ but Knox has $p \leq 0.045$ (which we consider as not significant). The interesting structure evident in Figure 8 suggests that calls about "person with a gun" predict shootings up to 200 feet away almost simultaneously *or* up to two days away, in the same location, but not both. This example suggests that perhaps the linear / discretization approach of Knox is unable to detect this kind of complicated space-time interaction.

The results in Table 2 are very interesting. Testing with HSIC showed that with the exception of gambling, interference with public officer, and kidnapping, each type of crime displayed significant space-time clustering. According to Knox, battery, criminal sexual assault, and sex offense were not significant, probably because of a misspecification of the critical space and time values. This could be further investigated with plots as in Figures 7 and 8. From a criminological point of view, the fact that there is space-time clustering for each of these crime types is perhaps not terribly surprising, and it matches the results found in [16] where robbery, burglary, and assault were each judged to exhibit space-time interaction. Of the crime types that did not exhibit space-time interaction, we can speculate that kidnapping is a singular, rare type of crime, while "interference with public officer" is a catch-all encompassing a variety of different events.

4.1 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 Diggle et al. in simulated and real data. The performance was comparable or better than the classical tests. We applied these tests to a predictive policing application, searching for leading indicators for shootings and homicides among 911 call types. From the list of 271 different 911 call types, we reported 60 which were significant for p-value $\leq .001$. The next step is to give this list to city officials for expert evaluation, refine the kernel and bandwidth choices, and investigate alternative specifications of the distance—should Euclidean distance be replaced by L_1 ("taxicab") distance? Future research could investigate predicting other types of crimes, or using crime incidents as predictors. A related question that these methods could answer is is whether the recent crime spike in Chicago can be explained through purely spatial and temporal trends, or whether there is a significant space-time interaction.

4.2 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.

911 call type	HSIC	Knox	911 call type (cont'd)	HSIC	Knox
10-1	0.002	0.0095	KIDNAPPING REPORT	0.925	0.01
ARSON REPORT	0.116	0.005	MENTAL UNAUTH ABSENCE	0.002	0.683
AUTO ACCIDENT PD	0.005	0.2285	MISSION	0.002	0.0105
AUTO THEFT IP	0.022	0.001	NOTIFY	0.056	0.001
BATTERY JO	0.006	0.701	OUTDOOR ROLL CALL	0.009	0.001
BATTERY VICTIM INJ.	0.001	0.267	PERSON SHOT	0.001	0.001
BEAT TEAM MEETING [OV]	0.002	0.3795	PERSON STABBED	0.522	0.007
CRIM DAM. TO PROP RPT	0.006	0.066	PERSON WITH A GUN	0.009	0.0445
CRIMINAL TRES. (OV)	0.024	0.0025	PICK UP CAR	0.245	0.001
DEATH REMOVAL	0.001	0.001	PLAN 1-5	0.001	0.001
EVIDENCE TECHNICIAN (PRI. 1)	0.001	0.001	POLLING PLACE CHECK	0.224	0.0085
EVIDENCE TECHNICIAN (PRI. 2)	0.001	0.001	ROBBERY VICTIM INJURED	0.004	0.6235
EVIDENCE TECHNICIAN (PRI. 3)	0.728	0.004	SHOTS FIRED	0.001	0.001
FOUND PROPERTY	0.074	0.003	SHOTS FIRED (OV)	0.001	0.001
GAMBLING	0.002	0.002	SUSPICIOUS PERSON (OV)	0.517	0.001
GANG DISTURBANCE	0.001	0.003	THEFT IP	0.001	0.5945
K9 REQUEST	0.69	0.002	TRANSPORT	0.364	0.0095

Table 1: Which 911 call types predict shootings / homicide? Using data from 2007 through the first half of 2010, we used the forward-in-time version of HSIC_{12} to test each pairing: 911 call type as points of type 1, shootings/homicides as points of type 2. We used a Gaussian RBF kernel with bandwidth 500 feet and 14 days. We also used the forward-in-time version of the Knox test, with critical values 500 feet and 14 days. In both cases, randomization testing with 1000 permutations was used. Results which were significant for $p < .01$ for either test are shown.

Crime type	Knox p-value	HSIC p-value
ARSON	0.01	0.01
ASSAULT	0.01	0.01
BATTERY	0.245	0.02
BURGLARY	0.01	0.01
CRIM SEXUAL ASSAULT	0.09	0.05
CRIMINAL TRESPASS	0.01	0.01
DECEPTIVE PRACTICE	0.01	0.01
GAMBLING	0.19	0.07
HOMICIDE	0.01	0.01
INTERFERENCE WITH PUBLIC OFFICER	0.095	0.06
KIDNAPPING	0.28	0.16
LIQUOR LAW VIOLATION	0.01	0.01
MOTOR VEHICLE THEFT	0.01	0.01
OFFENSE INVOLVING CHILDREN	0.01	0.01
PROSTITUTION	0.01	0.01
PUBLIC PEACE VIOLATION	0.01	0.01
ROBBERY	0.01	0.01
SEX OFFENSE	0.265	0.02
WEAPONS VIOLATION	0.01	0.01

Table 2: Space-time interactions for crime types (univariate) in Chicago in 2012. Using publicly available crime data, we used the Knox test and HSIC to test whether each individual type of crime could be considered "infectious." Bolded results in the Knox column highlight possible occasions in which Knox failed to reject the null hypothesis while HSIC rejected the null hypothesis (for $\alpha = .05$). The critical distance in space for Knox was 500 feet and the critical distance in time was 14 days. Kernel bandwidths of 500 feet and 14 days were used as well with the Gaussian RBF kernel.

References

[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. 1.1

[2] Niall H Anderson, Peter Hall, and DM Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54, 1994. 2.3

[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. 2.1

[4] Alfred Blumstein and Joel Wallman. *The crime drop in America*. Cambridge University Press, 2005. 1.1

[5] Anthony Allan Braga and David L Weisburd. *Policing problem places: Crime hot spots and effective prevention*. Oxford University Press, USA, 2010. 1.1

[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. 1.1

[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. 1.1

[8] Lin Cui. Foreclosure, vacancy and crime. *Available at SSRN 1773706*, 2010. 1

[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. 1, 2.1, 3.1, 4

[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. 1.1

[11] Richard M Dudley. *Real analysis and probability*, volume 74. Cambridge University Press, 2002. 2.3

[12] Ingrid Gould Ellen, Johanna Lacoe, and Claudia Ayanna Sharygin. Do foreclosures cause crime? *Journal of Urban Economics*, 2012. 1

[13] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander J Smola. A kernel statistical test of independence. 2008. 1, 2.3, 2.3

[14] 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. 2.3, 2.3

[15] Arthur Gretton, Bharath Sriperumbudur, Dino Sejdinovic, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, et al. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems 25*, pages 1214–1222, 2012. 2.4

[16] 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. 1.1, 4

[17] 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. 1, 2.1

[18] Pierre Legendre and Marie-Josée Fortin. Comparison of the mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources*, 10 (5):831–844, 2010. 2.4

[19] Pierre Legendre and Louis Legendre. *Numerical ecology*. Elsevier, 2012. 2.4

[20] 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. 1.1, 2.2

[21] Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220, 1967. 1, 2.1

[22] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493), 2011. 1.1

[23] J. Møller, A.R. Syversveen, and R.P. Waagepetersen. Log gaussian cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998. 3.1

[24] D.B. Neill and W.L. Gorr. Detecting and preventing emerging epidemics of crime. *Advances in Disease*, 2007. 1.1

[25] B.D. Ripley. Edge effects in spatial stochastic processes. *Statistics in theory and practice*, pages 247–262, 1982. 3

[26] Barry Rowlingson, Peter Diggle, adapted, packaged for R by Roger Bivand, pcp functions by Giovanni Petris, and goodness of fit by Stephen Eglen. *splancs: Spatial and Space-Time Point Pattern Analysis*, 2013. URL `http://CRAN.R-project.org/package=splancs`. R package version 2.01-32. 3.1

[27] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization and beyond*. the MIT Press, 2002. 2.3

[28] Matthew A Taddy. Autoregressive mixture models for dynamic spatial poisson processes: Application to tracking intensity of violent crime. *Journal of the American Statistical Association*, 105(492):1403–1417, 2010. 1.1

[29] James Q Wilson and George L Kelling. Broken windows. *Atlantic monthly*, 249(3):29–38, 1982. 1.1

[30] 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. 2.3

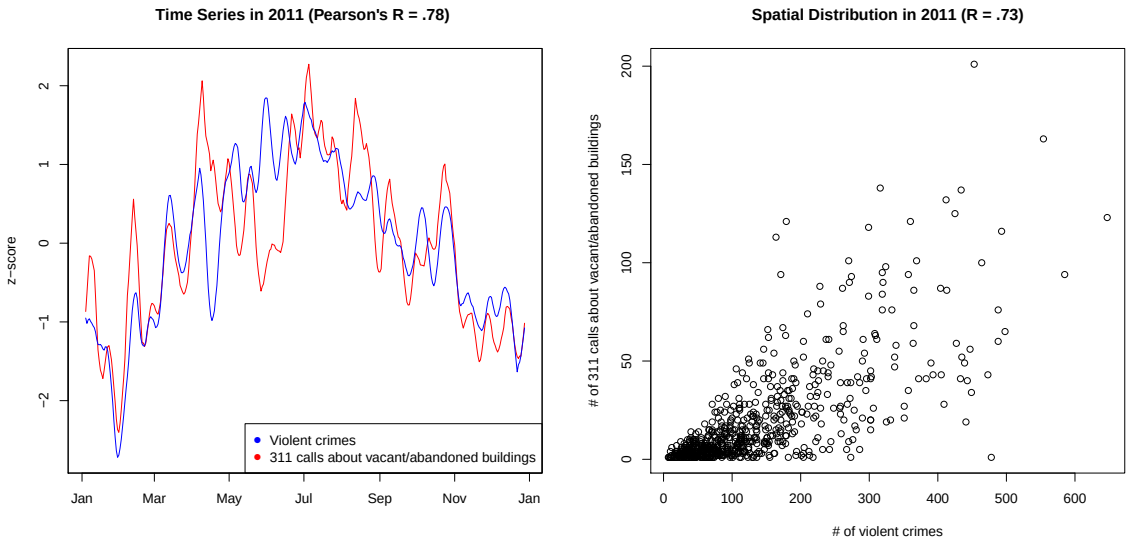
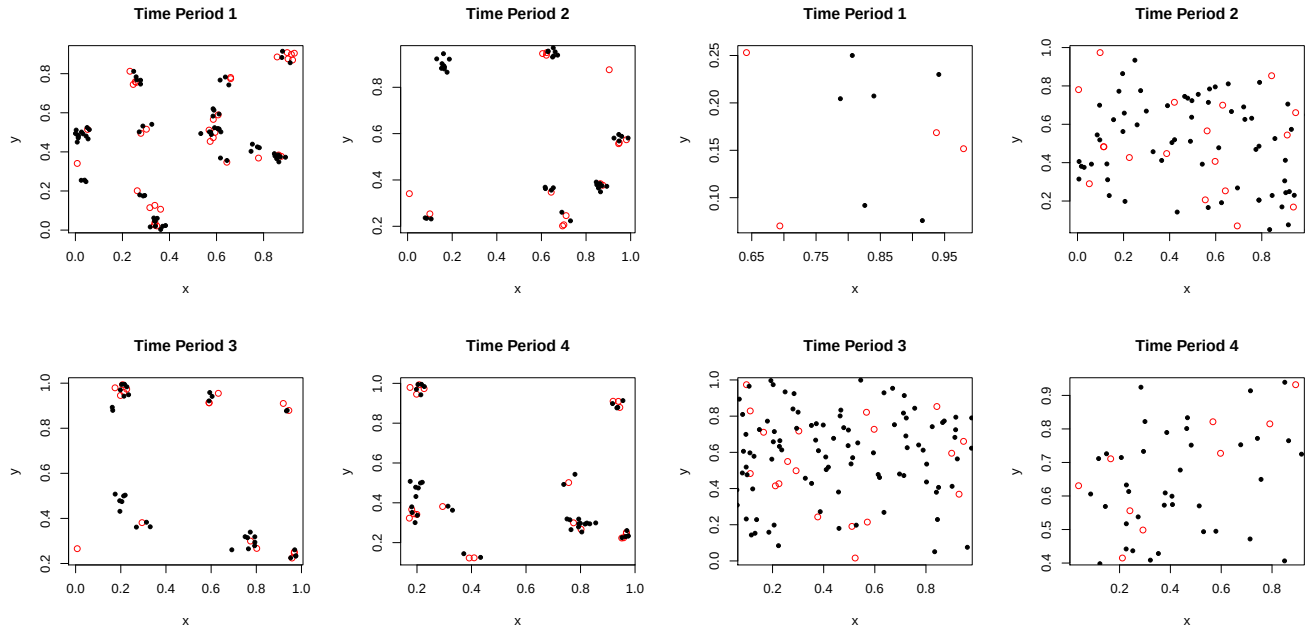


Figure 1: On the left, there is a strong correlation between the time series of violent crimes and calls to 311, Chicago's non-emergency services number, about vacant/abandoned buildings, due at least in part to common seasonal trends. On the right, calls about vacant/abandoned buildings and violent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



(a) (b)

Figure 2: Two different "infectious" Poisson cluster process with parents shown as red circles and children shown as black dots. Children are displaced from parents in space and (forward) in time by iid draws from $N(0, \sigma)$. In (a), $\sigma = .025$ and in (b) $\sigma = .2$. Visual inspection reveals space-time interaction in (a) but not easily in (b) while tests for space-time interaction reject the null hypothesis (of no space-time interaction) in both cases.

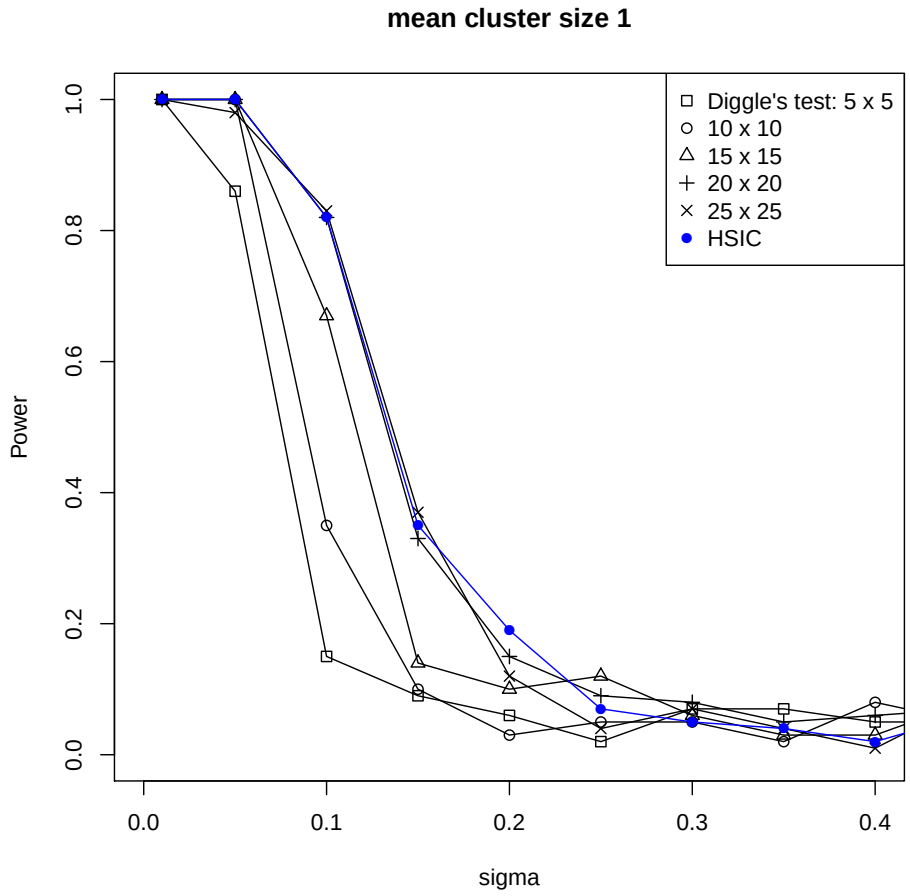


Figure 3: For each value of σ we created a simulated data set with 100 randomly located parent points and on average one child point per parent, displaced in space and time by independent draws $\sim N(0, \sigma)$. Space-time interaction was assessed with Diggle's test with grids of size $5 \times 5, \ldots, 25 \times 25$ (in each case, the distance between threshold values was .01) and HSIC. The y-axis shows the power, i.e. the fraction of times the test correctly rejected the null at $\alpha = .05$. HSIC, which outperforms most parameters settings of Diggle's test, was run with Gaussian RBF kernels with bandwidth .01 for the time kernel and .02 for the space kernel.

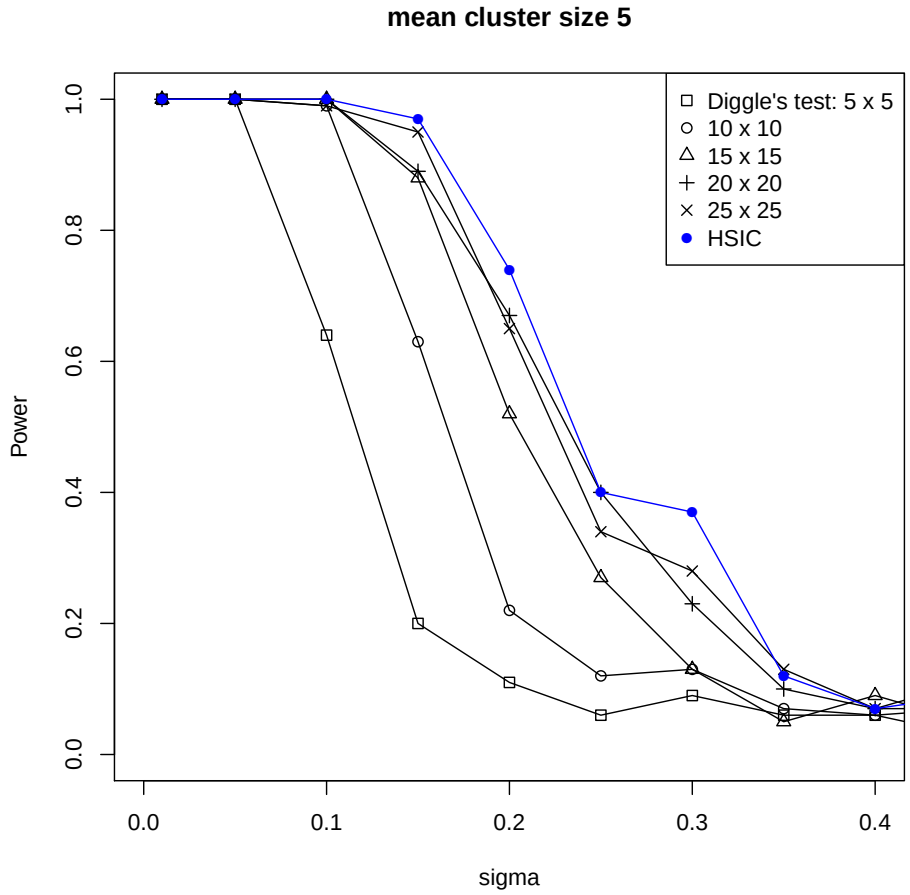


Figure 4: The setup is the same as in Figure 3 except that there are 40 parents and on average each parent has 5 children.

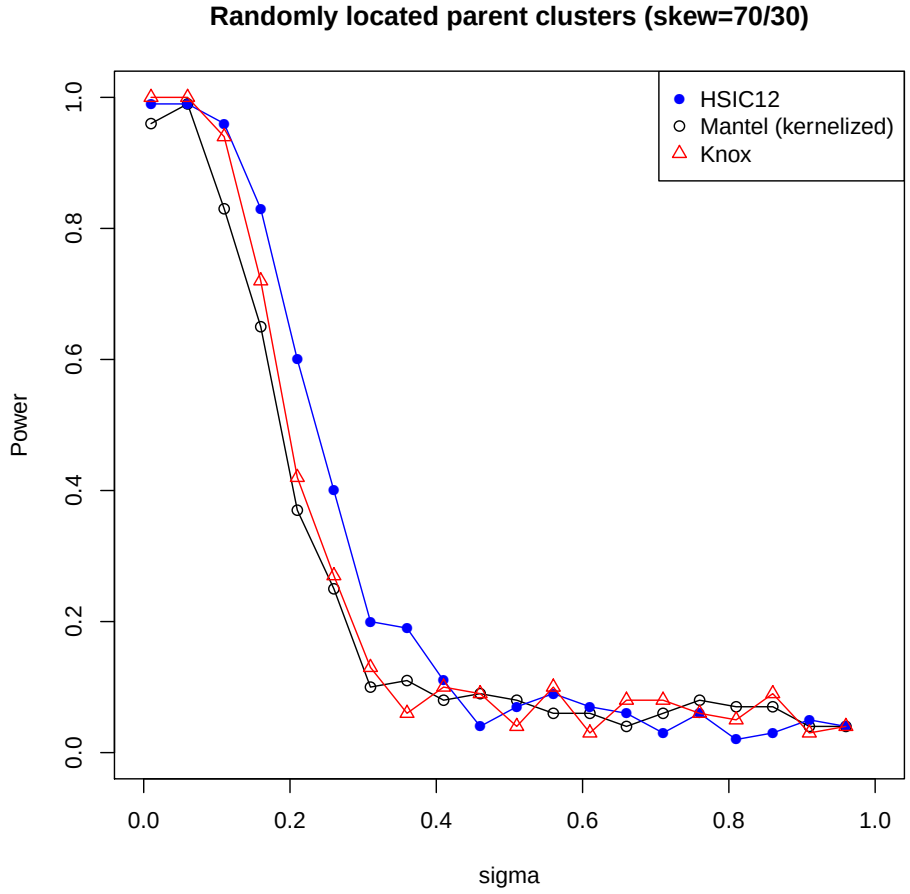


Figure 5: We drew 20 cluster centers and for each value of σ, children as in the setup in Figure 3. We randomly assigned each set of children to be of one of two types, with 70% of type 1 and 30% of type. We then threw away the cluster centers. For each value of σ we did this 100 times and obtained p-values for HSIC_{12}, kernelized Mantel, and Knox. We calculated the power as the fraction of times each test rejected at $\alpha = .05$.

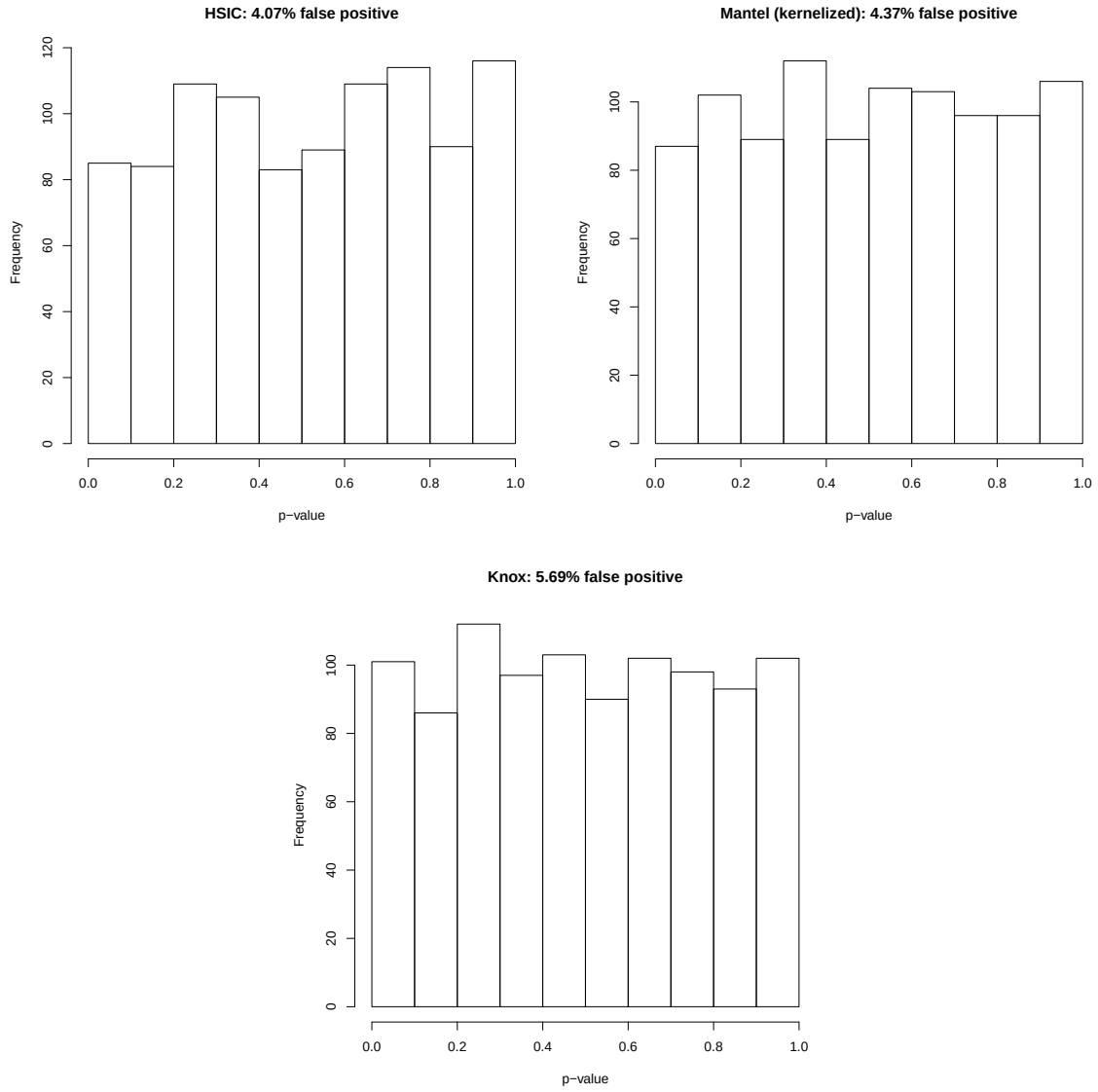


Figure 6: Type 1 error controls: 100 simulated datasets were generated from a log-Gaussian Cox Process model with a separable covariance function, meaning that space-time interaction tests should only reject $\alpha\%$ of the time for significance level α. Histograms of the p-values are shown, and they look relatively uniform, which is good. In the titles, the false positive rate (i.e. the percent of times that the test incorrectly rejected the null) is shown for $\alpha = 5\%$.

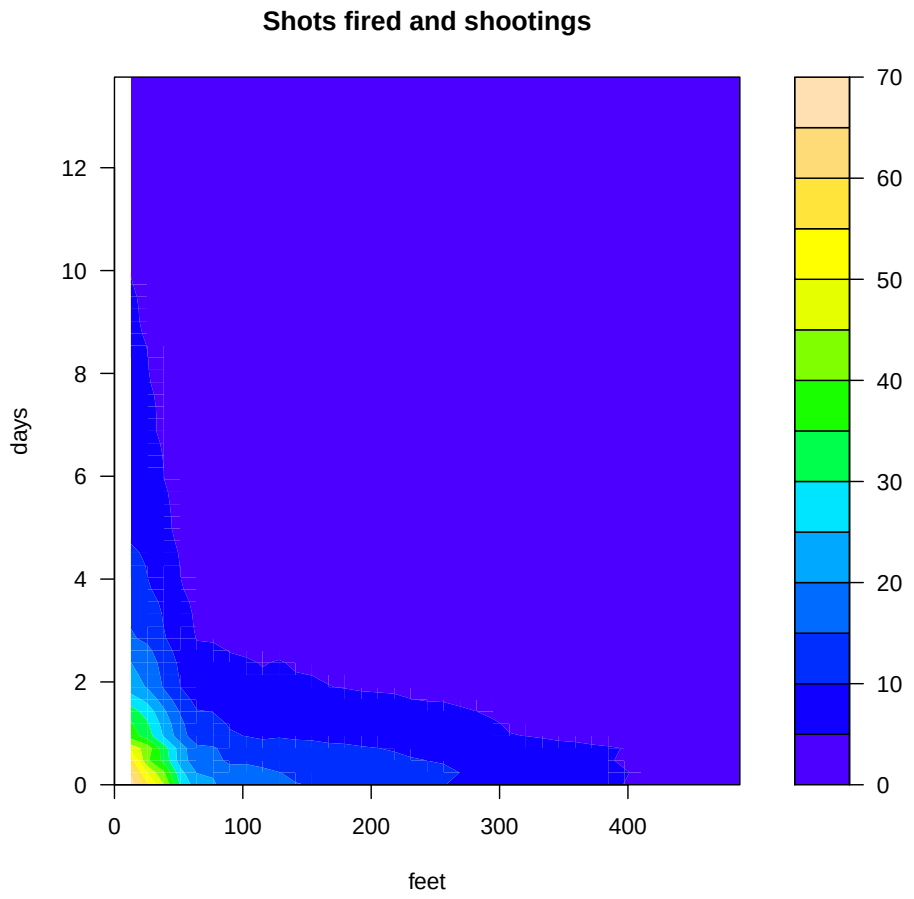


Figure 7: Space-time interaction plot between calls to 911 coded as type "shots fired" and shootings / homicides. Both the Knox test and HSIC rejected the null hypothesis of no space-time interaction with $p \leq 0.001$. The color-scale should be interpreted as 0 = no interaction to ≥ 1 = high interaction.

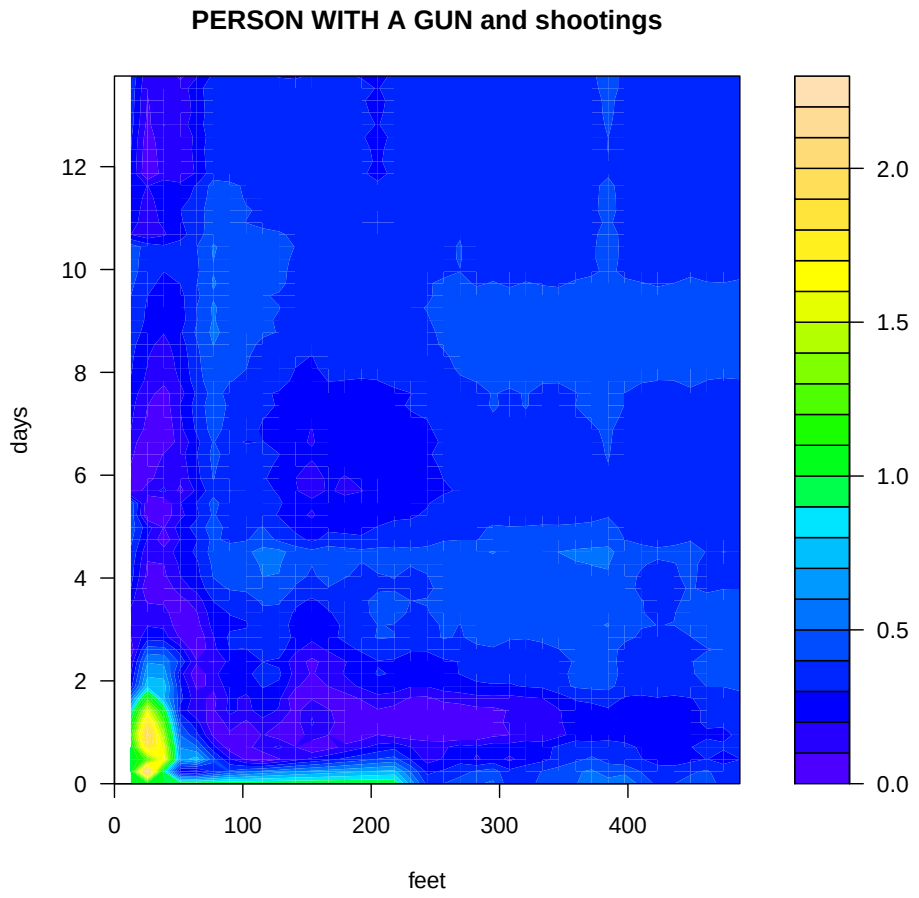


Figure 8: Space-time interaction plot between calls to 911 coded as type "person with a gun" and shootings / homicides. HSIC had p-value ≤ 0.001, while Knox had p-value ≤ 0.045 (which we interpreted as not significant).