A Kernel Independence Test for Geographical Language Variation

A Kernel Independence Test for Geographical Language Variation

Dong Nguyen
University of Twente

Jacob Eisenstein
Georgia Institute of Technology

Quantifying the degree of spatial dependence for linguistic variables is a key task for analyzing dialectal variation. However, existing approaches have important drawbacks. First, they are based on parametric models of dependence, which limits their power in cases where the underlying parametric assumptions are violated. Second, they are not applicable to all types of linguistic data: some approaches apply only to frequencies, others to boolean indicators of whether a linguistic variable is present. We present a new method for measuring geographical language variation, which solves both of these problems. Our approach builds on Reproducing Kernel Hilbert Space (RKHS) representations for nonparametric statistics, and takes the form of a test statistic that is computed from pairs of individual geotagged observations without aggregation into predefined geographical bins. We compare this test with prior work using synthetic data as well as a diverse set of real datasets: a corpus of Dutch tweets, a Dutch syntactic atlas, and a dataset of letters to the editor in North American newspapers. Our proposed test is shown to support robust inferences across a broad range of scenarios and types of data.[1]

1. Introduction

Figure 1 shows the geographical location of 1000 Twitter posts containing the word *hella*, an intensifier used in expressions like *I got hella studying to do* and *my eyes got hella big* (Eisenstein et al. 2014). While the word appears in major population centers throughout the United States, the map suggests that it enjoys a particularly high level of popularity on the west coast, in the area around San Francisco. But does this represent a real geographical difference in American English, or is it the result of chance fluctuation in a finite dataset?

Regional variation of language has been extensively studied in sociolinguistics and dialectology (Chambers and Trudgill 1998; Grieve, Speelman, and Geeraerts 2011, 2013; Lee and Kretzschmar Jr 1993; Nerbonne and Kretzschmar Jr 2013; Szmrecsanyi 2012). A common approach involves mapping the geographic distribution of a linguistic variable (e.g., the choice of *soda*, *pop*, or *coke* to refer to a soft drink) and identifying boundaries between regions based on the data. The identification of linguistic variables that exhibit regional variation is therefore the first step in many studies of regional dialects. Traditionally, this step has been based on the manual judgment of the researcher; depending on the quality of the researcher's intuitions, the most interesting or important variables might be missed.

The increasing amount of data available to study dialectal variation suggests a turn towards data-driven alternatives for variable selection. For example, researchers can mine social media data such as Twitter (Doyle 2014; Eisenstein et al. 2010; Huang

1 Code is available at `https://github.com/dongpng/geo-independence-testing`

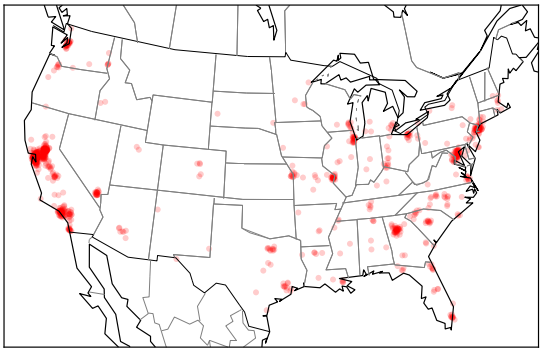


Figure 1: 1000 geolocated tweets containing the word *hella*

et al. 2016) or product reviews (Hovy, Johannsen, and Søgaard 2015) to identify and test thousands of dialectal variables. Despite the large scale of available data, the well-known "long tail" phenomenon of language ensures that there will be many potential variables with low counts. A statistical metric for comparing the strength of geographical associations across potential linguistic variables would allow linguists to determine whether finite geographical samples — such as the one shown in Figure 1 — reveal a statistically meaningful association.

The use of statistical methods to analyze spatial dependence has been only lightly studied in sociolinguistics and dialectology. Existing approaches employ classical statistics such as Moran's I (e.g., Grieve, Speelman, and Geeraerts (2011)), join count analysis (e.g., Lee and Kretzschmar Jr (1993)) and the Mantel Test (e.g., Scherrer (2012)); we review these statistics in section 2. These classical approaches suffer from a common problem: each type of test can capture only a specific parametric form of spatial linguistic variation. As a result, these tests can incorrectly fail to reject the null hypothesis if the nature of the geo-linguistic dependence does not match the underlying assumptions of the test.

To address these limitations, we propose a new test statistic that builds on a rich and growing literature on kernel embeddings for nonparametric statistics (Shawe-Taylor and Cristianini 2004). In these methods, probability distributions, such as the distribution over geographical locations for each linguistic variable, are embedded in a Reproducing Kernel Hilbert Space (RKHS). Specifically, we employ the Hilbert-Schmidt Independence Criterion (HSIC; Gretton et al. (2005a)). Due to its ability to compare arbitrarily high-order moments of probability distributions, HSIC can be used to compare arbitrary probability measures, by computing kernel functions on finite samples. Unlike prior approaches, HSIC is statistically consistent: in the limit of a sufficient amount of data, it will correctly determine whether the distribution of a linguistic feature is geographically dependent. As a further convenience, because it is built on kernel similarity functions, HSIC can be applied with equal ease to any type of linguistic data, as long as an appropriate kernel function can be constructed.

To validate this approach, we compare it against three alternative spatial statistics: Moran's I, the Mantel test, and join count analysis. For a controlled comparison, we use synthetic data to simulate different types of regional variation, and different types of linguistic data. This allows us to measure the capability of each approach to recover true geo-linguistic associations, and to avoid Type I errors even in noisy and sparse

data. Next, we apply these approaches to three real linguistic datasets: a corpus of Dutch tweets, a Dutch syntactic atlas and letters to the editor in North American newspapers.

To summarize, the contributions of this article are:

- We show how the Hilbert-Schmidt Independence Criterion can be applied to linguistic data. HSIC is a nonparametric test statistic, which can handle both frequency and categorical data. It requires no discretization of geographic data, and is capable of detecting arbitrary geo-linguistic dependencies (section 3).
- We use synthetic data to compare the power and calibration of HSIC against three alternatives: Moran's I, the Mantel Test, and join count analysis (section 4).
- We apply these methods to analyze dialectal variation in three empirical datasets, in both English and Dutch, across a variety of registers (section 5).

2. Prior Work

This section describes prior work on global methods for quantifying the degree of spatial dependence in a geotagged corpus.[2] While other global spatial statistics exist, we focus on the following three methods because they have been used in previous work on dialect analysis: Moran's I (Grieve, Speelman, and Geeraerts 2011), join count analysis (Lee and Kretzschmar Jr 1993) and the Mantel test (Scherrer 2012).

We define a consistent notation across methods. Let x_i represent a scalar linguistic observation for unit $i \in \{1 \dots n\}$ (typically, the presence or frequency of a linguistic variable), and let y_i represent a corresponding geolocation. For convenience, we define d_{ij} as the spatial distance between y_i and y_j. Suppose we have n observations, so that the data $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. Our goal is to test the strength of association between X and Y, against the null hypothesis that there is no association.

2.1 Moran's I

Grieve, Speelman, and Geeraerts (2011) introduced the use of Moran's I (Cliff and Ord 1981; Moran 1950) in the study of dialectal variation. To define the statistic, let $W = \{w_{ij}\}_{i,j \in \{1 \dots n\}}$ represent a *spatial weighting matrix*, such that larger values of w_{ij} indicate greater proximity, and $w_{ii} = 0$. In their application of Moran's I to a corpus of newspaper letters-to-the-editor, Grieve *et al.* define W as,

$$w_{ij} = \begin{cases} 1, & d_{ij} < \tau, i \neq j \\ 0, & d_{ij} \geq \tau, \text{ or } i = j \end{cases} \tag{1}$$

2 Global methods test for dependence over the entire dataset. In some cases, there will be local dependence in a few "hot spots", even when global dependence is not detected, and local autocorrelation statistics have been proposed to capture such dependences (Anselin 1995). For example, Grieve (2016) uses the Getis-Ord G_i statistic (Getis and Ord 1992) in his analysis of regional American English. Local statistics are particularly useful as an exploratory tool, but Grieve argues that the associated p-values are difficult to interpret due to the issue of multiple comparisons. We therefore focus on global tests in this paper. The adaptation of the proposed HSIC statistic into a local measure of dependence is an intriguing topic for future work.

where τ is some critical threshold (Grieve, Speelman, and Geeraerts 2011). When the spatial weighting matrix is defined in this way, Moran's I can be seen as a statistic that quantifies whether observations x_i and x_j are more similar when $w_{ij} = 1$ than when $w_{ij} = 0$.[3]

Moran's I is based on a hypothesized autoregressive process $X = \rho WX + \epsilon$, where X is a vector of the linguistic observations $x_1, \ldots x_n$, and ϵ is a vector of uncorrelated noise. Since X and W are given, the estimation problem is to find ρ so as to minimize the magnitude of ϵ. To take a probabilistic interpretation, it is typical to assume that ϵ consists of independent and identically distributed (IID) normal random variables with zero mean (Ord 1975). Under the null hypothesis of no spatial dependence between the observations in X, we would have $\rho = 0$. Note, however, that we may fail to reject the possibility that $\rho = 0$ even in the presence of spatial dependence, if the form of this dependence is not monotonic or nonlinear in W.

Because ρ is difficult to estimate exactly (Ord 1975), Moran's I is used as an approximation. It is computed as,

$$I = \frac{n}{\sum_i^n (x_i - \overline{x})^2} \frac{\sum_i^n \sum_j^n w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_i^n \sum_j^n w_{ij}}, \tag{2}$$

where $\overline{x} = \frac{1}{n} \sum_i x_i$. The ratio on the left is the inverse of the variance of X; the ratio on the right corresponds to the covariance between points i and j that are spatially similar. Thus, the statistic rescales a spatially-reweighted covariance (the ratio on the right of Equation 2) by the overall variance (the ratio on the left of Equation 2), giving an estimate of the overall spatial dependence of X. A compact alternative notation is to rewrite the statistic in terms of the vector of *residuals* $R = \{r_i\}_{i \in 1 \ldots n}$, where the residual $r_i = x_i - \overline{x}$. This yields the form $I = \frac{R^\top W R}{R^\top R}$, with $R^\top$ indicating the transpose of the column vector R, and with W assumed to be normalized so that $\sum_{i,j} w_{ij} = n$. Moran's I values often lie between -1 and 1, but the exact range depends on the weight matrix W, and is theoretically unbounded (de Jong, Sprenger, and van Veen 1984).

In hypothesis testing, our goal is to determine the p-value representing the likelihood that a value of Moran's I at least as extreme as the observed value would arise by chance under the null hypothesis. The expected value of Moran's I in the case of no spatial dependence is $-\frac{1}{n-1}$. Grieve *et al.* compute p-values from a closed-form approximation of the variance under the null hypothesis of total randomization. A nonparametric alternative is to perform a *permutation test*, calculating the empirical p-value by comparing the observed test statistic against the values that arise across multiple random permutations of the original data.

In either case, Moran's I does not test the null hypothesis of no statistical dependence between the linguistic features X and the geo-coordinates Y. Rather, it tests whether the estimated value of ρ would be likely to arise if there were no such dependence. But if the nature of the geo-linguistic dependence defies the assumptions of the statistic, then we risk incorrectly failing to reject the null hypothesis, a type II error. Put another way, there are forms of strong spatial dependence for which $\rho = 0$, such as non-monotonic spatial relationships. This risk can be somewhat ameliorated by careful

3 The matrix W can be defined in other ways. We can define a continuous-valued version of W by setting $w_{ij} = \exp(-\gamma d_{ij})$, with d_{ij} equal to the geographical distance between units i and j. Alternatively, we could define a topological spatial weighting matrix by setting $w_{ij} = 1$ when j is one of the k nearest neighbors of i (Getis and Aldstadt 2010).

choice of the spatial weighting matrix W, which could in theory account for non-linear or even non-monotonic dependencies. However an exhaustive search for some W that obtains a low p-value would clearly be an invalid hypothesis test, and so W must be fixed *before* any test is performed. In some cases, the researcher may bring substantive insights to the determination of W, and so the flexibility of Moran's I in this sense could be regarded as a positive feature. But there is little theoretical guidance, and a poor selection of W will result in inflated type II error rates.

From a practical standpoint, Moran's I is applicable to only some types of linguistic data. In the study of dialect, X typically represents the frequency or presence of some linguistic variable, such as the use of *soda* versus *pop*. We are unaware of applications of Moran's I to variables with more than two possibilities (e.g., *soda*, *pop*, *coke*). One possible solution would be to perform multiple tests, with each alternant pitted against all the others. But it is not clear how the p-values from these multiple tests should be combined. For example, selecting the minimum p-value across the alternants would mean that the null hypothesis would be more likely to be rejected for variables with more alternants; averaging the p-values across alternants would have the opposite problem.

2.2 Join Count Analysis

If the linguistic data X consists of discrete observations, *join count analysis* is another approach for detecting spatial dependence. For each pair of points (i, j), we compute $w_{ij}\delta(x_i = x_j)$, where $\delta(x_i = x_j)$ returns a value of 1 if x_i and x_j are identical, and 0 otherwise. As in Moran's I, w_{ij} is an element of a spatial weighting matrix, which could be binary or continuous. The global sum of the counts is computed as,

$$\text{num-agree} = \sum_i^n \sum_j^n w_{ij}(x_i x_j + (1 - x_i)(1 - x_j)) \tag{3}$$

$$= X^\top W X + (1 - X)^\top W (1 - X), \tag{4}$$

with $X^\top$ indicating the transpose of the column vector X. Note the similarity to the numerator of Moran's I, which can be written as $R^\top W R$. The number of agreements can be compared with its expectation under the null hypothesis, yielding a hypothesis test for global autocorrelation (Cliff and Ord 1981).

Join count analysis has been applied to the study of dialect by Lee and Kretzschmar Jr (1993), who take each linguistic observation $x_i \in \{0, 1\}$ to be a binary variable indicating the presence or absence of a dialect feature. They then build a binary spatial weighting matrix by performing a Delaunay triangulation over the geolocations of participants in dialect interviews, with $w_{ij} = 1$ if the edge (i, j) appears in the Delaunay triangulation. A nice property of Delaunay triangulation is that points tend to be connected to their closest neighbors, regardless of how distant or near those neighbors are: in high-density regions, the edges will tend to be short, while in low-density regions, the edges will be long. The method is therefore arguably more suitable to data in which the density of observations is highly variable — for example, between densely-populated cities and sparse-populated hinterlands.

Because join count statistics are based on counts of agreements, this form of analysis requires that each x_i is a categorical variable — possibly non-binary — rather than a frequency. In this sense, it is the complement of Moran's I, which can be applied to frequencies, but not to non-binary discrete variables. Thus, join count analysis is best

suited to cases where observations correspond to individual utterances (e.g., Twitter data, dialect interviews), rather than cases where observations correspond to longer texts (e.g., newspaper corpora).

2.3 The Mantel Test

The Mantel test can in principle be used to measure the dependence between any two arbitrary signals. In this test, we compute *distances* for each pair of linguistic variables, $d_x(x_i, x_j)$, and each pair of spatial locations, $d_y(y_i, y_j)$, forming a pair of distance matrices D_x and D_y. We then estimate the element-wise correlation (usually, the Pearson correlation) between these two matrices. Scherrer (2012) uses the Mantel test to correlate linguistic distance with geographical distance, and Gooskens and Heeringa (2006) correlate perceptual distance with linguistic distance. The Mantel test has also been applied to non-human dialect analysis, revealing regional differences in the call structures of Amazonian parrots by computing a linguistic distance matrix D_x directly from spectral measurements (Wright 1996).

Because it is built around distance functions, the Mantel test is applicable to binary, categorical, and frequency data — any kind of data for which a distance function can be constructed. For spatial locations, a typical choice is to compute the distance matrix based on the Euclidean distance between each pair of points. For binary or categorical linguistic data, the entries of the linguistic distance matrix can be set to 0 if $x_i = x_j$, and 1 otherwise. For linguistic frequency data, we use the absolute difference between the frequency values.

The role of hypothesis testing in the Mantel test is to determine the likelihood that the observed test statistic — in this case, the correlation between the distance matrices D_x and D_y — could have arisen by chance under the null hypothesis. In the ideal case of perfect correlation, twice as much geographical distance should imply twice as much as linguistic distance. But this situation is highly unlikely to obtain in practice. In fact, as noted by Grieve (2014), such a perfect correlation *cannot* arise from any linguistic data involving a single variable: even if a linguistic variable obeys a perfect dialect continuum (e.g., varying in frequency from east to west), the distances in the orthogonal north-south direction would diminish the resulting correlation. In realistic settings in which the geo-linguistic dependence is obscured by noise, this can dramatically diminish the power of the test. Note that even non-linear transformations of the distance metric would not correct this issue. The key problem, as identified by Legendre, Fortin, and Borcard (2015), is that the Mantel test is not designed to test for independence between X and Y, but rather, the correlation between *distances* on X and Y. When distances are the primary units of analysis — as, for example, in the work of Gooskens and Heeringa (2006) — the test is applicable. But for the task of determining whether a specific linguistic variable is geographically dependent, the test is incorrectly applied; as we show in section 4, this results in inflated Type II error rates.

2.4 Other Related Work

Several computational studies attempt to characterize linguistic differences across geographical regions, although most of these studies do not perform hypothesis testing on geographical dependence. A common approach is to aggregate geotagged social media content into geographical bins. Some studies rely on politically defined units such as nations and states (Hovy, Johannsen, and Søgaard 2015); however, *isoglosses* (the geographical boundaries between linguistic features) need not align with politically-

defined geographical units (Nerbonne and Kretzschmar Jr 2013). Other approaches rely on automatically defined geographical units, induced by computational methods such as geodesic grids (Wing and Baldridge 2011), KD-trees (Roller et al. 2012), Gaussians (Eisenstein et al. 2010), and mixtures of Gaussians (Hong et al. 2012). While these approaches offer insights about the nature of geographical language variation, they do not provide test statistics that allow us to quantify the geographical dependence of various linguistic features.

As described in the next section, our approach is based on Reproducing Kernel Hilbert Spaces, which enable us to nonparametrically compare probability distributions. Another way in which kernel methods can be applied to spatial analysis is in Gaussian Processes, which are often used to represent spatial data (Cressie 1988; Ecker and Gelfand 1997). Specifically, we can define a kernel over space, so that a response variable is distributed as a Gaussian with covariance defined by the kernel function. For example, it might be possible to model the popularity of linguistic features as a Gaussian Process, using the spatial covariance kernel to make smooth predictions at unknown locations. Our approach in this paper is different, as we are interested in hypothesis testing, rather than modeling and prediction. Another difference is that we apply kernels to both the geographical and linguistic data sources, while a Gaussian Process approach would make the parametric assumption that the linguistic signal is Gaussian distributed with covariance defined by the spatial covariance kernel.

3. Hilbert-Schmidt Independence Criterion (HSIC)

Moran's I, join count analysis, and the Mantel test share an important drawback: they do not directly test the independence of language and geography, but rather, they test for autocorrelation between X and Y, or between distances on these variables. Moran's I tests whether the parameter of a linear autoregressive model is nonzero; the Mantel test is performed on the correlations between pairwise distances; join count statistics enable tests of whether spatially adjacent units tend to have the same linguistic features. In each case, rejection of the null hypothesis implies dependence between the geographical and linguistic signals. However, each test can incorrectly fail to reject the null hypothesis if its assumptions are violated, even if given an arbitrarily large amount of data.

We propose an alternative approach: directly test for the independence of geographical and linguistic variables, $P_{XY} \stackrel{?}{=} P_X P_Y$. Our approach, which is based on the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al. 2005a, 2008), makes no parametric assumptions about the form of these distributions. The proposed test is *consistent*, in the sense that it will always reach the right decision, if provided enough data (Fukumizu et al. 2007).[4]

To test independence for arbitrary distributions P_{XY}, P_X, and P_Y, HSIC employs the framework of Reproducing Kernel Hilbert Spaces (RKHS). This framework will be familiar to the computational linguistics community through its application to support vector machines (Collins and Duffy 2001; Lodhi et al. 2002), where *kernel similarity functions* between pairs of instances are used to induce classifiers with highly nonlinear decision boundaries. HSIC is a kernelized independence test, and it offers an analogous advantage: by computing kernel similarity functions on pairs of observations, it is possible to implicitly compare probability distributions across high-order

4 HSIC was introduced as a test of independence by Gretton et al. (2008). In this paper, we present the first application to computational linguistics.

moments, enabling nonparametric independence tests that are statistically consistent. An additional advantage of the RKHS framework is that it can be applied to arbitrary linguistic data — including dichotomous, polytomous, continuous, and vector-valued observations — as long as an appropriate kernel similarity function can be defined.

Intuitively, HSIC tests independence by approximating a measure of the discrepancy between the joint geo-linguistic distribution P_{XY} and the product of independent distributions $P_X P_Y$. The forms of these distributions are unknown; for example, P_Y might be Gaussian, or it might be some complicated multimodal distribution. The *maximum mean discrepancy* is a scalar function of the discrepancy between a pair of distributions, which makes no assumption about the distributions' parametric forms. The maximum mean discrepancy will be large when linguistic similarity tends to co-occur with geographical similarity, indicating an association between language and geography that is unlikely to arise by chance. The key insight is that it is possible to approximate the maximum mean discrepancy from a finite sample of observations, by rewriting it as a sum of kernel similarity functions. These kernel functions should quantify the similarity between each pair of instances as a scalar; they must also obey some more technical properties, enumerated in section 3.3. If the kernel functions are appropriately chosen, then the approximation is asymptotically consistent, meaning that it will approach the exact maximum mean discrepancy in the limit of infinite data, regardless of the forms of P_X, P_Y, and P_{XY}. (This property is not shared by the Mantel test, which is superficially similar in that it operates on distances between pairs of observations.) We now present the mathematical details of the method.

3.1 Comparing Probability Distributions

The *maximum mean discrepancy* (MMD) is a nonparametric statistic that compares two arbitrary probability distributions. In the HSIC test, this statistic is used to compare the joint distribution P_{XY} with the product of marginal distributions $P_X P_Y$. The MMD is defined as,

$$\text{MMD}(P, Q) = \sup_f (\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)]), \tag{5}$$

where we take the supremum f over a set of possible functions, and compute the difference in the expected values under the distributions P and Q. Clearly, if $P = Q$, then $\text{MMD} = 0$, but the challenge is to estimate MMD for arbitrary P and Q, based only on finite samples from these distributions.

To explain how to do this, we introduce some concepts from Reproducing Kernel Hilbert Spaces (RKHS). Let $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}_+$ denote a kernel function, mapping from pairs of observations (x_i, x_j) to reals. A classical example is the radial basis function (RBF) kernel on vectors, where $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||_2^2\right)$. Many other kernel functions are possible; the conditions for valid kernels are enumerated in section 3.3.

For any instance x, the kernel function k defines a corresponding *feature map* $k(\cdot, x) : \mathcal{X} \mapsto \mathbb{R}_+$, which is the function that arises by fixing one of the arguments of the kernel function to the value x. The "reproducing" property of RKHS implies an identity between kernel functions and inner products of feature maps:

$$k(x_i, x_j) = \langle k(\cdot, x_i), k(\cdot, x_j) \rangle. \tag{6}$$

Thus, even though the feature map may be an arbitrarily complex function of x, we can compute inner products of feature maps by computing the kernel similarity function over the associated instances. The MMD can be expressed in terms of such inner products, and thus, can be computed in terms of kernel similarity functions.

For a probability measure P, the *mean element* of P is defined as the expected feature map, $\mu_P = \mathbb{E}_P k(\cdot, x)$. The MMD can then be computed in terms of kernel functions of the mean elements,

$$\text{MMD}^2(P, Q) = \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2\langle \mu_P, \mu_Q \rangle. \tag{7}$$

If $\mu_P = \mu_Q$, then the MMD is zero. The key observation is that $\mu_P = \mu_Q$ if and only if $P = Q$, so long as an appropriate kernel similarity function is chosen (Fukumizu et al. 2007); see section 3.3 for more on the choice of kernel functions.

Each of the inner products in Equation 7 corresponds to an expectation that can be estimated empirically from finite samples $\{x_1, x_2, \ldots, x_m\}$ and $\{y_1, y_2, \ldots, y_n\}$,

$$\text{MMD}^2(P, Q) = \mathbb{E}_{x,x' \sim P} k(x, x') + \mathbb{E}_{y,y' \sim Q} k(y, y') - 2\mathbb{E}_{x \sim P, y \sim Q} k(x, y) \tag{8}$$

$$\widehat{\text{MMD}^2}(P, Q) = \frac{1}{m^2} \sum_{i,j}^{m} k(x_i, x_j) + \frac{1}{n^2} \sum_{i,j}^{n} k(y_i, y_n) - \frac{2}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j). \tag{9}$$

The full derivation is provided by Gretton et al. (2008). Having shown how to estimate a statistic on whether two probability measures are identical, we now use this statistic to test for independence.

3.2 Derivation of HSIC

To construct an independence test over random variables X and Y, we test the MMD between the joint distribution P_{XY} and the product of marginals $P_X P_Y$. In this setting, each observation i corresponds to a pair (x_i, y_i), so we require a kernel function on paired observations, $k((x_i, y_i), (x_j, y_j))$. We define this as a *product kernel*,

$$k((x_i, y_i), (x_j, y_j)) = k_{\mathcal{X}}(x_i, x_j) k_{\mathcal{Y}}(y_i, y_j), \tag{10}$$

where $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are kernels for the linguistic and geographic observations respectively.

Using the product kernel, we can define mean embeddings for the distributions P_{XY} and $P_X P_Y$, enabling the application of the MMD estimator from Equation 9. The Hilbert-Schmidt Independence Criterion (HSIC) is precisely this application of maximum mean discrepancy to compare the joint distribution against the product of marginal distributions.

Concretely, let us define the *Gram matrix* K_x so that $(K_x)_{i,j} = k_{\mathcal{X}}(x_i, x_j)$ for all pairs i, j in the sample. Analogously, $(K_y)_{i,j} = k_{\mathcal{Y}}(y_i, y_j)$. Then the HSIC can be estimated

from a finite sample of m observations as

$$\widehat{\text{HSIC}} = \frac{1}{n^2}\sum_{i,j}^{m}(K_x)_{i,j}(K_y)_{i,j} + \frac{1}{n^4}\sum_{i,j,q,r}^{m}(K_x)_{i,j}(K_y)_{q,r} - \frac{2}{n^3}\sum_{i,j,q}^{m}(K_x)_{i,j}(K_y)_{i,q} \quad (11)$$

$$= \frac{\mathbf{tr} K_{\mathcal{X}} H K_{\mathcal{Y}} H}{n^2}, \quad (12)$$

where $\mathbf{tr}$ indicates the matrix trace and H is a centering matrix, $H = \mathbb{I}_m - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. With this definition of H, we have,

$$(K_{\mathcal{X}} H)_{ij} = k_{\mathcal{X}}(x_i, x_j) - \frac{1}{n}\sum_{j'} k_{\mathcal{X}}(x_i, x_{j'}) \quad (13)$$

$$(K_{\mathcal{Y}} H)_{ij} = k_{\mathcal{Y}}(y_i, y_j) - \frac{1}{n}\sum_{j'} k_{\mathcal{Y}}(y_i, y_{j'}). \quad (14)$$

These two terms can be seen as mean-centered Gram matrices. By computing the trace of their matrix product, we obtain a cross-covariance between the Gram matrices. This trace is directly proportional to the maximum mean discrepancy between P_{XY} and $P_X P_Y$. If X and Y are dependent, then large values of $k_{\mathcal{X}}(x_i, x_j)$ will imply large values of $k_{\mathcal{Y}}(y_i, y_j)$ — similar geography implies similar language — and so the cross-covariance will be greater than zero. If X and Y are independent, then large values of $k_{\mathcal{X}}(x_i, x_j)$ are not any more likely to correspond to large values of $k_{\mathcal{Y}}(y_i, y_j)$, and so the expectation of this cross-covariance will be zero.

3.3 Kernel Functions

To apply HSIC to the problem of detecting geo-linguistic dependence, we must define the kernel functions $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$. In the RKHS framework, valid kernel functions must be symmetric and positive semi-definite. To ensure consistency of the kernel-based estimator for MMD, the kernel must also be *characteristic*, meaning that it induces an injective mapping between probability measures and their corresponding mean elements (Fukumizu et al. 2007): each probability measure P must correspond to a single unique mean element μ_P. Muandet *et al.* elaborate these and other properties of several well-known kernels (Muandet et al. 2016, Table 3.1).

For the spatial kernel $k_{\mathcal{Y}}$, we employ a Gaussian radial basis function (RBF), which is a widely used choice for vector data. Specifically, we define $k_{\mathcal{Y}}(y_i, y_j; \gamma) = \exp(-\gamma d_{i,j}^2)$, where d_{ij}^2 is the squared Euclidean distance between y_i and y_j, and γ is a parameter of the kernel function. We also employ the RBF in $k_{\mathcal{X}}$ when the linguistic observations take on continuous values, such as frequencies or phonetic data. The RBF kernel is symmetric, positive semi-definite, and characteristic.[5]

5 Flaxman recently proposed a "kernelized Mantel test", in which correlations are taken between kernel similarities rather than distances (Flaxman 2015). The resulting test statistic is similar, but not identical to HSIC. Specifically, while HSIC centers the kernel matrix against the local mean kernel similarities for each point, the kernelized Mantel test centers against the global mean kernel similarity. This makes the test more sensitive to distant outliers. We implemented the kernelized Mantel test, and found its performance to be similar to the classical Mantel test, with lower statistical power than HSIC. Flaxman made similar observations in his analysis of the spatiotemporal distribution of crime events.

The parameter γ corresponds to the "length-scale" of the kernel. Intuitively, as this parameter increases, the kernel similarity drops off more quickly with distance. In this paper, we follow the popular heuristic of setting γ to the median of the data $(y_1, \dots y_n)$, as proposed by Gretton et al. (2005b). We empirically test the sensitivity of HSIC to this parameter in section 4. More recent work offers optimization-based approaches for setting this parameter (Gretton et al. 2012), but we do not consider this possibility here.

Linguistic data is often binary or categorical. In this case, we use a Delta kernel (also sometimes called a Dirac kernel). This kernel is simply defined as $k_{\mathcal{X}}(x_i, x_j) = 1$ if $x_i = x_j$ and 0 otherwise. The Delta kernel has been used successfully in combination with HSIC for high-dimensional feature selection (Song et al. 2012; Yamada et al. 2014), and is symmetric, positive semi-definite, and characteristic for discrete data. For continuous or vector-valued linguistic variables, the RBF kernel can again be applied.

3.4 Scalability

The size of each Gram matrix is the square of the number of observations. For large datasets, this will be too expensive to compute. Following Gretton et al. (2005a), we employ a low-rank approximation to each Gram matrix, using the incomplete Cholesky decomposition (Bach and Jordan 2002). Specifically, we approximate the symmetric matrices $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$ as low-rank products, $K_{\mathcal{X}} \approx AA^T$ and $K_{\mathcal{Y}} \approx BB^T$, where $A \in \mathbb{R}^{n \times r_A}$ and $B \in \mathbb{R}^{n \times r_B}$. The approximation quality is determined by the parameters r_A and r_B, which are set to ensure that the magnitudes of the residuals $K - AA^T$ and $L - BB^T$ are below a predefined threshold. HSIC may then be approximated as:

$$\widehat{\text{HSIC}} = \frac{\mathbf{tr} K_{\mathcal{X}} H K_{\mathcal{Y}} H}{n^2}, \tag{15}$$

$$\approx \frac{\mathbf{tr}(AA^T)H(BB^T)H}{n^2}, \tag{16}$$

$$= \frac{\mathbf{tr}(B^T(HA))(B^T(HA))^T}{n^2} \tag{17}$$

where the matrix product HA can be computed without explicitly forming the $n \times n$ matrix H, due to its simple structure. Alternative methods for scaling the computation of HSIC are discussed in a recent note by Zhang et al. (2016).

4. Synthetic Data

Real linguistic datasets lack ground truth about which features are geographically distinct, making it impossible to use such data to quantitatively evaluate the proposed approaches. We therefore use synthetic data to compare the power and sensitivity of the various approaches described above. Our main goals are: (1) to calibrate the p-values produced by each approach in the event that the null hypothesis is true, using completely randomized data; (2) to test the power of each approach to capture spatial dependence, particularly under conditions in which the spatial dependence is obscured by noise.

We compare HSIC with specific instantiations of the methods described in section 2, focusing on previous published applications of these methods to dialect analysis. Specifically, we consider the following methods:

Moran's I We follow Grieve, Speelman, and Geeraerts (2011), using a binary spatial weighting matrix with a distance threshold τ, usually set to the median of the distances between points in the dataset[6]. This method is not applicable to categorical data.

Join counts We follow the approach of Lee and Kretzschmar Jr (1993), who define a binary spatial weighting matrix from a Delaunay triangulation, and then compute join counts for linked pairs of observations. This method is not applicable to frequency data.

Mantel test We use Euclidean distance for the geographical distance matrix. For continuous linguistic data, we also use Euclidean distance; for discrete data, we use a delta function.

For all approaches, a one-tailed significance test is appropriate, since in nearly all conceivable dialectological scenarios we are testing only for the possibility that geographically proximate units are *more* similar than they would be under the null hypothesis. For some methods, it is possible to calculate a p-value from the test statistic using a closed form estimate of the variance. However, for consistency, we employ a permutation approach to characterize the null distribution over the test statistic values. We permute the linguistic data x, breaking any link between geography and the language data, and then compute the distribution of the test statistic under many such permutations.

4.1 Data Generation

To ensure the verisimilitude of our synthetic data, we target the scenario of geo-tagged tweets in the Netherlands. For each municipality i, we stochastically determine the number and location of the tweets as follows:

Number of data points For each municipality, the number of tweets n_i is chosen to be proportional to the population, as estimated by Statistics Netherlands (CBS). Specifically, we draw $\tilde{n}_i \sim \text{Poisson}(\mu_{obs} \times \text{population}_i)$ and then set $n_i = \tilde{n}_i + 1$, ensuring that each municipality has at least one data point. The parameter μ_{obs} controls the frequency of the linguistic variable. For example, a common orthographic variable (e.g., "g-deletion") might have a high value of μ_{obs}, while a rare lexical variable (e.g., *soda* versus *pop*) might have a much lower value. Note that μ_{obs} is shared across all municipalities.

Locations Next, for each tweet t, we determine the location y_t by sampling without replacement from the set of real tweet locations in municipality i (the dataset is described in section 5.3). This ensures that the distribution of geo-locations in the synthetic data matches the real geographical distribution of tweets, rather than drawing from a parametric distribution which may not match the complexity of true geographical population distributions. Each location is represented as a latitude and longitude pair.

For each variable, each municipality is assigned a frequency vector θ_i, indicating the relative frequency of each variable form: e.g., 70% *soda*, 30% *pop*. We discuss methods for setting θ_i below, which enable the simulation of a range of dialectal phenomena.

6 Other types of spatial weighting matrices might give different results. We leave this for future work.

(a) Angle: 0 degrees (east to west)

(b) Angle: 120 degrees

Figure 2: Synthetic frequency data with dialect continua in two different angles

We simulate both counts data and frequency data. In counts data — such as geotagged tweets — the data points in each instance in municipality i are drawn from a binomial or multinomial distribution with parameter θ_i. In frequency data, we observe only the relatively frequency of each variable form for each municipality. In this case, we draw the frequency from a Dirichlet distribution with expected value equal to θ_i, drawing $\phi_t \sim \text{Dirichlet}(s\theta_i)$, where the scale parameter s controls the variance within each municipality.

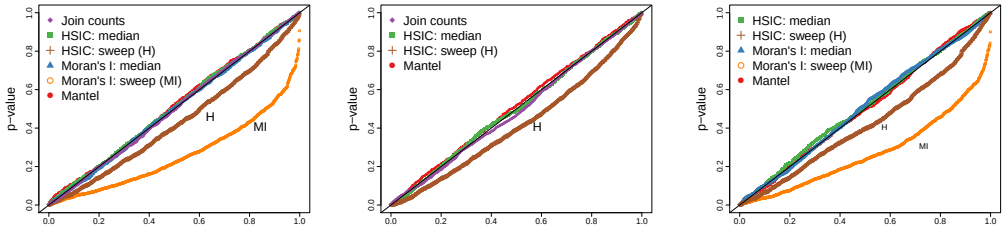
4.2 Calibration

Our first use of synthetic data is to examine the p-values obtained from each method when the null hypothesis is true — that is, when there is no geographical variation in the data. The p-value corresponds to the likelihood of seeing a test statistic at least as extreme as the observed value, under the null hypothesis. Thus, if we repeatedly generate data under the null hypothesis, a well-calibrated test will return a distribution of p-values that is uniform in the interval $[0, 1]$. We would expect to observe $p < .05$ in exactly 5% of cases, corresponding to the allowed rate of Type I errors (incorrect rejection of the null hypothesis) at the threshold $\alpha = 0.05$.

To measure the calibration of each of the proposed tests, we generate 1,000 random datasets using the procedure described above, and then compute the p-values under each test. In these random datasets, the relative frequency parameters θ_i are the same for all municipalities, which is the null hypothesis of complete randomization. To generate the binary and categorical data, we use $\mu_{obs} = 10^{-5}$, meaning that the expected number of observations is one per hundred thousand individuals in the municipality; for comparison, this corresponds roughly to the tweet frequency of the lengthened spelling *hellla* in the 2009-2012 Twitter dataset gathered by Eisenstein et al. (2014).

To visualize the calibration of each test, we use quantile-quantile (Q-Q) plots, comparing the obtained p-values with a uniform distribution. A well-calibrated test should give a diagonal line from the origin to $(1, 1)$. Figure 3 shows the Q-Q plots obtained from each method on each relevant type of data (recall that not all methods can be applied to all types of data, as described in the previous section).

HSIC and Moran's I each have tuning parameters that control the behavior of the test: the kernel bandwidth in HSIC and the distance cutoff in Moran's I. A simple heuristic is to use the median Euclidian distance $\overline{d}$: in Moran's I, we use $\overline{d}$ as the distance threshold for constructing the neighborhood matrix W; in HSIC, we use $\frac{1}{\overline{d}^2}$ as the kernel bandwidth parameter. Figure 3 shows that by basing these parameters on the



(a) Binary data (b) Categorical data (c) Frequency data

Figure 3: Quantile-quantile plots comparing the distribution of the obtained p-values with a uniform distribution. The y-axis is the p-value returned by the tests. The x-axis shows the corresponding quantile for a uniform distribution on the range [0,1]. The approaches that optimize the parameters, i.e., the cutoff for Moran's I (MI) and the bandwidth for HSIC (H), lead to a skewed distribution of p-values.

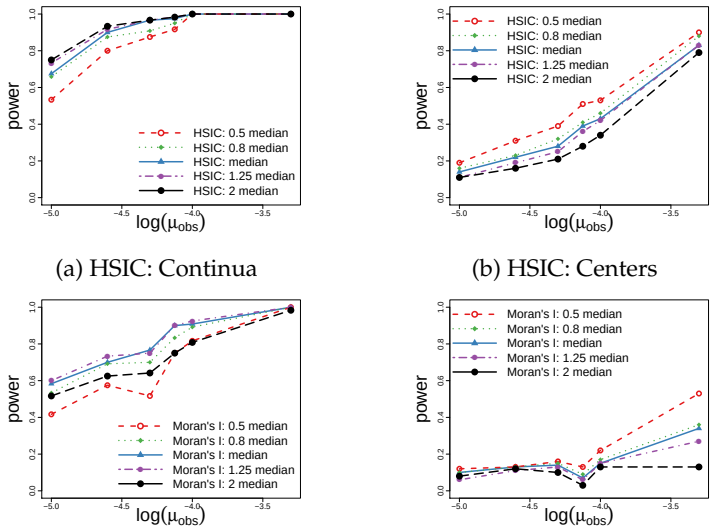
median distance between pairs of points, we get well-calibrated results. However, some prior work takes an alternative approach, sweeping over parameter values to obtain the most significant results (Grieve, Speelman, and Geeraerts 2011). In our experiments we sweep across the distance cutoff for Moran's I, and the bandwidth for the spatial distances in HSIC. This distorts the calibration, meaning that the resulting p-values are not reliable. This is most severe for Moran's I with type I error rates of 11.7% (binary data) and 14.3% (frequency data) when the significance threshold α is set to 5%. Given that such parameter sweeps are explicitly designed to maximize the number of positive test results — and not the overall calibration of the test — this is unsurprising. We therefore avoid parameter sweeps in the remainder of this article, and rely instead on median distance as a simple heuristic alternative.

4.3 Power

Next, we consider synthetic data in which there is geographical variation by construction. We assess the *power* of each approach by computing the fraction of simulations for which the approaches correctly rejected the null hypothesis of no spatial dependence, given a significance threshold of $\alpha = 0.05$. We again use the Netherlands as the stage for all simulations, and consider two types of geographical variation.

Dialect continua We generate data such that the frequency of a linguistic variant increases linearly through space, as in a dialect continuum (Heeringa and Nerbonne 2001). In most of the synthetic data experiments below, we average across a range of angles, from $0°$ to $357°$ with step sizes of $3°$, yielding 120 distinct angles in total. Each angle aligns differently with the population distribution of the Netherlands, so we also assess sensitivity of each method to the angle itself. Figure 2 shows two synthetic datasets with dialect continua in different angles.

Geographical centers Second, we consider a setting in which variation is based on one or more geographical *centers*. In this setting, all cities within some specified range of the center (or one of the centers) have some maximal frequency value θ_i; in other cities, this value decreases as distance from the nearest center grows. This corresponds to the dialectological scenario in which a variable form is centered on one specific city, as in, say, the association of the word *hella* with the San Francisco



(a) HSIC: Continua (b) HSIC: Centers

(c) Moran's I: Continua (d) Moran's I: Centers

Figure 4: Power across different parameter settings. Higher values indicate a greater likelihood of correctly rejecting the null hypothesis.

metropolitan area. We average across twenty five possible centers: the capitals of each of the twelve provinces of the Netherlands; the national capital of the Netherlands (Amsterdam); the *two* most populous cities in each of the twelve provinces. For each setting, we randomly generate synthetic data four times, resulting in a total of 100 synthetic datasets for this condition.

Parameter settings. We use these data generation scenarios to test the sensitivity of HSIC and Moran's I to their hyperparameters, by varying the kernel bandwidths in HSIC (Figures 4a and 4b) and the distance threshold in Moran's I (Figures 4c and 4d). The sensitivity of HSIC to the bandwidth value decreases as the number of data points increases (as governed by μ_{obs}), especially in the case of dialect continua. The sensitivity of Moran's I to the distance cutoff value decreases with the amount of data in the case of dialect continua, but in the case of center-based variation, Moran's I becomes *more* sensitive to this parameter as there is more data. For both methods, the same trends regarding the best performing parameters can be observed. In the case of dialect continua, larger cutoffs and bandwidths perform best, but in the case of variation based on centers, smaller cutoffs and bandwidths lead to higher power. Overall, there is no single best parameter setting, but the median heuristics perform reasonably well for both types of variation.

Direction of dialect continua. We simulate dialect continua by varying the frequency of linguistic variables linearly through space. Due to the heterogeneity of population density, different spatial angles will have very different properties: for example, one choice of angle would imply a continuum cutting through several major cities, while another choice might imply a rural-urban distinction. Figure 5 shows the power of the methods on binary data (there are two variant forms, and each instance contains exactly

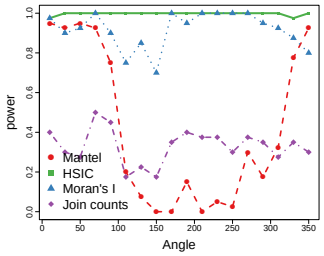
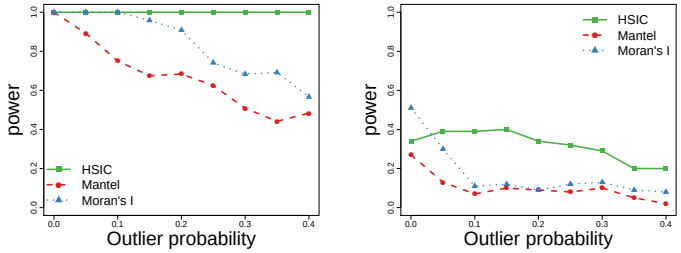


Figure 5: Relationship between the statistical power of each test and the angle of the dialect continuum across the Netherlands.



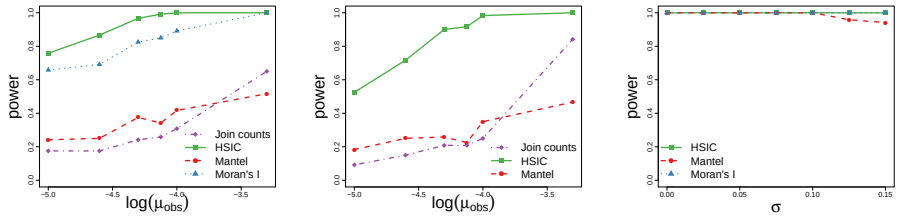
(a) Dialect continua (b) Centers

Figure 6: Results on synthetic frequency data ($\sigma = 0.1$) with outliers

one of them), in which we vary the angle of the continuum. HSIC is insensitive to the angle of variation, demonstrating the advantage of this kernel nonparametric method. Moran's I is relatively robust, while join count analysis performs poorly across the entire range of settings. The Mantel test is remarkably sensitive to the angle of variation, attaining nearly zero power for some scenarios of dialect continua. This is caused by the complex interaction between the underlying linguistic phenomenon and the east-west variation of the population density of the Netherlands. For example, when the dialect continuum is simulated at an angle of 105 degrees, the south east of the Netherlands has a higher usage of the variable, but this is only a very small region due to the shape of the country. The Mantel test apparently has great difficulty in detecting geographical variation in such cases.

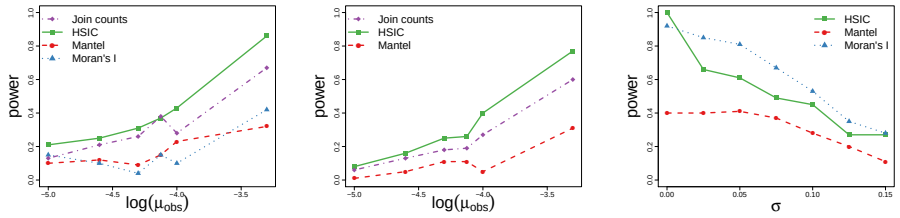
Outliers. In the frequency-based synthetic data, each instance uses each variable form with some continuous frequency — this is based on the scenario of letters-to-the-editors of regional newspapers, as explored in prior work (Grieve, Speelman, and Geeraerts 2011). We test the robustness of each approach by introducing *outliers*: randomly selected data points whose variable frequencies are replaced at random with extreme values of either 0 or 1. As shown in Figure 6, HSIC is the most robust against outliers, while the performance of the Mantel test is the most affected by outliers (recall that join count analysis applies only to discrete observations, so it cannot be compared on this measure).

Overall. We now compare the methods by averaging across various settings simulating dialect continua (Figure 7) and variation based on centers (Figure 8). To generate the



(a) Binary variable (b) Categorical variable (c) Frequency variable

Figure 7: Dialect continua



(a) Binary variable (b) Categorical variable (c) Frequency variable

Figure 8: Centers

categorical data, we vary μ_{obs} in our experiments, with a higher μ_{obs} resulting in more tweets and consequently less variation on the municipality level. As expected, the power of the approaches increases as μ_{obs} increases in the experiments on the categorical data, and the power of the approaches decreases as σ increases in the experiments on the frequency data. The experiments on the binary and categorical data show the same trends: HSIC performs the best across all settings. Join count analysis does well when the variation is based on centers, and Moran's I does best for dialect continua. Moran's I performs best on the frequency data, especially in the case of variation based on centers.

4.4 Summary

In this section, we evaluated each statistical test for geographical language variation on a battery of synthetic data. HSIC and the Mantel test are the only approaches applicable to all data types (binary, categorical and frequency data). Overall, HSIC is more effective than the Mantel test, which is much more sensitive to the specifics of the synthetic data scenario, such as the angle of the dialect continuum. HSIC is robust against outliers, and performs particularly well when the number of data points increases. Join count analysis is suitable for capturing non-linear variation, but its power is low compared to other approaches in the analysis of dialect continua. Conversely, when the linguistic data is binary, Moran's I performs well on dialect continua, but its power is low in situations of variation based on centers. In our experiments on frequency data, Moran's I performs well in both scenarios.

5. Empirical Data

We now assess the spatial dependence of linguistic variables on three real linguistic datasets: letters to the editor (English), a syntactic atlas of the Dutch dialects, and Dutch geotagged tweets. In each dataset, we compute statistical significance for the geo-linguistic dependence of multiple linguistic variables. To adjust the significance thresholds for multiple comparisons, we use the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) to bound the overall false discovery rate (FDR).

5.1 Letters to the Editor

In their application of Moran's I to English dialects in the United States, Grieve, Speelman, and Geeraerts (2011) compile a corpus of letters-to-the-editors of newspapers to measure the presence of dialect variables in text. To compute the frequency of the lexical variables, letters are aggregated to core-based statistical areas (CBSA), which are defined by the United States to capture the geographical region around an urban core. The frequency of 40 manually selected lexical variables is computed for each of 206 cities.

We use the Mantel test, HSIC, and Moran's I to assess the spatial dependence of variables in this dataset. Join count analysis was excluded, because it is not suitable for frequency data. We verified our implementation of Moran's I by following the approach taken by Grieve *et al.*: we computed Moran's I for cutoffs in the range of 200 to 1000 miles and selected the cutoff that yielded the lowest p-value. The obtained cutoffs and test statistics closely followed the values reported in the analysis by Grieve *et al.*, with slight deviations possibly due to our use of a permutation test rather than a closed-form approximation to compute the p-values.

After adjusting the p-values using the false discovery rate (FDR) procedure, a 500-mile cutoff results in three significant linguistic variables.[7] However, recall that the approach of selecting parameters by maximizing the number of positive test results tends to produce a large number of Type I errors. When setting the distance cutoff to the median distance between data points, none of the linguistic variables were found to have a significant geographical association. Similarly, HSIC and the Mantel test also found no significant associations after adjusting for multiple comparisons. Figure 9 shows the proportion of significant variables according to Moran's I based on different thresholds. The numbers vary considerably depending on the threshold. The figure also suggests that the median distance (921 miles) may not be a suitable threshold for this dataset.

5.2 Syntactic Atlas of the Dutch Dialects (SAND)

Dialect atlases are frequently used in the study of dialect. In this section we demonstrate the use of the discussed methods on SAND (Barbiers et al. 2005, 2008), an online electronic atlas that maps syntactic variation of Dutch varieties in the Netherlands, Belgium, and France.[8] The data was collected between the years of 2000 and 2005. SAND has been used to measure the distances between dialects, and to discover dialect regions (Spruit 2006; Tjong Kim Sang 2015). To our knowledge, we are the first to use statistical

7 Grieve *et al.* report five significant variables. In our analysis, there are two variables with FDR-adjusted p-values of 0.0559

8 http://www.meertens.knaw.nl/sand/

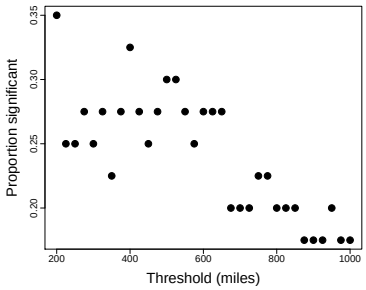


Figure 9: The proportion of variables detected to be significant $(p < .05)$ by Moran's I by varying the distance cutoff (without adjusting for multiple comparisons).

methods to quantify the degree of spatial dependence of the linguistic variables in this atlas. Compared to the other empirical datasets that we consider, this data is smaller and many variables contain more than two variants.

In our experiments, we consider only locations within the Netherlands (157 locations). The number of variants per linguistic variable ranges from one (due to our restriction to the Netherlands) to eleven. We do not include Moran's I in our experiments, since it is not applicable to linguistic variables with more than two variants. We apply the remaining methods to all linguistic variables with twenty or more data points and at least two variants, resulting in a total of 143 variables.

Table 1 lists the 10 variables with the highest HSIC values. Statistical significance at a level of $\alpha = 0.05$ is detected for 65.0% of the linguistic variables using HSIC, 78.3% when using join count analysis, and 52.4% when using the Mantel test. The three methods agree on 99 out of the 143 variables, and HSIC and join count analysis agree on 118 variables. From manual inspection, it seems that the non-linearity of the geographical patterns may have caused difficulties for the Mantel test. Figure 10 is an example of a variable where HSIC and join count analysis both had an FDR-adjusted $p < .05$, but the Mantel test did not detect a significant association.

5.3 Twitter

Our Twitter dataset consists of 4,039,786 geotagged tweets from the Netherlands, written between January 1, 2015 and October 31, 2015. We manually selected a set of linguistic variables (Table 2), covering examples of lexical variation (e.g., two different words for referring to french fries), phonological variation (e.g., t-deletion), and syntactic variation (e.g., *heb gedaan* ('have done') vs. *gedaan heb* ('done have')). We are not aware of any previous work on dialectal variation in the Netherlands that uses spatial dependency testing on Twitter data. The number of tweets per municipality varies dramatically, and for the less frequent linguistic variables there are no tweets at all in some municipalities. In our computation of Moran's I, we only include municipalities with at least one tweet.

Table 2 shows the output of each statistical test for this data. Some of these linguistic variables exhibit strong spatial variation, and are identified as statistically significant by all approaches. An example is the different ways of referring to french fries (*friet* versus *patat*, Figure 11a), where the figure shows a striking difference between the south and the north of the Netherlands. Another example is Figure 11b, which shows two different

Table 1: Highest ranked variables by HSIC. All methods had an adjusted p-value $< .05$ for all variables, except variable 2:30b (Mantel: $p = .118$.

Map id	Description
1:84b	Free relative, complementizer following relative pronoun
1:84a	Short subject and object relative, complementizer following relative pronoun
1:80b	ONE pronominalisation
1:33a	Complementizer agreement 3 plural
1:76a	Reflexive pronouns; synthesis
2:36b	Form of the participle of the modal verb willen 'want'
1:29a	Complementizer agreement 1 plural
1:69a	Correlation weak reflexive pronouns
2:30b	Interruption of the verbal cluster; synthesis I
2:61a	Forms for iemand 'somebody'

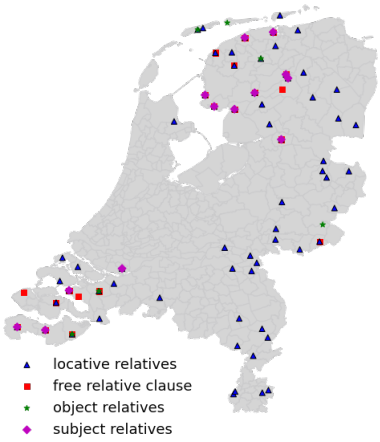


Figure 10: SAND map 16b, book 1. Finite complementizer(s) following relative pronoun (N=112). HSIC: p=.024; Mantel: p=.506; Join counts: p=.002.

ways of saying 'for a little while' (*efkes* versus *eventjes*). The less common form, *efkes* is mostly used in Friesland, a province in the north of the Netherlands.

Examples of linguistic variables where the approaches disagree are shown in Figure 12. The first case (Figure 12a) is an example of lexical variation, with two different ways of saying *bye* in the Netherlands. A commonly used form is *doei*, while *houdoe* is known to be specific to North-Brabant, a Dutch province in the south of the Netherlands. HSIC and join count analysis both detect a significant pattern, but Moran's I and the Mantel test do not. The trend is less strong than in the previous examples, but the figure does suggest a higher usage of *houdoe* in the south of the Netherlands.

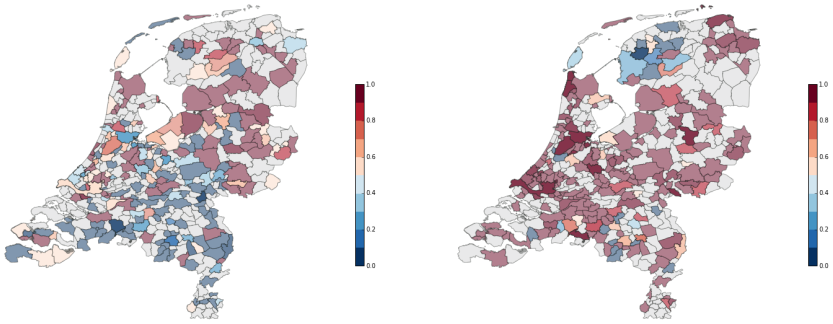
Another example is t-deletion for a specific phrase (*niet meer* versus *nie meer*), as shown in Figure 12b. Previous dialect research has found that geography is the most important external factor for t-deletion in the Netherlands, with contact zones, such as the Rivers region in the Netherlands (at the intersection of the dialects of the southern province of North-Brabant, the south-west province of Zuid-Holland and the Veluwe region), having high frequencies of t-deletion (Goeman 1999). Both HSIC and join count analysis report an FDR-adjusted $p < .05$, while for Moran's I, the geographical association does not reach the threshold of significance.

We also present preliminary results on using HSIC as an exploratory tool on the same Twitter corpus. To focus on active users who are most likely tweeting on a personal basis, we exclude users with 1,000 or more followers and users who have fewer than 50 tweets, resulting in 8,333 users. We exclude infrequent words (used by fewer than 100 users) and very frequent words (used by 1000 users or more), resulting in a total of 5,183 candidate linguistic variables. We represent the usage of a word by each author as a binary variable, and use HSIC to compute the level of spatial dependence for each word.

The top 10 words with the highest HSIC scores are *groningen* (city), *zwolle* (city), *eindhoven* (city) *arnhem* (city), *breda* (city), *enschede* (city), *nijmegen* (city), *leiden* (city),

Linguistic variables	Description	N	Moran's I	HSIC	Mantel	Join counts
Friet / patat	french fries	842	0.0004	0.0002	0.0003	0.0003
Proficiat / gefeliciteerd	congratulations	14,474	0.0004	0.0002	0.0080	0.0003
Iedereen / een ieder	everyone	13,009	0.8542	0.0002	0.8769	0.0432
Doei / aju	bye	4,427	0.7163	0.0050	0.2570	0.3868
Efkes / eventjes	for a little while	969	0.0036	0.0002	0.0003	0.0003
Naar huis / naar huus	to home	3, 942	0.8542	0.1090	0.1245	0.9426
Niet meer / nie meer	not anymore	11,596	0.0793	0.0002	0.5590	0.0329
Of niet / of nie	or not	1,882	0.8357	0.1010	0.4191	0.9426
-oa- / -ao-	e.g., *jao* versus *joa*	754	0.0004	0.0002	0.0003	0.0003
Even weer / weer even	for a little while again	921	0.0004	0.0002	0.0003	0.0003
Have + participle	e.g., *heb gedaan* ('have done') vs. *gedaan heb* ('done have')	1,122	0.8587	0.2849	0.6668	0.0255
Be + participle	e.g., *ben geweest* ('have been') vs. *geweest ben* ('been have')	1,597	0.0793	0.2849	0.7862	0.0051
Spijkerbroek / jeans	jeans	1,170	0.7796	0.0002	0.0080	0.0003
Doei/ houdoe	bye	4,491	0.5016	0.0002	0.6668	0.0047
Bellen / telefoneren	to call by telephone	4,689	0.2730	0.0003	0.9781	0.5941

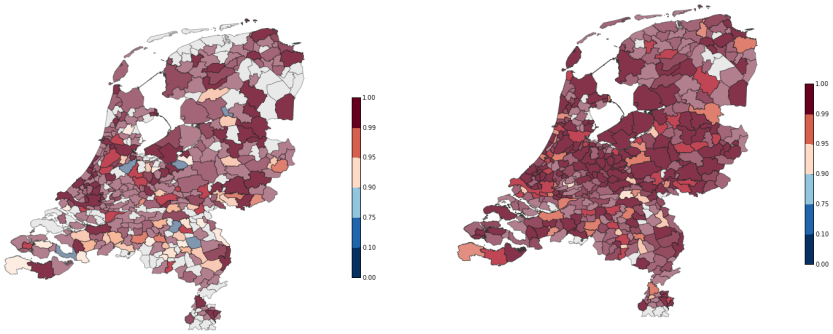
Table 2: Twitter results. The p-values were calculated using 10,000 permutations and corrected for multiple comparisons.



(a) French fries (*friet* versus *patat*), N=844

(b) For a little while (*efkes* versus *eventjes*), N=970

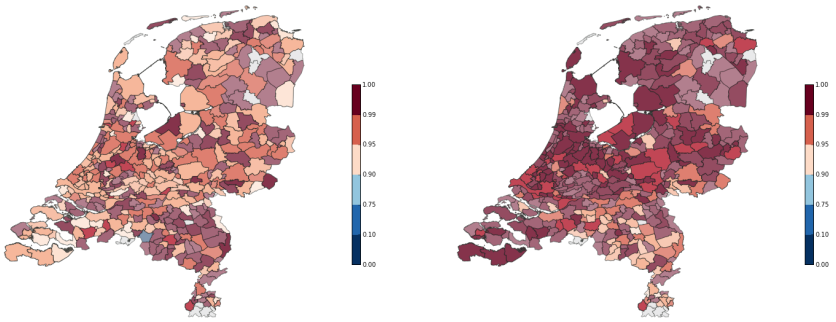
Figure 11: Highly significant linguistic variables on Twitter. Grey indicates areas with no data points. The intensity indicates the number of data points.



(a) Bye (*doei* versus *houdoe*), N=4,491

(b) t-deletion (*niet meer* vs. *nie meer*), N=11,596

Figure 12: Linguistic variables on Twitter where tests disagreed



(a) Usage of *joh* (b) Usage of *dadelijk*

Figure 13: Linguistic features on Twitter

twente (region) and *delft* (city). While these words do not reflect dialectal variation as it is normally construed, we expect their distribution to be heavily influenced by geography. Manual inspection revealed that many English words (e.g., *his, very*) have high geographical dependence. English speakers are more likely to visit tourist and commercial centers, so it is unsurprising that these words should show a strong geographical association. The top-ranked non-topical word is *proficiat*, occurring at rank 34 according to HSIC. *Proficiat* had previously been identified as a candidate dialect variable, and was included in our analysis in Table 2; this replication of prior dialectological knowledge validates the usage of HSIC as an exploratory tool. Less well known are *joh* (an interjection) and *dadelijk* ('immediately'/'just a second'), which are ranked respectively at #60 and #71 by HSIC. The geographical distributions of these words are shown in Figures 13a and 13b; both seem to distinguish the southern part of the Netherlands from the rest of the country. The identification of these words speaks to the potential of HSIC to guide the study of dialect by revealing geographically-associated terms.[9]

6. Conclusion

We have reviewed four methods for quantifying the spatial dependence of linguistic variables: Moran's I, which is perhaps the best-known in sociolinguistics and dialectology; join count analysis; the Mantel test; and the Hilbert-Schmidt Independence Criterion (HSIC), which we introduce to linguistics in this paper. Of these methods, only HSIC is consistent, meaning that it converges to an accurate measure of the statistical dependence between X and Y in the limit of sufficient data. In contrast, the other approaches are based on parametric models. When the assumptions of these models are violated, the power to detect significant geo-linguistic associations is diminished. All three of these methods can be modified to account for various geographical distributions: for example, the spatial weighting matrix employed in Moran's I and join count analysis can be constructed as a non-linear or non-monotonic function of distance (Getis and Aldstadt 2010), the distances in the Mantel test can be censored at some maximum

9 The top 10 words with the highest Moran's I scores are similar: *groningen*, *eindhoven*, *friesland* (province), *leeuwarden* (city), *zwolle*, *proficiat*, *drachten* (city), *carnaval* (a festival), *brabant* (province), *enschede* (city).

value (Legendre, Fortin, and Borcard 2015), and so on. However, such modifications require the user to have strong prior expectations of the form of the geolinguistic dependence, and open the door to p-value hacking through iterative "improvements" to the test. In contrast, HSIC can be applied directly to any geotagged corpus, with minimal tuning. By representing the underlying probability distributions in a Hilbert space, HSIC implicitly makes a comparison across high-order moments of the distributions, thus recovering evidence of probabilistic dependence without parametric assumptions.

These theoretical advantages are borne out in an analysis of synthetic data in section 4. We consider a range of realistic scenarios, finding that the power of Moran's I, the Mantel test, and join count analysis depends on the nature of the geographical variation (e.g., dialect continua versus centers), and in some cases, even on the direction of variation. Overall, we find that HSIC, while not the most powerful test in every scenario, offers the broadest applicability and the least potential for catastrophic failure of any of the proposed approaches.

We then showed how to apply these tests to a diverse range of real datasets in section 5: frequency observations in letters to the editor, binary and categorical observations in a dialect atlas, and binary observations in social media. We find that previous results on newspaper data were dependent on the procedure of selecting the geographical distance cutoff to maximize the number of positive test results; using all other test procedures, the significance of these results disappears (section 5.1). On the dialect atlas, we find that the fraction of statistically significant variables ranges from 55.2% to 78.3% depending on the statistical approach (section 5.2). On the social media data, we obtain largely similar results from the four different tests, but HSIC detects the largest number of significant associations, identifying cases in which geography and population density were closely intertwined (section 5.3).

To conclude, we believe that kernel embeddings of probability measures offer a powerful new approach for corpus analysis. In this paper, we have focused on measuring geographical dependence, which can be used to test and discover new dialectal linguistic variables. But the underlying mathematical ideas may find application in other domains, such as tracking change over time (Popescu and Strapparava 2014; Štajner and Mitkov 2011), or between groups of authors (Koppel, Argamon, and Shimoni 2002). Of particular interest for future research is the use of structured kernels, such as tree kernels (Collins and Duffy 2001) or n-gram kernels (Lampos et al. 2015), which could test for structured linguistic phenomena such as variation in syntax (Johannsen, Hovy, and Søgaard 2015) or phonological change (Bouchard-Côté et al. 2007).

Acknowledgments

Thanks to Jack Grieve for sharing the corpus of dialect variables from Letters to the Editor in North American newspapers, Arthur Gretton for advice about how best to use HSIC, Erik Tjong Kim Sang for help on using the SAND data, the DB group of the University of Twente for sharing the Dutch geotagged tweets, and Leonie Cornips and Sjef Barbiers for advice on selecting the Dutch linguistic variables. The first author was supported by the Netherlands Organization for Scientific Research (NWO), grant 640.005.002 (FACT). This material is based upon work supported by the National Science Foundation under Grant Number RI-1452443, and by a grant from the Air Force Office of Scientific Research to the second author. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Anselin, Luc. 1995. Local indicators of spatial association–LISA. *Geographical analysis*, 27(2):93–115.

Bach, Francis R. and Michael I. Jordan. 2002. Kernel independent component analysis. *The Journal of Machine Learning Research*, 3:1–48.

Barbiers, Sief, Hans Bennis, Gunther De Vogelaer, Magda Devos, Margreet van der Ham, Irene Haslinger, Marjo van Koppen, Jeroen Van Craenenbroeck, and Vicky Van den Heede. 2005. *Syntactic Atlas of the Dutch Dialects: Volume I*. Amsterdam University Press.

Barbiers, Sjef, Johan van der Auwera, Hans Bennis, Eefje Boef, Gunther De Vogelaer, and Margreet van der Ham. 2008. *Syntactic Atlas of the Dutch Dialects: Volume II*. Amsterdam University Press.

Benjamini, Yoav and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bouchard-Côté, Alexandre, Percy Liang, Thomas Griffiths, and Dan Klein. 2007. A probabilistic approach to diachronic phonology. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 887–896, Prague, Czech Republic.

Chambers, Jack K. and Peter Trudgill. 1998. *Dialectology*. Cambridge University Press.

Cliff, Andrew D. and J. Keith Ord. 1981. *Spatial processes: models & applications*, volume 44. Pion London.

Collins, Michael and Nigel Duffy. 2001. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems 14*, pages 625–632, Vancouver, British Columbia, Canada.

Cressie, Noel. 1988. Spatial prediction and ordinary kriging. *Mathematical geology*, 20(4):405–421.

Doyle, Gabriel. 2014. Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 98–106, Gothenburg, Sweden.

Ecker, Mark D. and Alan E. Gelfand. 1997. Bayesian variogram modeling for an isotropic spatial process. *Journal of Agricultural, Biological, and Environmental Statistics*, 2(4):347–369.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE*, 9(11):e113114.

Flaxman, Seth R. 2015. *Machine Learning in Space and Time*. Ph.D. thesis, Carnegie Mellon University.

Fukumizu, Kenji, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. 2007. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496, Vancouver, B.C., Canada.

Getis, Arthur and Jared Aldstadt. 2010. Constructing the spatial weights matrix using a local statistic. In *Perspectives on spatial data analysis*. Springer, pages 147–163.

Getis, Arthur and J. Keith Ord. 1992. The analysis of spatial association by use of distance statistics. *Geographical analysis*, 24(3):189–206.

Goeman, Ton. 1999. *T-deletie in Nederlandse dialecten; kwantitatieve analyse van structurele, ruimtelijke en temporele variatie*. Ph.D. thesis, Vrije Universiteit Amsterdam.

Gooskens, Charlotte and Wilbert Heeringa. 2006. The relative contribution of pronunciational, lexical, and prosodic differences to the perceived distances between Norwegian dialects. *Literary and Linguistic Computing*, 21(4):477–492.

Gretton, Arthur, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005a. Measuring statistical dependence with Hilbert-Schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Algorithmic Learning Theory*, volume 3734 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 63–77.

Gretton, Arthur, Kenji Fukumizu, Choon H. Teo, Le Song, Bernhard Schölkopf, and Alex J. Smola. 2008. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592, Vancouver, British Columbia, Canada.

Gretton, Arthur, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. 2005b. Kernel methods for

measuring independence. *Journal of Machine Learning Research*, 6:2075–2129.
Gretton, Arthur, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K. Sriperumbudur. 2012. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems 25*, pages 1205–1213, Lake Tahoe, Nevada, USA.
Grieve, Jack. 2014. A comparison of statistical methods for the aggregation of regional linguistic variation. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*. Berlin/New York: Walter de Gruyter, pages 53–88.
Grieve, Jack. 2016. *Regional Variation in Written American English*. Cambridge University Press.
Grieve, Jack, Dirk Speelman, and Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23(02):193–221.
Grieve, Jack, Dirk Speelman, and Dirk Geeraerts. 2013. A multivariate spatial analysis of vowel formants in American English. *Journal of Linguistic Geography*, 1(1):31–51.
Heeringa, Wilbert and John Nerbonne. 2001. Dialect areas and dialect continua. *Language Variation and Change*, 13(03):375–400.
Hong, Liangjie, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the Twitter stream. In *Proceedings of the 21st international conference on World Wide Web (WWW '12)*, pages 769–778, Lyon, France.
Hovy, Dirk, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*, pages 452–461, Florence, Italy.
Huang, Yuan, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255.
Johannsen, Anders, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China.
de Jong, P., C. Sprenger, and F. van Veen. 1984. On extreme values of Moran's I and Geary's c. *Geographical Analysis*, 16(1):17–24.
Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.
Lampos, Vasileios, Elad Yom-Tov, Richard Pebody, and Ingemar J. Cox. 2015. Assessing the impact of a health intervention via user-generated internet content. *Data Mining and Knowledge Discovery*, 29(5):1434–1457.
Lee, Jay and William A. Kretzschmar Jr. 1993. Spatial analysis of linguistic data with GIS functions. *International Journal of Geographical Information Science*, 7(6):541–560.
Legendre, Pierre, Marie-Josée Fortin, and Daniel Borcard. 2015. Should the Mantel test be used in spatial analysis? *Methods in Ecology and Evolution*, 6(11):1239–1247.
Lodhi, Huma, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research*, 2(Feb):419–444.
Moran, Patrick A. P. 1950. Notes on continuous stochastic phenomena. *Biometrika*, 37(1-2):17–23.
Muandet, Krikamol, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. 2016. Kernel mean embedding of distributions: A review and beyond. *arXiv preprint arXiv:1605.09522*.
Nerbonne, John and William A. Kretzschmar Jr. 2013. Dialectometry++. *Literary and Linguistic Computing*, 28(1):2–12.
Ord, Keith. 1975. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70(349):120–126.
Popescu, Octavian and Carlo Strapparava. 2014. Time corpora: Epochs, opinions and changes. *Knowledge-Based Systems*, 69:3–13.
Roller, Stephen, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510, Jeju Island, Korea.

Scherrer, Yves. 2012. Recovering dialect geography from an unaligned comparable corpus. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 63–71, Avignon, France.
Shawe-Taylor, John and Nello Cristianini. 2004. *Kernel methods for pattern analysis*. Cambridge University Press.
Song, Le, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. 2012. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13:1393–1434.
Spruit, Marco René. 2006. Measuring syntactic variation in Dutch dialects. *Literary and Linguistic Computing*, 21(4):493–506.
Štajner, Sanja and Ruslan Mitkov. 2011. Diachronic stylistic changes in British and American varieties of 20th century written English language. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage at RANLP*, pages 78–85, Hissar, Bulgaria.
Szmrecsanyi, Benedikt. 2012. *Grammatical variation in British English dialects: a study in corpus-based dialectometry*. Cambridge University Press.
Tjong Kim Sang, Erik. 2015. Discovering dialect regions in syntactic dialect data. In *Workshop European Dialect Syntax VIII - Edisyn 2015*, Zurich, Switserland.
Wing, Benjamin and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 955–964, Portland, Oregon, USA.
Wright, Timothy F. 1996. Regional dialects in the contact call of a parrot. *Proceedings of the Royal Society of London B: Biological Sciences*, 263(1372):867–872.
Yamada, Makoto, Wittawat Jitkrittum, Leonid Sigal, Eric P. Xing, and Masashi Sugiyama. 2014. High-dimensional feature selection by feature-wise kernelized lasso. *Neural computation*, 26(1):185–207.
Zhang, Qinyi, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. 2016. Large-scale kernel methods for independence testing. *arXiv preprint arXiv:1606.07892*.