Chapter 13

Assessing smoothing parameters in dialectometry

Jack Grieve
Aston University

This paper considers an approach that was suggested by John Nerbonne for assessing how to best set parameters for the smoothing of dialect maps using statistical methods. This approach involves correlating the smoothed maps for a regional linguistic variable generated under various different settings of a parameter to the underlying raw map and then graphing the results in order to estimate a reasonable value for that parameter. In order to test this method, the relative frequencies of numerous words were mapped across the counties of the contiguous United States based on an 8.9 billion word corpus of geocoded Tweets. These relative frequency maps were then smoothed using a Getis-Ord G_i^* local spatial autocorrelation analysis based on a nearest neighbor spatial weights matrix, where the number of nearest neighbors was varied from between 1 and 200 locations. The analysis suggests that setting the value of this parameter at between 25 and 50 nearest neighbors, or alternatively at approximately 10% of the total locations over which the variable was measured, generally yields acceptable results.

1 Introduction

One of the longest standing methodological problems in dialectology is how to make sense out of the complex patterns of regional variation that are generally exhibited by linguistic variables when mapped. The traditional solution to this problem has been to draw isoglosses by hand that divide the map into regions where the different values of the variable are more or less common. An alternative solution is to use statistical methods to automatically smooth dialect maps, including a Getis-Ord G_i^* local spatial autocorrelation analysis (Getis & Ord 1992; Grieve, Speelman & Geeraerts 2011; Grieve 2016).

Basically, a Getis-Ord G_i^* local spatial autocorrelation analysis is a statistical method that identifies underlying patterns of spatial clustering in the values of a quantitative variable that has been measured across a set of locations. A Getis-Ord G_i^* analysis functions by comparing the values of a variable around each location

over which it is measured. If the values of the variable tend to be relatively high, then that central location is assigned a positive z-score, whereas if those values tend to be relatively low, then that location is assigned a negative z-score. These z-scores are then mapped to identify underlying regional clusters of high and low value locations, much like drawing an isogloss.

This type of statistical approach to the analysis of dialect maps has several advantages over drawing isoglosses by hand. Most important, it allows for consistent, efficient, and replicable analyses. This is especially useful when analyzing and comparing maps for many different linguistic variables, for example when the results of dialect surveys are used to identify common patterns of regional linguistic variation. In such studies, there is a very real possibility that dialectologist bias will substantially affect the results of the analysis. On any given map, most dialectologists will usually broadly agree on the placement of isoglosses, but when this procedure is repeated for many different variables, for example as a first step toward the identification of common patterns of regional linguistic variation, small variations in how isoglosses are drawn can become amplified, possibly leading to major differences when isoglosses are aggregated, which can reflect the preconceptions of the dialectologist about where important dialect boundaries lie.

The use of statistical methods for identifying underlying regional signals in dialect maps therefore greatly limits the influence of dialectologist bias, but it does not eliminate this bias entirely. As is generally the case with all but the simplest statistical methods, including most of the methods that are commonly applied in dialectometry, there are numerous parameters that must be set by the dialectologist. Most notably, an important step in conducting a Getis-Ord G_i^* analysis is to define a spatial weights matrix, which specifies the relationship between every pair of locations over which the variable is measured. Essentially, the spatial weights matrix defines what constitutes a nearby location. For example, in the most basic type of spatial weights matrix, two locations are assigned a weight of 1 if they are considered to be nearby to each other and a weight of 0 if the are not considered to be nearby to each other. Proximity can be defined in various ways, including by the number of nearest neighbors, where for each location the n nearest neighboring locations are assigned a weight of 1 and all other locations are assigned a weight of 0. How one chooses to set the number of nearest neighbors is an important decision that affects the smoothness of the resultant maps. Specifically, the smaller the number of nearest neighbors taken into consideration, the more similar the resultant map will be to the original map. Of course, the goal of applying a local spatial autocorrelation analysis in the first place is to smooth the map and so it is always necessary to this set this parameter to a value larger than 1, but otherwise setting this value is an important and often challenging decision (Getis 2009), where dialectologist bias can enter into the analysis. Perhaps most important, if these parameters are set too liberally, over-smoothing can result, where the smoothed map no longer accurately reflects the underlying regional pattern visible in the map for the variable under analysis.

Fortunately, it is possible to record and scrutinize the effects of these decisions, which is impossible when isoglosses are drawn by hand. Furthermore, given that a

local spatial autocorrelation analysis produces quantitative results, it is also possible to compare the smoothed maps to each other and to the original raw map in order to assess the degree of smoothing, and in particular to consider whether the maps have been over-smoothed. This paper considers one such approach to assessing smoothing parameters, which was suggested to the author by John Nerbonne at the 2014 Methods in Dialectology conference that he hosted in Groningen. Specifically, the suggestion was to assess the degree of smoothing of dialect maps by measuring the correlation between the raw map and the smoothed maps generated using different parameter settings.

2 Analysis

The corpus analyzed in this study consists of 8.9 billion words of geo-coded American mobile Twitter data, totaling 980 million tweets written by 7 million users from across the contiguous United States, downloaded between October 11th, 2013 and November 22nd, 2014 using the Twitter API (see Huang et al. 2016; Grieve, Nini & Guo 2016). To analyze patterns of regional linguistic variation in this variety of language, the corpus was geographically stratified by county using the longitude and latitude provided with each Tweet. In total the corpus contains 3,075 county equivalents out of a total of 3,108 county equivalents in the contiguous United States. On average, the corpus contains 2 million words per county, but the number of words per county ranges from 300 to 300 million words. Overall, 98% of the counties are represented by at least 10,000 words and 79% of the counties are represented by at least 100,000 words. Twitter provides a uniquely large and accessible source of geo-coded natural language data, which is also a highly informal variety of language that is participated in by millions of people from across the United States, making it a valuable source of data for dialectologists.

To test the effect of varying the number of nearest neighbors used to define the spatial weights matrix for a Getis-Ord G_i^* analysis of dialect maps, the relative frequencies of a series of words were measured across the counties in the corpus. This is not the type of linguistic variable commonly analyzed in dialectology, where lexical variation tends to be measured as alternations between equivalent forms (e.g. *pail* vs. *bucket*, *pop* vs. *soda* vs. *coke*); nevertheless, word frequencies still do generally show regional patterns and are therefore as suitable as any other type of linguistic variable for testing the effect of varying smoothing parameter settings. For example, the relative frequency map for the word *love*–the most common content word in the corpus–is presented in the first cell of Figure 1, showing that the usage of this word is relatively more common in the Upper South. Similarly, the relative frequency map for the word *know*–the second most common content word in the corpus–is presented in the first cell of Figure 2, showing that the usage of this word is relatively more common in the Deep South.

Next, smoothed maps for each of these words were generated using a Getis-Ord G_i^* analysis based on a series of 200 different nearest neighbors spatial weights matrices defined for between 1 and 200 nearest neighbors. As discussed above, each of these

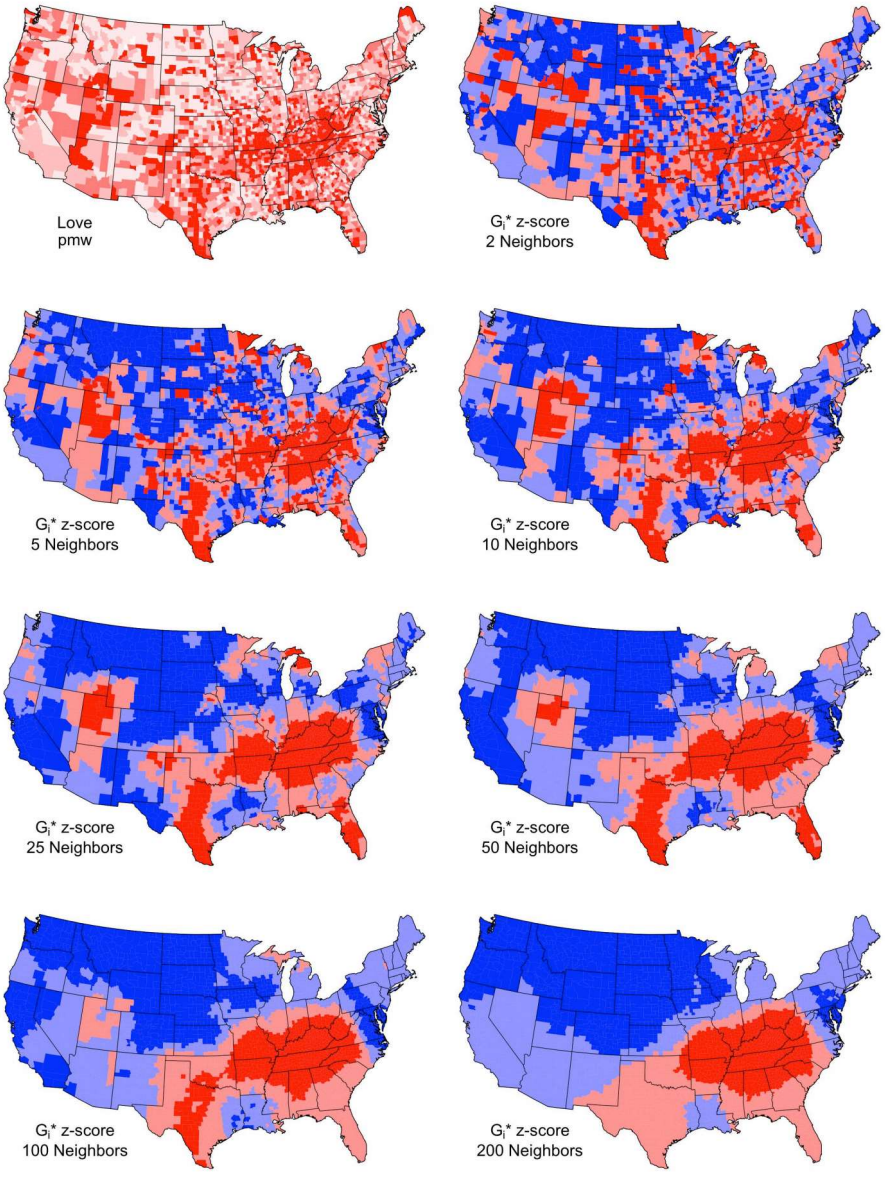

Figure 1: Relative frequency and local spatial autocorrelation maps for *love*.

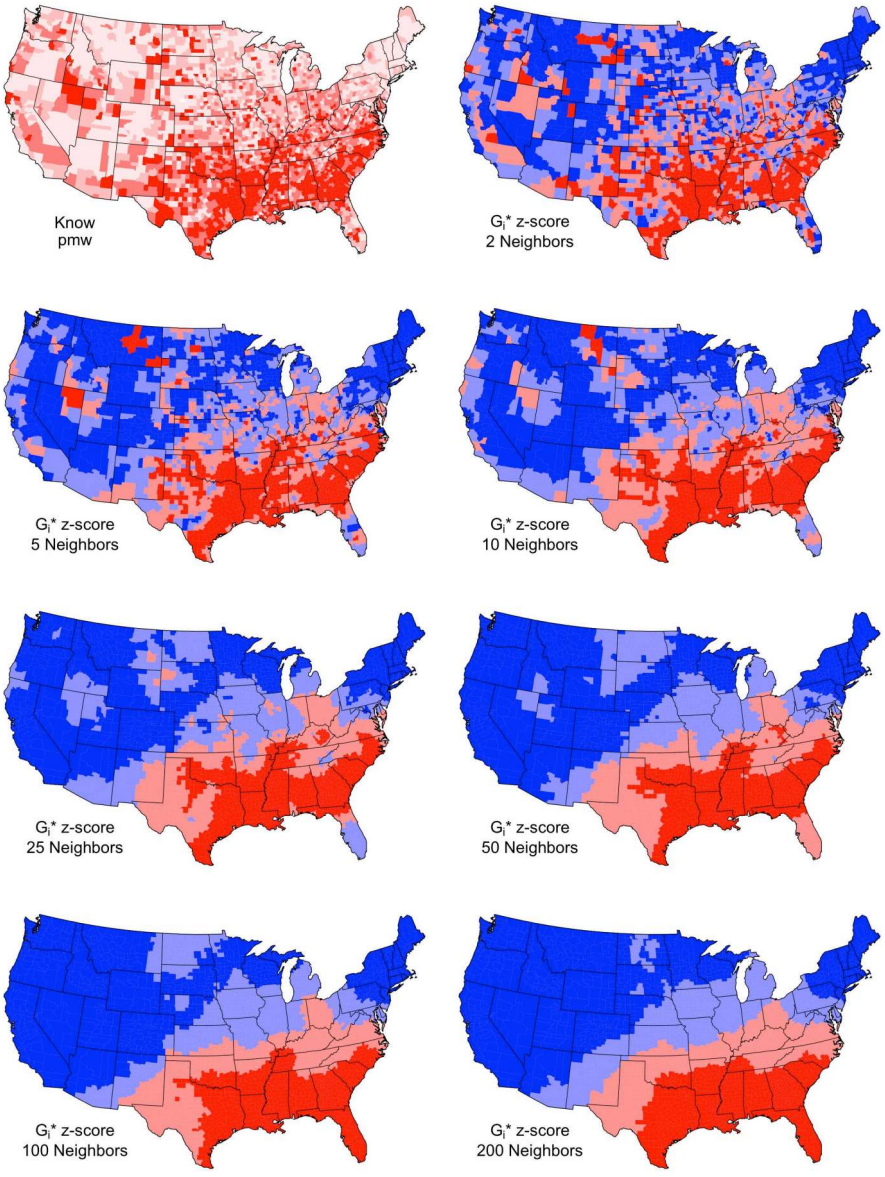

Figure 2: Relative frequency and local spatial autocorrelation maps for *know*.

analyses generates a *z*-score for each location over which the variable is measured, which can then be mapped to visualize the patterns of spatial clustering identified by the analysis. For example, the remaining cells in Figure 1 show the smoothed maps for *love* generated based on 2, 5, 10, 25, 50, 100, and 200 nearest neighbors spatial weights matrices, while the remaining cells in Figure 2 show the smoothed maps for *know* for these same parameter settings. When the number of nearest neighbors is set very low, the smoothed maps basically reproduce the raw data, whereas when the number of nearest neighbors is set very high, there is clearly over-smoothing present. For example, looking at the final smoothed map in the series for *love* (i.e. for 200 nearest neighbors), all of Utah is incorrectly identified as being a relatively low value region. Similarly, looking at the final smoothed map in the series for *love*, all of Florida is incorrectly identified as being a relatively high value region. It is therefore necessary to set the number of nearest neighbors somewhere between these two extremes in order to produce usefully smoothed maps that still accurately reflect the underlying patterns present in the raw data. In particular, looking at both sets of local spatial autocorrelation maps reproduced in Figures 1 and 2, it would appear that local spatial autocorrelation analysis based on between 25 and 50 nearest neighbors is ideal as values in this range strike a balance between over- and under-smoothing.

To assess how similar the local spatial autocorrelation maps are to the raw relative frequency maps upon which they are based, each local spatial autocorrelation map (i.e. the Getis-Ord G_i^* *z*-scores measured over the 3,075 counties) were correlated to the corresponding raw maps (i.e. the relative frequencies measured over the 3,075 counties), following the suggestion made by John Nerbonne. The resulting Pearson correlation coefficients were then plotted against the number of nearest neighbors. The resulting graph for *love* is presented in the first cell of Figure 3 and the resulting graph for *know* is presented in the second cell of Figure 3. As one would expect, both graphs show that as the number of nearest neighbors increases the correlation between the raw map and the smoothed maps decreases, although overall the strength of the correlation remains substantial. The decrease, however, is not linear. Rather, the decrease starts off very steep and then gradually flattens. Furthermore, there notably appears to be an inflection point in both graphs between 25 and 50 nearest neighbors, which corresponds to the impressionistic analysis of the smoothed maps described above, where it was argued that conducting the local spatial autocorrelation analysis on spatial weights matrix based on between 25 and 50 nearest neighbors was best.

The other cells in Figure 3 present the results of the same analysis repeated for several other words, which were selected to represent a range of word frequencies and degrees of spatial clustering. Remarkably, all these graphs show very similar patterns, with inflections points falling in the same range, i.e. between 25 and 50 nearest neighbors, suggesting that this value represents a consistently applicable parameter setting for the smoothing of maps based on this dataset. This conclusion is supported by an analysis of the maps for these variables (not shown), which exhibit similar results to the smoothed maps presented for *love* and *know* in Figures 1 and 2—clearly exhibiting over-smoothing at higher parameter settings. Given that there are 3,075

locations in this dataset, it would therefore appear that using a number equal to approximately 10% of the total locations to set the spatial weights matrix is in general a reasonable value for this parameter.

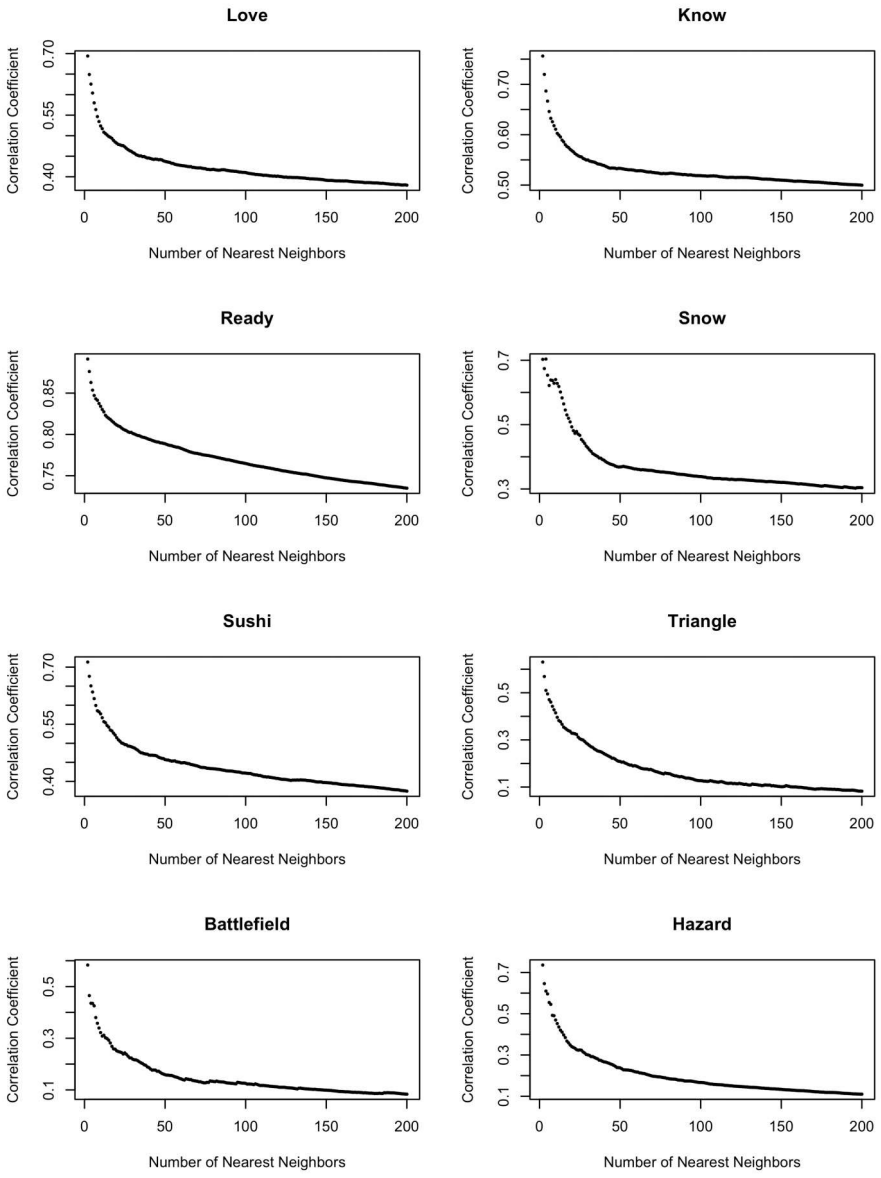

Figure 3: Nearest neighbor vs. correlation coefficient graphs.

3 Conclusion

This paper has briefly explored a general method for setting smoothing parameters for the analysis of individual patterns of regional linguistic variation in dialect maps, which was conceived by John Nerbonne. This method involves correlating maps smoothed using different parameter settings to the underlying raw maps and graphing these results. By inspecting the resultant graph, an approximate point of inflection is estimated and the smoothing parameter under consideration is then set to this value. In particular, this method was used in this paper to assess the number of nearest neighbors used to generate a nearest neighbor spatial weights matrix, an important step in conducting a Getis-Ord G_i^* local spatial autocorrelation analysis, which is an increasingly common method for smoothing dialect maps in dialectometry. Based on this approach, this study found evidence suggesting that using a number of nearest neighbors equal to approximately 10% of the total number of locations under analysis is a reasonable way to set this parameter for a Getis-Ord G_i^* analysis—generating usefully smoothed maps, while guarding against over-smoothing. Considerably more analysis both within this dataset and across other dialect datasets, however, is necessary to fully support this claim. In addition, it is important to test the applicability of this approach for setting the parameters associated with other types of spatial weights matrices as well as to test the applicability of this approach more generally for setting the parameters associated with other methods for smoothing used in dialectometry. Nevertheless, this relatively simple method appears to be a promising approach for helping to resolve an important modern methodological problem in dialectometry.

References

Getis, Arthur. 2009. Spatial weights matrices. *Geographical Analysis* 41. 404–410.

Getis, Arthur & J. K. Ord. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24. 189–206.

Grieve, Jack. 2016. *Regional variation in written American English.* Cambridge University Press.

Grieve, Jack, Andrea Nini & Diansheng Guo. 2016. Analyzing lexical emergence in American English online. *English Language and Linguistics.* To appear.

Grieve, Jack, Dirk Speelman & Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23. 193–221.

Huang, Yuan, Diansheng Guo, Alice Kasakoff & Jack Grieve. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems.* To appear.