@ifisc_mallorca www.facebook.com/ifisc http://ifisc.uib-csic.es

American cultural regions mapped through the linguistic analysis of social media

Thomas Louf[1], José J. Ramasco[1], David Sánchez[1], Bruno Gonçalves[2] and Jack Grieve[3]

[1] IFISC (CSIC-UIB) Palma de Mallorca – Spain. [2] Data For Science, Inc., New York, USA. [3] University of Birmingham, UK.

thomaslouf@ifisc.uib-csic.es

Introduction

Various theories of American cultural regions have been proposed, based on many different factors, including patterns in settlement, ethnicity, religion, politics, and economics. An accurate picture of American cultural regions is important because it can help explain and predict national trends in a range of human behaviors, but it is difficult to choose between existing theories because they have been based primarily on the judgment of geographers. Quantitative data from the Census and elections have sometimes been taken into consideration, but the selection and weighting of these factors has always been subjective. Assuming, however, that cultural regions are generally reflected in all forms of cultural expression, then the analysis of regional patterns in topics of discussion should provide a more objective basis for their mapping.

Aim: Automatically detect cultural regions and identify what topics define them.
How: Data-driven approach with geo-tagged tweets, methods of dimensionality reduction and data clustering.

Measuring regional variations

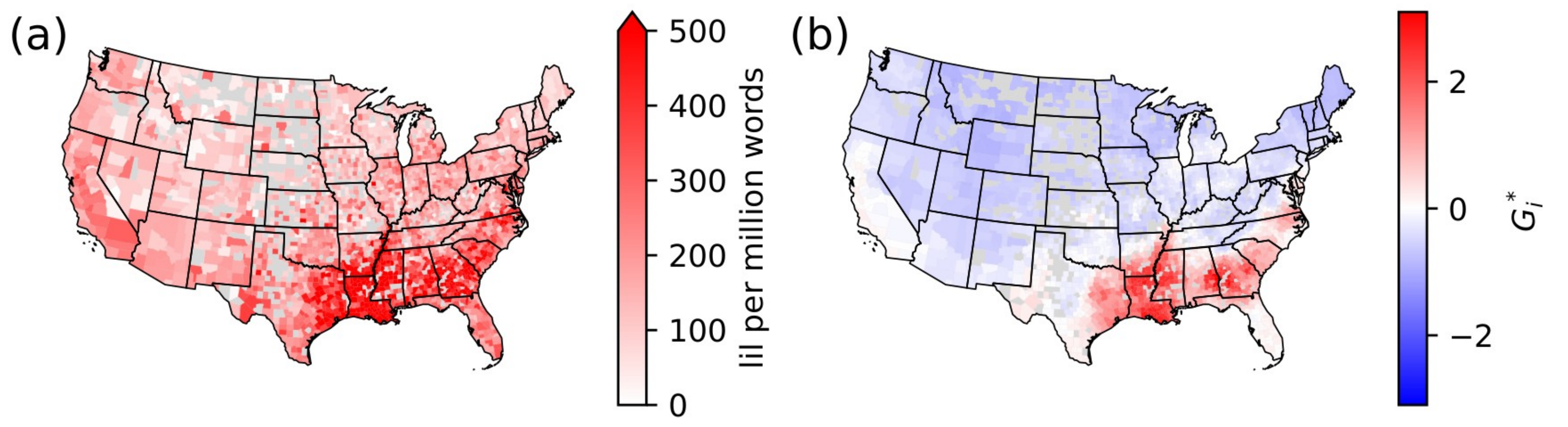
Geo-tagged tweets from 2015 to 2021 are filtered and matched to a county → frequency of word w by county c, $f_{c,w}$, for the top 10,000 words.

Getis-Ord G_i^* measure of local spatial autocorrelation:

$$G^*_{c,w} = \frac{\sum_{c'} w_{c,c'}(f_{c',w} - \bar{f}_w)}{\sigma_w \sqrt{\frac{n_c \sum_{c'} w^2_{c,c'} - \left(\sum_{c'} w_{c,c'}\right)^2}{n_c - 1}}},$$

With σ_w the variance of w's frequencies, n_c the number of counties and $w_{c,c'}$ the spatial weights (=1 for 10 nearest neighbors of c).

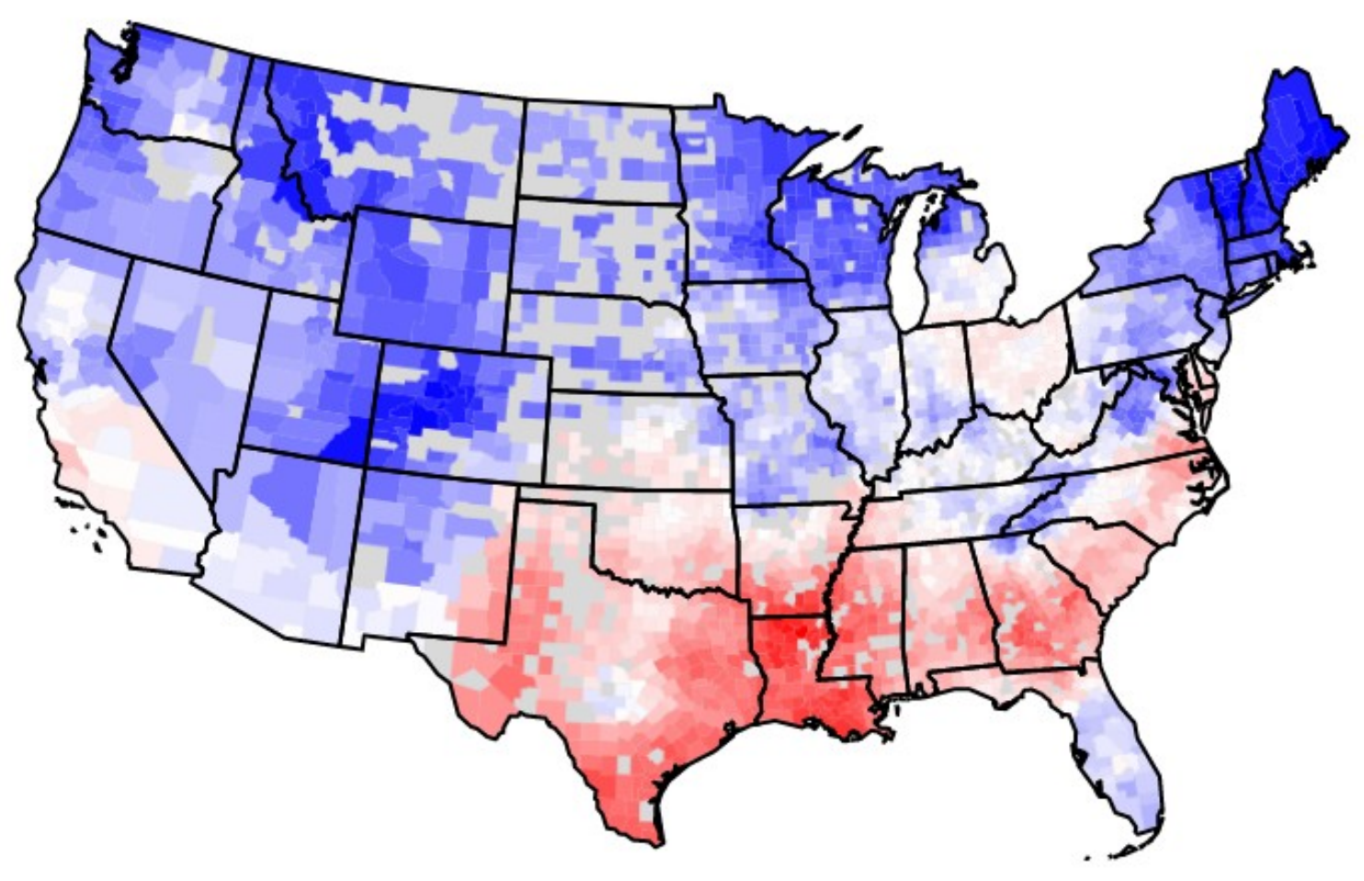
Smooths fluctuations and highlights word hotspots: where over and under-used:

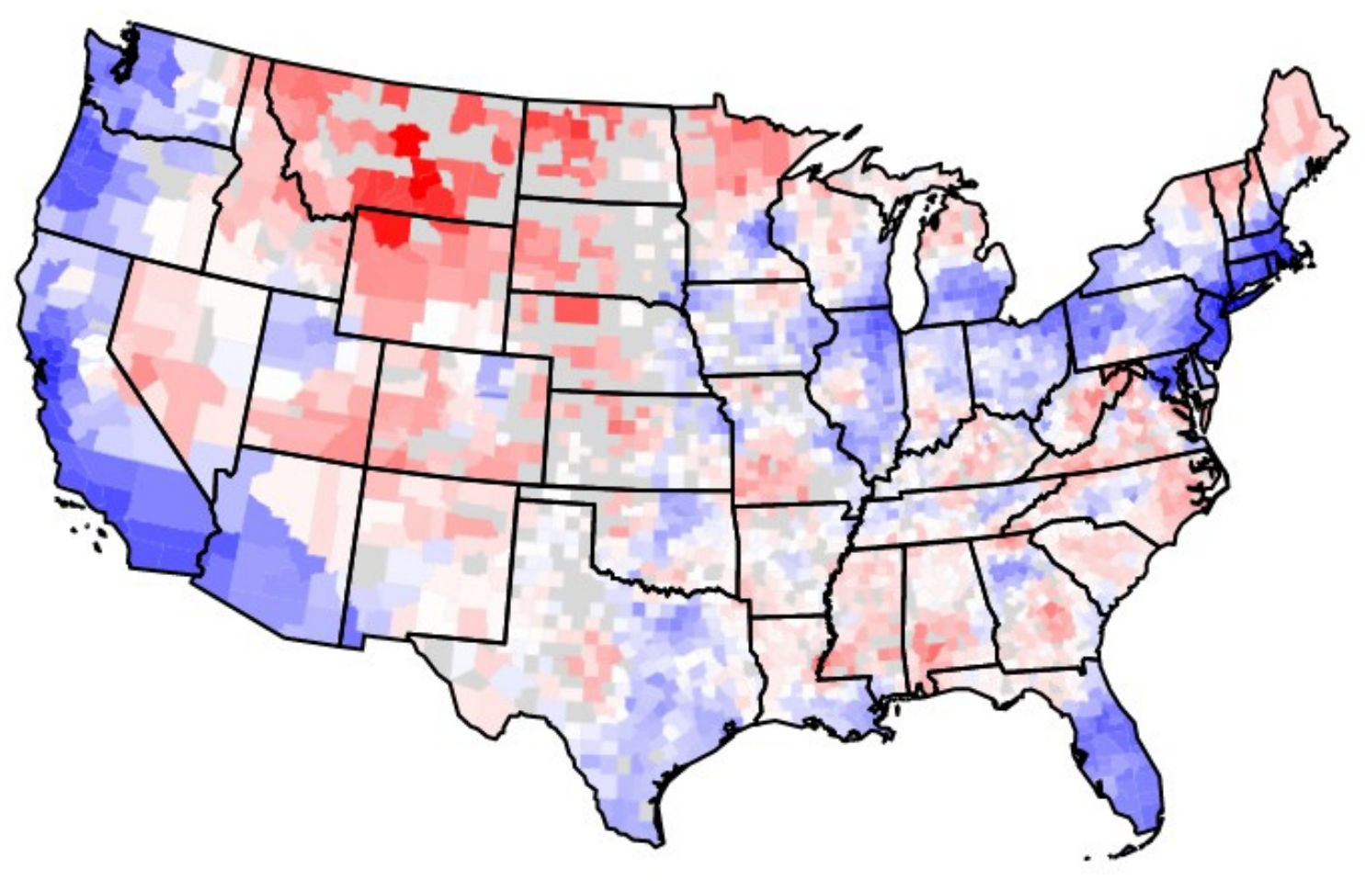


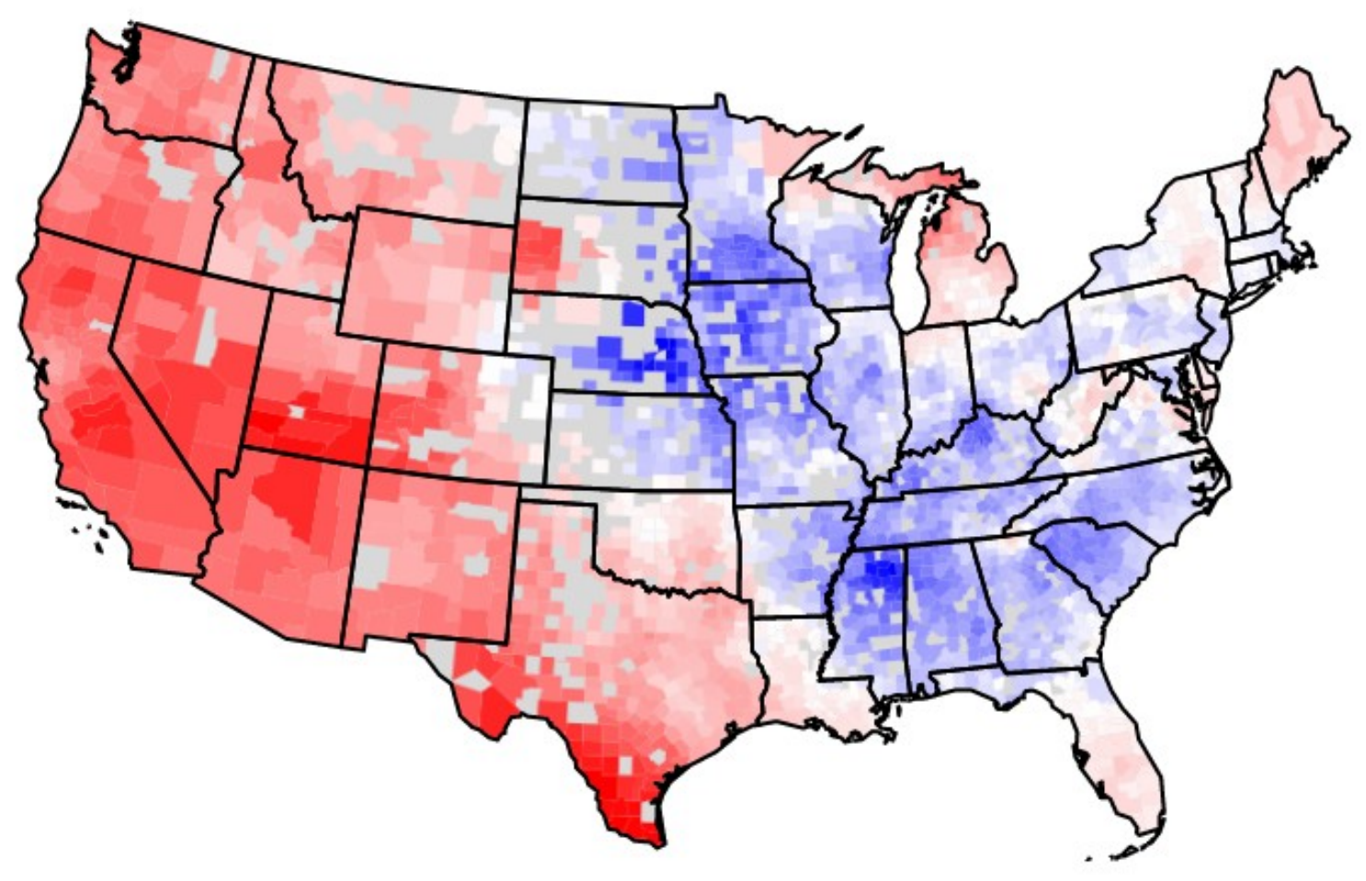
Obtaining the principal dimensions of regional variations

Extract the most important dimensions of regional lexical variation with Principal Component Analysis (PCA) → word-G_i^*-space of 10,000 dimensions to PC-space of ~300 dimensions. To avoid assuming noise/signal ratio, number of components selected using broken-stick model.

Three examples of the first components in terms of explained variance:


First component: North / South with African American influence


Urban / rural areas

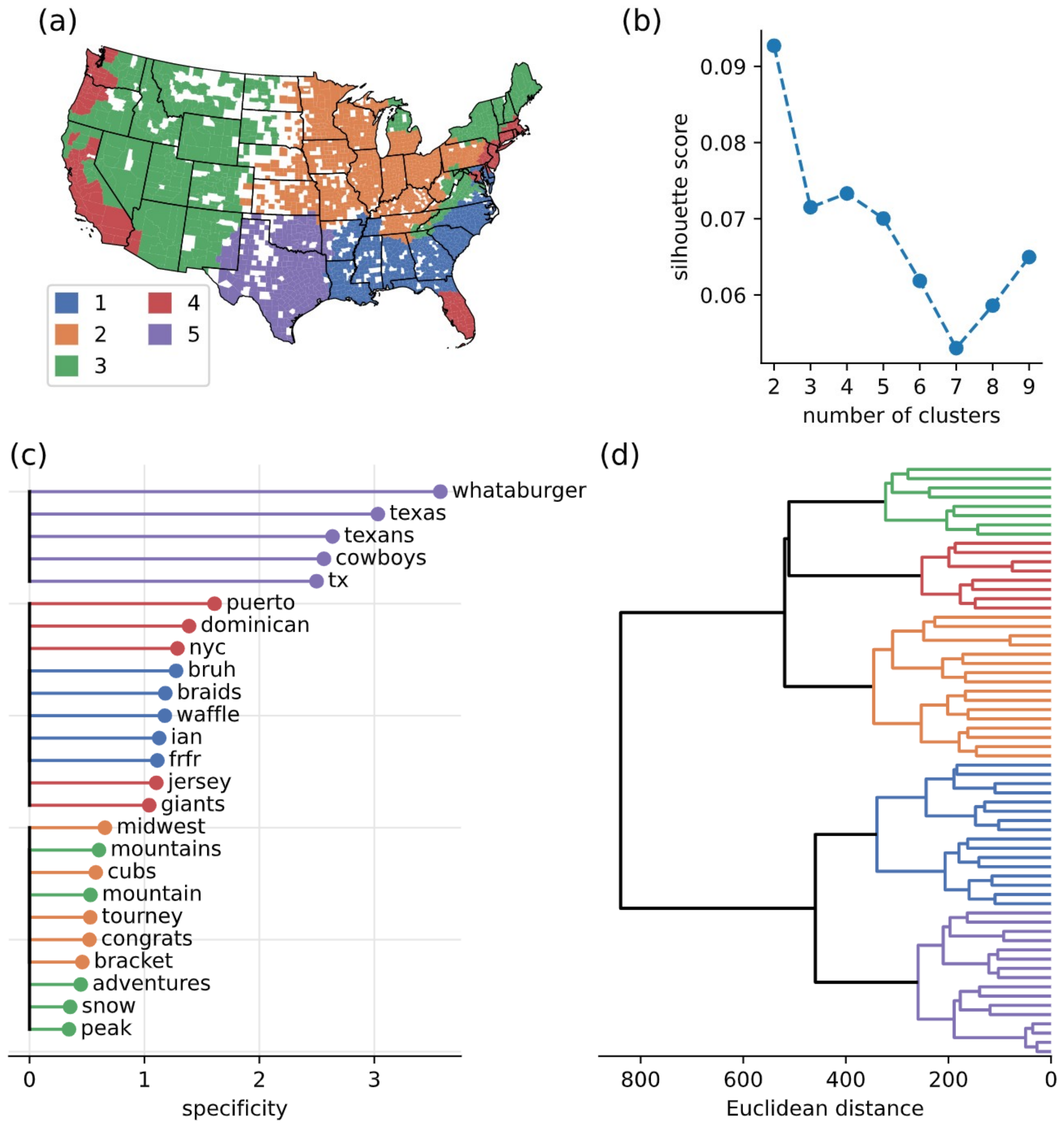

East / West

Inferring cultural regions

Cluster together counties with similar lexical signature with a hierarchical clustering → cultural regions defined from our corpus.

Meaningful number of clusters selected considering silhouette score evolution (b) + dendrogram (d). The latter shows which clusters are joined together at higher levels, and thus how North/South division is strongest.

We also identify the words which are most specific of each cluster (c): African American vernacular English in Deep South, sports-related in Midwest, outdoors lexical field for wilder areas.



Conclusions

- From a corpus of 3 billion geo-tagged tweets collected over 7 years, we are able to measure the regional variations of popular word forms (including vernacular forms).
- This allows for an automatic detection of cultural regions, here in the US, without any previous choice of topics / other linguistic features.
- The US has a very distinct North / South distinction, and from our method can hardly be divided into more than 5 regions. Further divisions may depend on data accuracy and method limitations.
- This can further be expanded to study cultural differences across countries that share a same language, like those speaking a so-called world language like English, French, Spanish, Arabic...