
Jack Grieve
University of Leuven
jack.grieve@arts.kuleuven.be

The use of spatial autocorrelation statistics for the analysis of regional linguistic variation

Introduction

In geolinguistics, there is no standard method for testing if an individual linguistic variable, measured across a set of locations, exhibits a statistically significant regional pattern. In traditional dialectology regional patterns in the values of individual linguistic variables are identified through subjective analyses (Kurath 1949; Labov et al. 2006), whereas in modern dialectometry regional patterns in the values of individual linguistic variables are generally ignored (Séguy 1973; Goebel 2006; Nerbonne 2006), as the focus of this field is aggregated regional linguistic variation. This paper introduces two measures of spatial autocorrelation and demonstrates how these statistics can be used to identify significant patterns of regional linguistic variation in the values of individual linguistic variables.

Spatial autocorrelation

Spatial autocorrelation (Odland 1988; Lee & Kretzschmar 1993) is a measure of spatial dependency that quantifies the degree of spatial clustering or dispersion in the values of a variable measured across a set of locations. There are two basic types of spatial autocorrelation statistics: global measures identify whether the values of a variable exhibit a significant overall pattern of regional clustering, whereas local measures identify the location of significant high and low value clusters.

In order to determine if the values of a linguistic variable exhibit significant spatial clustering across a set of locations, global Moran's I (Moran 1948; Odland 1988) can be used to test for significant levels of positive global spatial autocorrelation. The value of Moran's I ranges from -1 to +1, where a significant negative value indicates that nearby locations tend to have different values (i.e. spatial dispersion), an

insignificant value indicates that nearby locations tend to have random values, and a significant positive value indicates that nearby locations tend to have similar values (i.e. spatial clustering). By testing for significant levels of positive global spatial autocorrelation, it is therefore possible to statistically identify the presence of regional clustering in the values of individual linguistic variables.

In order to determine the location of high and low value clusters in the values of a linguistic variable, local Getis-Ord G_i^* (Ord & Getis 1995) can be used to test each variable for local spatial autocorrelation. Unlike measures of global spatial autocorrelation, which return *one value for each variable* indicating the degree of regional clustering across the *entire* distribution of that variable, measures of local spatial autocorrelation return *one value for each location* for each variable indicating the degree to which that particular location is part of a high or low value cluster. The results of a local spatial autocorrelation analysis can then be mapped across the set of locations in order to identify the position of high and low value clusters.

Application

In order to demonstrate the application of these two measures of spatial autocorrelation, a regional analysis of continuous grammatical, lexical and phonetic variation will be presented based on two American datasets: the phonetic data gathered for the Atlas of North American English (Labov et al. 2006), and lexical-grammatical data extracted from a 25 million word corpus of letters to the editor (Grieve 2009). It will be argued that spatial autocorrelation statistics are particularly useful for identifying regional patterns in the type of geolinguistic data that tends to be obtained when continuously measured linguistic variables are analyzed, which often does not produce the types of clear cut regional patterns that are common in traditional categorical analyses.

The utility of the spatial autocorrelation statistics is demonstrated in the two figures reproduced below.

FIGURE 1. *Do Not* CONTRACTION RATE.

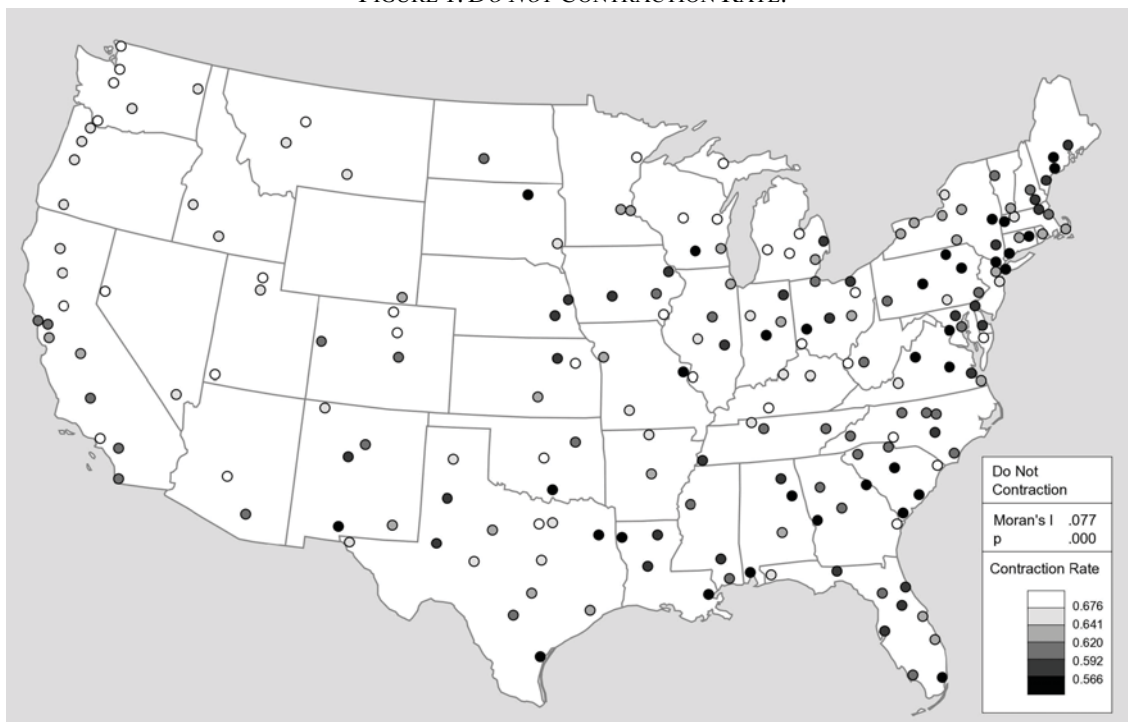
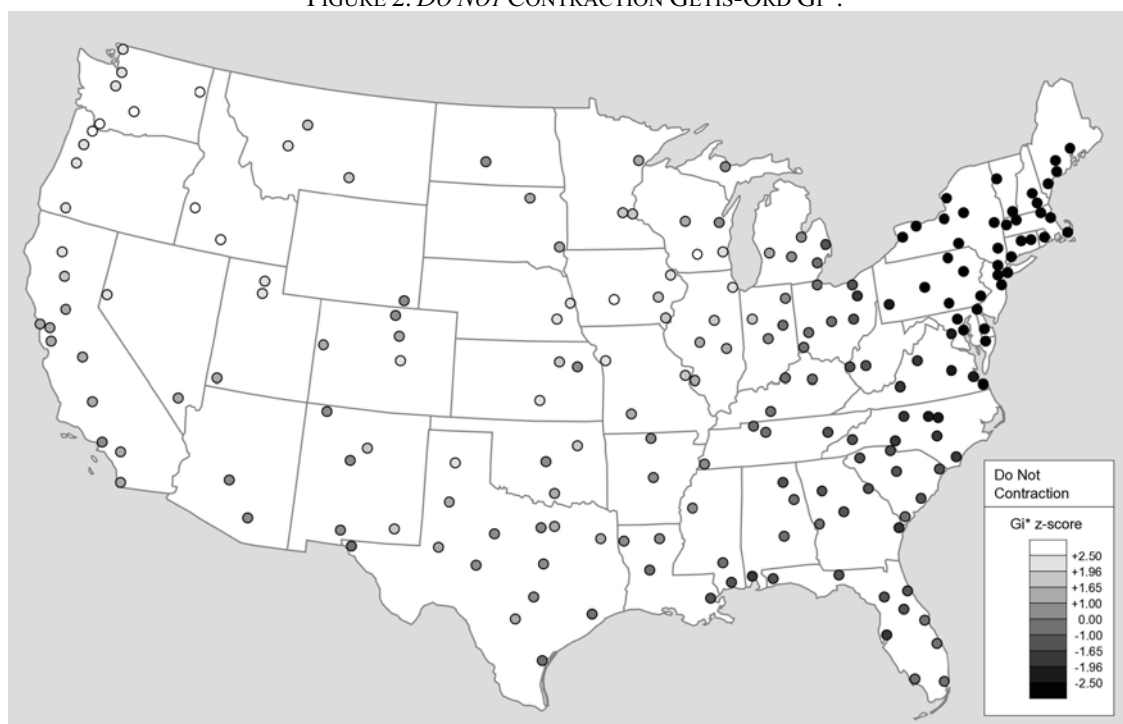


Figure 1 plots *do not* contraction rate (e.g. *doesn't* vs. *does not*) across the 206 cities represented in the letter to the editor corpus. Despite the fact that there is no clear regional pattern in the raw values of this variable, at least by traditional standards, the results of the Moran's I analysis ($I = .077$, $p < .0001$) show that the variable actually exhibits highly significant spatial clustering. The location of these clusters can then be determined by conducting an analysis of local spatial autocorrelation. The results of the Getis-Ord G_i^* analysis are presented in Figure 2, which plots the Getis-Ord G_i^* z-scores for *do not* contraction rate across the 206 cities, showing quite clearly that *do not* contraction is relatively more common in the Northwest and Western Midwest, and relatively less common in the Northeast.

FIGURE 2. *DO NOT* CONTRACTION GETIS-ORD G_i^* .



References

- Goebel, H. 2006. Recent Advances in Salzburg Dialectometry. *Literary and Linguistic Computing* 21, 411-435.
- Grieve, J. 2009. *A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English*. Ph.D. Dissertation. Northern Arizona University
- Kurath, H. 1949. *Word Geography of the Eastern United States*. University of Michigan Press.
- Labov, W., Ash, S. and Boberg, C. 2006. *Atlas of North American English: Phonetics, Phonology, and Sound Change*. New York: Mouton de Gruyter.
- Lee, J. and Kretzschmar, W. 1993. Spatial Analysis of Linguistic Data with GIS Functions. *International Journal of Geographical Information Systems* 7, 541-60.
- Moran, P.A.P. 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B* 37, 243-251.
- Nerbonne, J. 2006. Identifying Linguistic Structures in Aggregate Comparison. *Literary and Linguistic Computing* 21, 463-475.
- Odland, J.D. 1988. *Spatial Autocorrelation*. Sage Publications.
- Ord, J.K. and Getis, A. 1995. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis* 27, 286-306.

Séguy, J. 1973. La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane*, 37, 1-24.
