Global Autocorrelation and Dialect Studies: The Role of Significance

Costanza Asnaghi
Università Cattolica del Sacro Cuore and QLVL, KU Leuven

Research Question

Is it sensible to include or better to exclude variables that do not exhibit statistically significant patterns in dialect studies?

Dataset

The argumentation is based on the results of a previous regional lexical variation survey of written Standard California English, which examined 45 continuous lexical alternation variables in 334 online newspapers across 273 California locations [1]. The frequencies of the variables were gathered through site-restricted web searches [3].

Methodology

Language in neighboring locations is likely to be more similar than language in locations which are far apart: with a reciprocal weighting function, the results are highlighted based on the distances between cities. Moran's *I* [5] is used to analyze global spatial autocorrelation, and Getis-Ord *Gi* [2, 6] is used to analyze local spatial autocorrelation.

Results

The list of 45 variables is ranked according to their significance score (*p*-value). 30 out of 45 variables display significant patterns of autocorrelation on the global level.

Multivariate Analysis

Factor analysis is calculated for the autocorrelated values in two ways:
- on the 30 significant variables only (Fig. 1-3);
- on the complete set of 45 variables (Fig. 4-6).

Results Comparison

The maps from the two sets of variables reasonably align, after swapping factor 3 with factor 1. It should be noted also that Fig. 3 and Fig. 6 are comparable if one reverses the color correspondence (red shades in Fig. 3 correspond to blue shades in Fig. 6).
A more cohesive pattern is detected in Fig. 4-6 for the complete set of variables rather than in Fig. 1-3 for the significant variables only:
- Fig 1 and Fig. 4: main differences in the Los Angeles Area and in Northern California (east and south of the San Francisco Bay Area);
- Fig 2 and Fig. 5: main difference in the central coastal locations;
- Fig 3 and Fig. 6: main differences in the upper part of Northern California and in the lower part of Southern California.

Why is the Complete Set Better?

Although a more cohesive pattern is not necessarily sound proof of a more accurate representation of language usage, the choice for the all variable representation is more reasonable from a technical point of view: including more variables allows for a bigger matrix and therefore an ampler variance. Variance provides information, and the more variable one feeds the model with, the more the model is informative.
Even if a variable does not return a significant Moran's *I*, it can still show clear patterns of local spatial autocorrelation. As a general rule, the stronger the Moran's *I* score is, the stronger the local autocorrelation clusters are; nonetheless, the elimination of lower Moran's *I* variables leads to an exclusion of important patterns that the local autocorrelation analysis identified. In fact, not significant Moran's *I* variable patterns side with other significant variables, strengthening important information.

North/South Distinction

From Fig. 1 and 4 a language usage distinction between north and south emerges.
A historical motivation underlies this distinction: in the mid-nineteenth century Northern California was growing rapidly as a consequence of the Gold Rush, while Southern California continued to be a pastoral Hispanic region until the 1880s, when, with the development of irrigation and the aqueduct system, the Imperial Valley saw the increase of the farm population. The year-round favorable weather and the relaxed lifestyle drew people to Southern California, incrementing the real estate industry [4, 7]. While San Francisco was the most populated city in the state until 1880, Los Angeles grew considerably in the following years and became three times as big as San Francisco in 1950. Therefore, the residents of Northern California have been settled for longer than the residents of Southern California, thus resulting in a different use of the language on historical basis.

Rural/Urban Distinction

From Factor 2 map (Fig. 2, 5) a language distinction between the rural and the urban areas of California emerges.
The reasons for this distinction is that the socio-economical structure of a metropolis influences people's lifestyles in a different way if compared to what has an impact on people's habits in the agricultural cities. Thus, also the language used in the urban regions differs from the language used in the rural regions.
Notably, although California encompasses five urban agglomerations (Greater Los Angeles, the San Francisco Bay, San Diego-Tijuana, Greater Sacramento, and Metropolitan Fresno), factor 2 maps indicate that only two of those agglomerations are involved in a different use of the language compared to the rest of the state (i.e. Greater Los Angeles and the San Francisco Bay). Greater Los Angeles and the San Francisco Bay are in fact different from the rest of the California metropolitan areas from a variety of points of view, e.g. the gross domestic product (GDP) of Greater Los Angeles and the San Francisco Bay is much higher than the GDP of any other California urban agglomerations.

Socio-Economic Rationale

A comparison between Fig. 3 and 6 shows that in the first case (Fig. 3) the language choices made in the Greater Los Angeles Area are somewhat comparable to those made in the San Diego Area, while in the second case (Fig. 6) Los Angeles is distinct from San Diego.
On one hand, San Diego is at the head of one of the four metropolitan areas of California. On the other hand, this metropolis preserves a provincial and conservative character. San Diego is heavily populated with military retirees, and it is more conservative than the nearby Los Angeles in its outlook. Moreover, metropolitan San Diego sits on top of the Mexican border. As a consequence, the Latino culture is stronger than anywhere else in the state. The discontinuity and dissemination of the urban spatial design makes it de facto a more provincial town, characterized by a beach culture. The language spoken in San Diego reflects these characteristics and is therefore more rural than the language in Los Angeles. On this socio-economic basis, Fig. 6 seems more accurate than Fig. 3, providing a sociolinguistic rationale as a further criterion for the selection of the all-variable model.

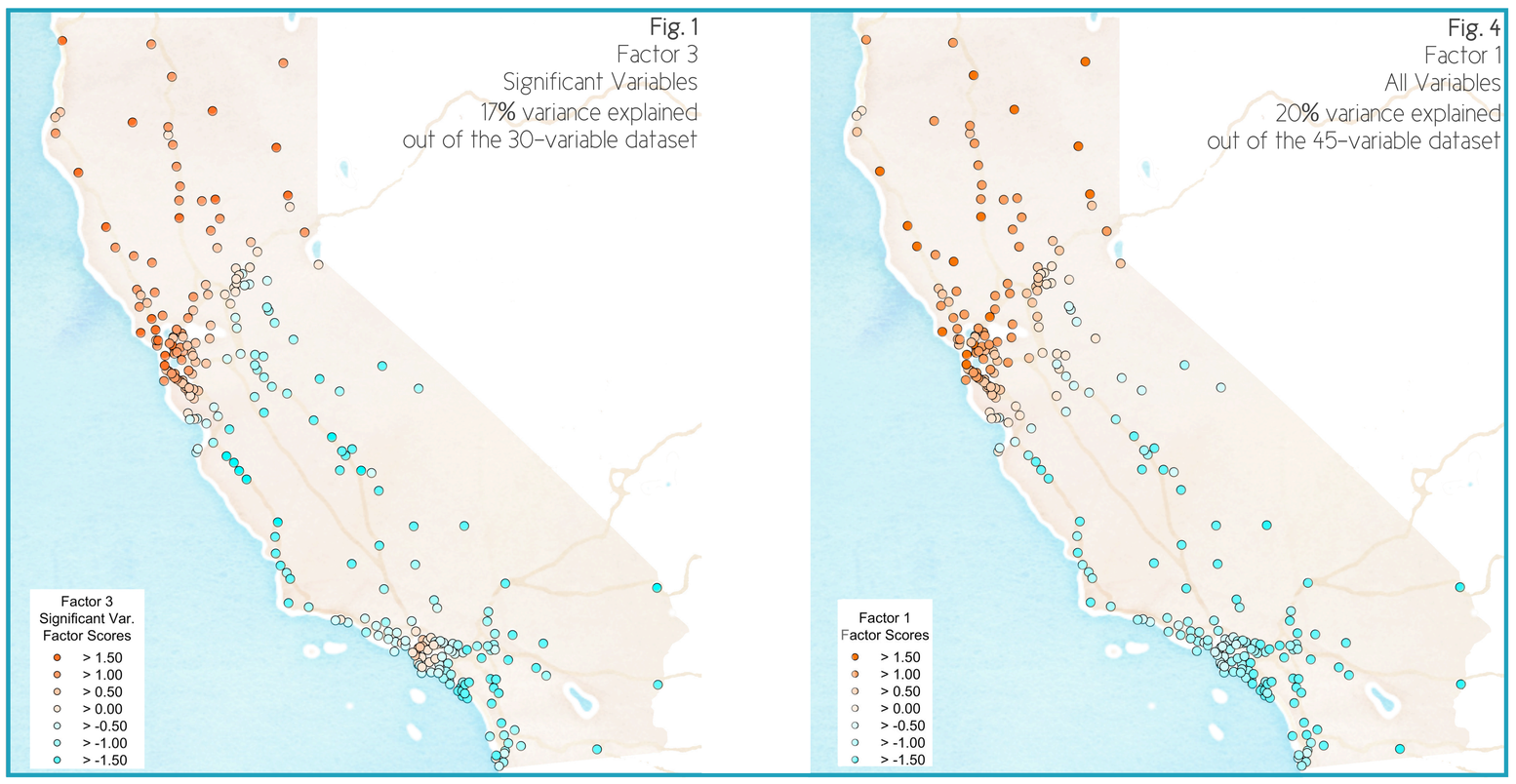

Fig. 1
Factor 3
Significant Variables
17% variance explained
out of the 30-variable dataset

Fig. 4
Factor 1
All Variables
20% variance explained
out of the 45-variable dataset

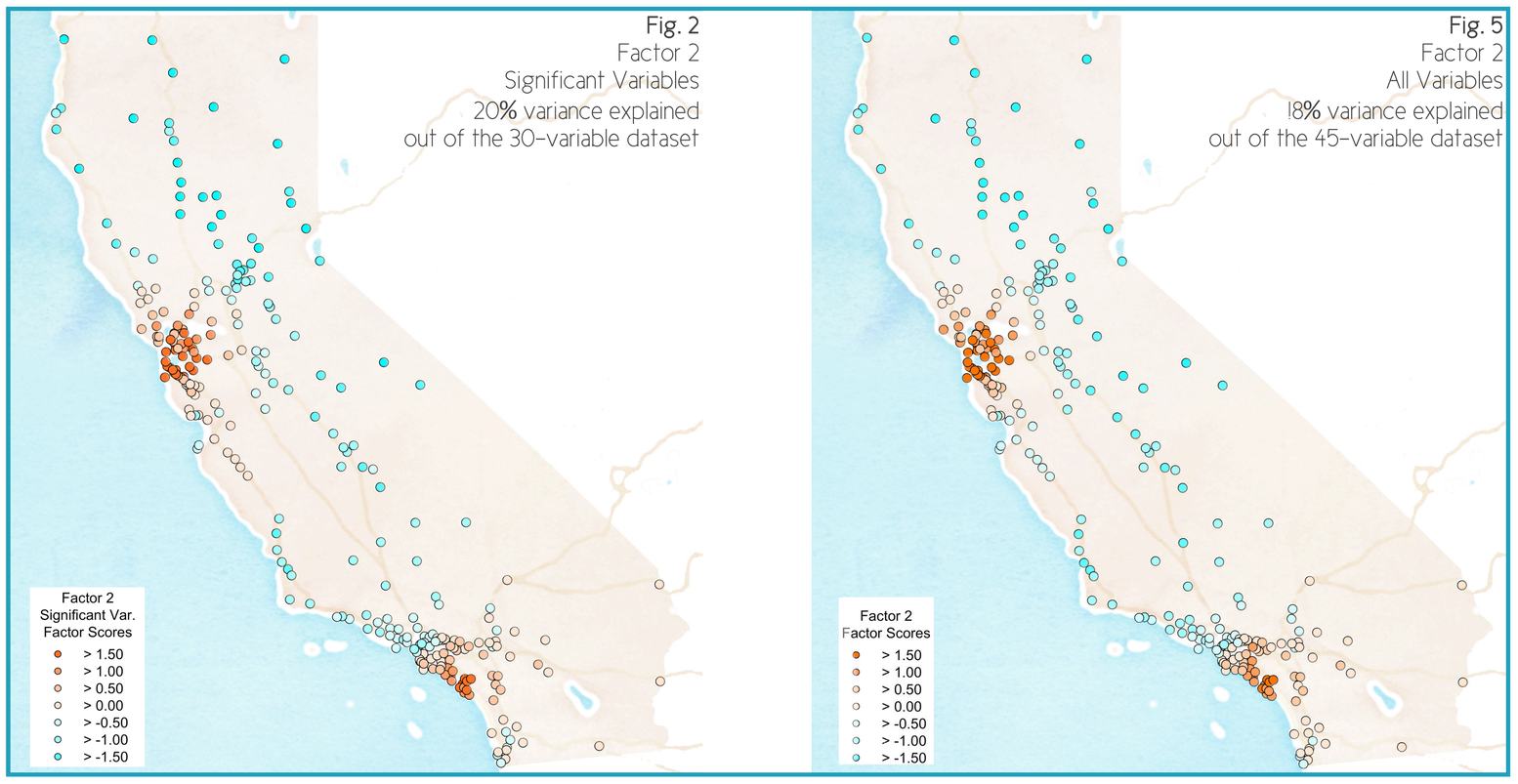

Fig. 2
Factor 2
Significant Variables
20% variance explained
out of the 30-variable dataset

Fig. 5
Factor 2
All Variables
18% variance explained
out of the 45-variable dataset

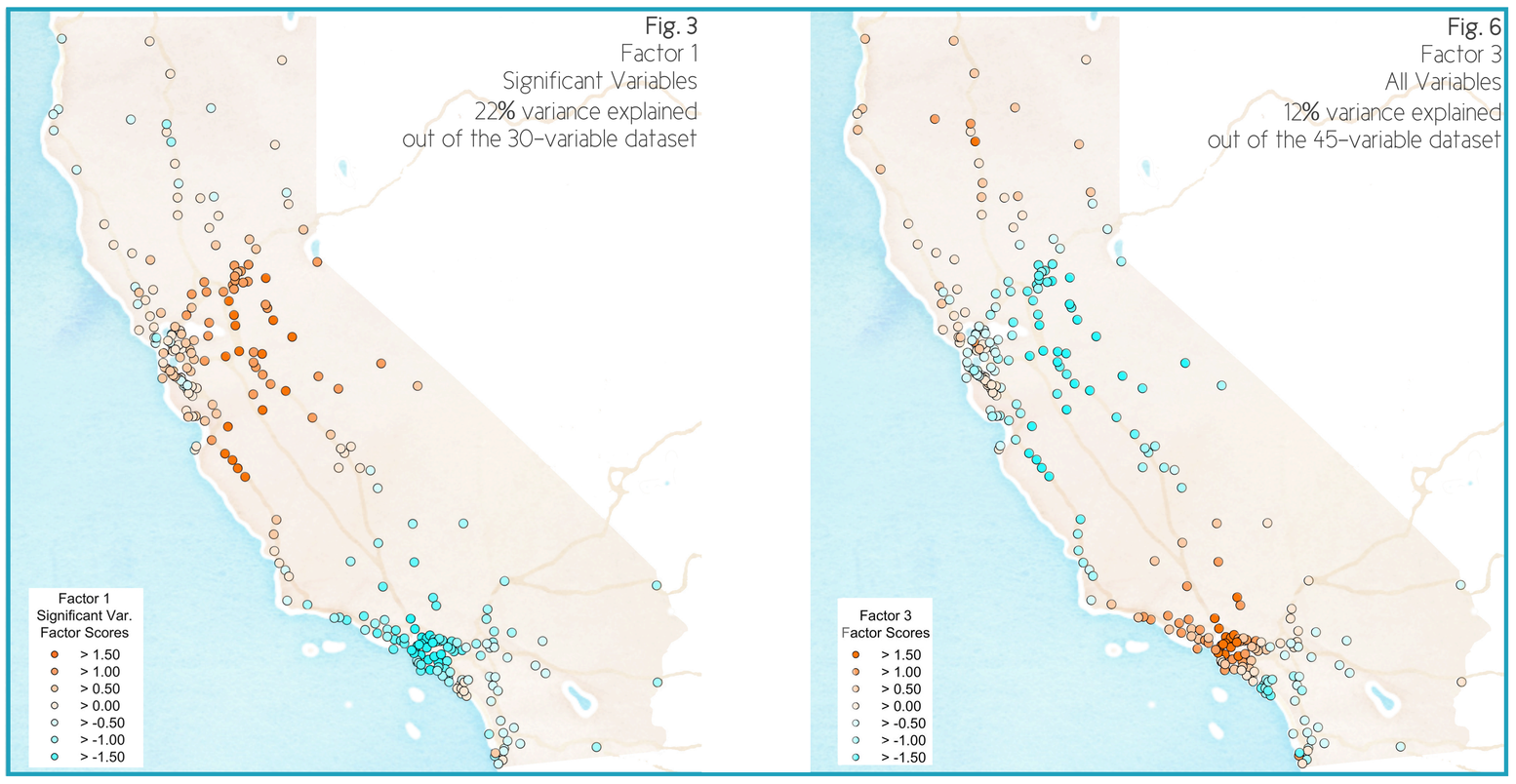

Fig. 3
Factor 1
Significant Variables
22% variance explained
out of the 30-variable dataset

Fig. 6
Factor 3
All Variables
12% variance explained
out of the 45-variable dataset

References

1. Asnaghi, C. 2013. An Analysis of Regional Lexical Variation in California English Using Site-Restricted Web Searches. Ph.D. thesis, Università Cattolica del Sacro Cuore and KU Leuven.
2. Grieve, J. 2011. A Regional Analysis of Contraction Rate in Written Standard American English. International Journal of Corpus Linguistics 16 (4): 514-546.
3. Grieve, J., C. Asnaghi, and T. Ruette. Submitted. Site-Restricted Web Searches for Data Collection in Regional Dialectology.
4. Hayes, D. and U. of California Press (2007). *Historical Atlas of California: With Original Maps*. University of California Press.
5. Moran, P. 1948. The Interpretation of Statistical Maps. Journal of the Royal Statistical Society. Series B (Methodological) 10 (2): 243-251.
6. Ord, J.K., and A. Getis 1995. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geographical Analysis 27 (4): 286-306.
7. Starr, K. and B. Procter (1973). Americans and the California Dream, 1850-1915. *History: Reviews of New Books* 1(9), 201-201.

Costanza Asnaghi, costanza.asnaghi@unicatt.it
Department of Linguistic Sciences and Foreign Literatures, Università Cattolica del Sacro Cuore, Milano
Quantitative Lexicology and Variational Linguistics, Faculty of Arts, KU Leuven