Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-01611-3 OPEN

American cultural regions mapped through the lexical analysis of social media

Thomas Louf[1✉], Bruno Gonçalves[2], José J. Ramasco[1], David Sánchez[1✉] & Jack Grieve[3,4]

Cultural areas represent a useful concept that cross-fertilizes diverse fields in social sciences. Knowledge of how humans organize and relate their ideas and behavior within a society can help us to understand our actions and attitudes toward different issues. However, the selection of common traits that shape a cultural area is somewhat arbitrary. What is needed is a method that can leverage the massive amounts of data coming online, especially through social media, to identify cultural regions without ad-hoc assumptions, biases, or prejudices. This work takes a crucial step in this direction by introducing a method to infer cultural regions based on the automatic analysis of large datasets from microblogging posts. The approach presented here is based on the principle that cultural affiliation can be inferred from the topics that people discuss among themselves. Specifically, regional variations in written discourse are measured in American social media. From the frequency distributions of content words in geotagged tweets, the regional hotspots of words' usage are found, and from there, principal components of regional variation are derived. Through a hierarchical clustering of the data in this lower-dimensional space, this method yields clear cultural areas and the topics of discussion that define them. It uncovers a manifest North–South separation, which is primarily influenced by the African American culture, and further contiguous (East–West) and non-contiguous divisions that provide a comprehensive picture of modern American cultural areas.

[1] Institute for Cross-Disciplinary Physics and Complex Systems IFISC (UIB-CSIC), Palma de Mallorca, Spain. [2] ISI Foundation, Turin, Italy. [3] Department of English Language and Linguistics, University of Birmingham, Birmingham, UK. [4] The Alan Turing Institute, London, UK. ✉email: thomaslouf@ifisc.uib-csic.es; david.sanchez@uib.es

Introduction

Cultural identity is an elusive notion because it depends on a wide range of different cultural factors—including politics, religion, ethnicity, economics, and art, among countless other examples—which will generally differ across individuals, with the cultural background of every individual ultimately being unique. Nevertheless, individuals from the same region can generally be expected to share some cultural traits, reflecting the shared cultural values and practices associated with the region (Broek et al., 1973). Identifying the cultural regions of a nation—regions whose populations are characterized by relative cultural homogeneity compared to the populations of other regions within the nation—is very valuable information across a wide range of domains. For example, it is important for governments to understand geographical variation in the values of their population so as to better meet their educational, social, and welfare needs. Similarly, from an economic standpoint, it is important to identify where certain services and products are most required and how best to engage with populations in different regions of the country. In general, defining the cultural regions of a nation is therefore a crucial part of understanding the complex landscape of human behavior that a nation encompasses, providing an accessible and broad classification of the populations of a country (Lane and Ersson, 2016).

Mapping cultural regions have been a particularly active area of research in the US, where there has long been debate over the cultural geography of the country, with a wide range of theories of American cultural regions having been proposed. Seven of the most prominent theories (Elazar, 1970; Garreau, 1996; Gastil, 1975; Lieske, 1993; Odum, 1936; Woodard, 2012; Zelinsky, 1973) are mapped in Fig. 1, showing considerable disagreement. For example, five major cultural regions—New England, the Midland, the South, the Middle West, and the West—have been identified (Zelinsky, 1973) based on a synthesis of regional patterns in a wide range of cultural factors, including ethnicity, religion, economics, and settlement history. An alternative proposal (Gastil, 1975), drawing on a similar but more extensive range of cultural factors, identified 13 major cultural regions, offering a more complex theory than Zelinsky, including by dividing Zelinsky's Midland, Middle West, and West regions. These two studies illustrate two basic limitations with these types of approaches that subjectively synthesize a range of data to infer cultural regions. First, it is unclear exactly how relevant cultural factors should be identified. Zelinsky considers fewer factors than Gastil, which may explain his simpler proposal. Second, it is unclear how these different factors should be synthesized to produce a single overall map of cultural regions. Zelinsky places greater emphasis on the importance of initial settlement, which may also explain his simpler proposal.

Given the subjectivity underlying these studies, the lack of agreement over the number and location of American cultural regions (as illustrated in Fig. 1) is not surprising. Only a distinction between the North and South, reflecting the Union-Confederacy border, and a distinction between the East and West, reflecting the path of the Rocky Mountains, are common to these most influential theories of American cultural regions (Elazar, 1970; Fischer, 1989; Garreau, 1996; Gastil, 1975; Lieske, 1993; Odum, 1936; Woodard, 2012; Zelinsky, 1973). Otherwise, between 4 and 12 primary cultural areas have been mapped, typically including the Northeast (Elazar, 1970, Fischer, 1989; Garreau, 1996; Gastil, 1975; Lieske, 1993; Odum, 1936; Zelinsky, 1973), the South (Elazar, 1970; Fischer, 1989; Garreau, 1996; Gastil, 1975; Lieske, 1993; Odum, 1936; Woodard, 2012; Zelinsky, 1973), the West (Elazar, 1970; Garreau, 1996; Gastil, 1975; Odum, 1936; Woodard, 2012; Zelinsky, 1973), and the Midwest (Elazar, 1970; Garreau, 1996; Gastil, 1975; Odum, 1936; Zelinsky, 1973).

In large part, the debate over the geography of American cultural regions has been about which types of cultural factors should be given precedence, and how these factors should be combined. Crucially, these decisions have generally been left entirely to the judgment of the analyst. Quantitative data from the census and elections have sometimes been taken into consideration (e.g. Gastil, 1975; Lieske, 1993; Woodard, 2012; Zelinsky, 1973), but less often subjected to statistical analysis (e.g. Lieske, 1993), while the selection and weighting of these factors have always been subjective. For example, religion and politics are undoubtedly important cultural factors, but they can be measured in various ways, and it is unclear how important they are relatively speaking, and whether their importance varies across the United States.

A basic question is therefore how can we infer general American cultural regions in an objective way? In particular, how can we both identify a complete or at least representative range of relevant cultural factors and somehow combine these factors so as to map American cultural regions? Defining such regions does not mean that they do not contain internal variation or that they are separated by hard borders—culture is dynamic and complex and humans are highly mobile—but that we can find areas where the cultural practice and values of the people who live within that region are more similar to each other than to those of people who live outside that region.

The goals of this paper are therefore to address these issues, by (i) proposing a novel method for discovering cultural regions by identifying regional patterns in topics of conversation, and by then (ii) proposing a theory of American cultural regions derived from the application of this method to a large corpus of geolocated social media data.

Our starting premise is that cultural regions will necessarily be reflected by regional variation in the topics that people choose to discuss in their everyday lives. If the cultural geography of the US was broadly homogenous, we would expect topics of conversation to be largely the same across the country, aside from different uses of place names and other such relatively superficial and necessarily regionalized vocabulary items. However, if people from different regions exhibit distinct and systematic cultural characteristics—for example, in politics, religion, music, sport, and fashion—as research on American cultural geography has consistently shown, then these patterns of cultural variation will necessarily manifest themselves as patterns of topical variation in the language used by people from these regions (Kramsch, 2014). For example, if hip-hop music, baseball, tattoos, or some other cultural practice is especially popular in some parts of the country, we would expect more discussion on that topic in large samples of everyday language use originating from that region, including on social media. Furthermore, cultural characteristics are often regionally patterned and inter-related. For example, regional patterns in ethnicity and religion often reflect settlement patterns, which can in turn help explain regional patterns in politics. Consequently, analyzing these regional topical patterns in the aggregate can be used to infer broad cultural regions.

Crucially, there is no need to predefine what these topical patterns are or how much they matter: the topics themselves and their relative importance can be inferred through the analysis of everyday language as well. We, therefore, introduce an automated method for identifying cultural regions based on the automated identification of patterns of regional variation in topics of discussion in very large corpora of geotagged everyday language use. Our method is specially intended to take advantage of the incredibly large amount of geotagged social media data that online communication now provides us with for the first time, although our method could be used to identify cultural regions

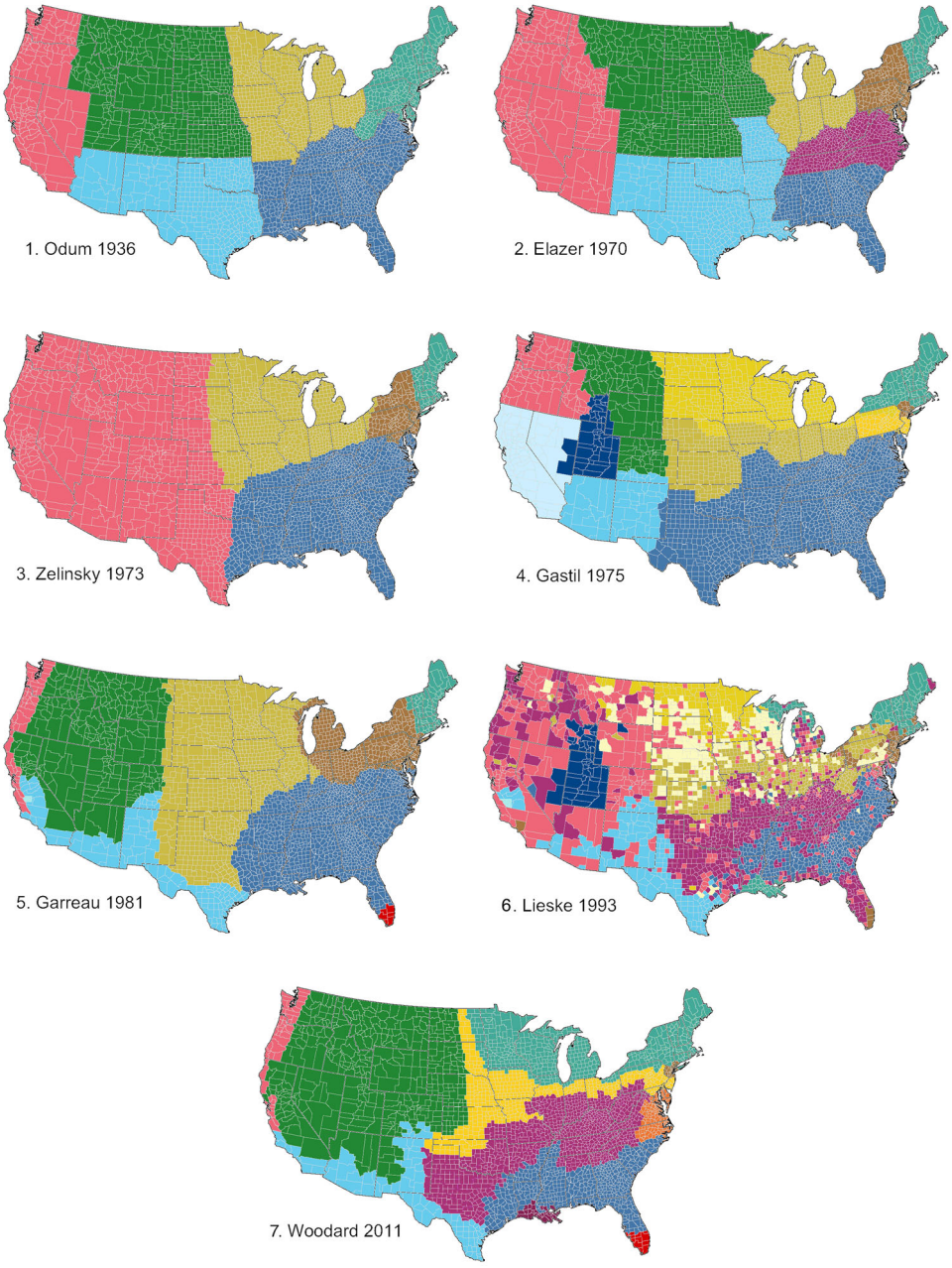

Fig. 1 American cultural regions identified in previous studies. These maps are generated from a compilation of results given in eight previous studies (Elazar, 1970; Garreau, 1996; Gastil, 1975; Lieske, 1993; Odum, 1936; Woodard, 2012; Zelinsky, 1973).

within any area based on any substantial source of regionalized everyday language use.

Specifically, to map modern American cultural regions, we identify regional patterns in the topics that Americans tend to discuss on social media through a quantitative analysis of 10,000 lexical items in over 3.3 billion geotagged tweets from across the US, collected between 2015 and 2021. Large corpora of geotagged Twitter data have been used frequently in computational sociolinguistics (Nguyen et al., 2016) to map patterns of dialect variation (Abitbol et al., 2018; Donoso and Sánchez, 2017; Eisenstein et al., 2014; Gonçalves et al., 2018; Gonçalves and Sánchez, 2014; Grieve, 2016; Grieve et al., 2019, 2011; Huang et al., 2016), while others have leveraged methods such as Latent Dirichlet Allocation to identify regional topical patterns (Funkner et al., 2021; Koylu, 2018). Despite this wealth of research that has used large corpora of social media to identify regional patterns in language use, we are aware of no research that has used this type of information to infer the location of general cultural regions.

Of course, social media or any other form of language can only provide a partial picture of regional patterns in overall topics of discussion in a region. In general, big data corpora generated from microblogging platforms certainly present a number of biases: incomplete demographic representativeness (Mislove et al., 2011), particularly for users geotagging their tweets (Pavalanathan and Eisenstein, 2015), non-homogeneous spatio-temporal distribution (Steiger et al., 2015), or severe topic differences with the offline world (Diaz et al., 2016). However, if cultural regions are real and pervasive, then we should expect these regions to manifest themselves in any large sample of everyday language that encompasses a large proportion of the population, even if the specific topics of interest vary across these different domains. Furthermore, right now, Twitter is the only

variety of geotagged natural language data available in sufficient amounts to allow reliable automatic analyses, and is a very popular social media platform used regularly by millions of people from across the US, mostly in interactive contexts (Auxier and Anderson, 2021), serving as a perfect domain to apply our data-driven approach for automatically mapping cultural regions.

Our main finding is that the modern US can be divided into five primary cultural areas, each defined by its own topical patterns. We emphasize that this result stems from quantitative analysis in contrast to previous proposals based on more or less informative (qualitative) approaches. Further, beyond the specific number of regions it is most relevant to note that our method yields the list of words and topics that define those regions, which highlights the differences in interests, habits, and backgrounds that distinguishes each cultural region from the others. Crucially, by means of dynamic analysis, we show that the cultural regions of the US are relatively stable over the past few years, offering further evidence that cultural areas are real phenomena that pervade American society.

The rest of the article is structured as follows. The results of the work are first introduced by a description of the dataset collection and pre-processing methodology. Regional variations of word usage observed from this dataset are then explored, before obtaining the principal dimensions of these variations. The main result of the work, the cultural regions of the US, and the main topics of discussion that define them are then presented in detail. The possibility of a variation with the time of the results is then explored. Finally, a discussion of the insights brought by the analysis and also of where future works could build on it comes to conclude the work.

Results

Dataset. We analyze geotagged tweets collected through the streaming API of Twitter, more specifically, using the filtered stream endpoint: https://developer.twitter.com/en/docs/twitter-api/tweets/filtered-stream. This endpoint provides a sample of tweets in real-time matching suitable filters. This allows us to gather 3.3 billion geotagged tweets from the contiguous US, posted from January 1, 2015, to December 31, 2021. Importantly, we discard users tweeting at an inhuman rate, which we define to be any rate superior to 10 tweets per hour over one's whole tweeting span. We also discard users tweeting from any platform that is not a Twitter mobile application or their website. In our dataset, we thus retain 17 million users. We strip tweets of any link, hashtag, or user mention, and only keep those that still have more than 4 words after this filter. Hashtags were discarded out of precaution: some of them may be content words, but they may also be related to short-lived trends for instance. As we found that the content of hashtags accounted for <5% of the content of the tweets we collected, they can anyway be safely discarded. We subsequently use the Chromium Compact Language Detector (CLD) (Al-Rfou and Solomon, 2014) to eliminate tweets written in a language other than English. To attach geolocation to tweets, they are geotagged with either the precise GPS coordinates of the device of the user or "places", which can be an administrative region, a city, or a place of interest. Then, as these geotags may be places of the size of a state, we also remove tweets with a geotag that did not allow for reliable assignment to our unit areas, which are the US counties and county equivalents (3108 in total). Certainly, counties vary in both size and population but most of them form a useful division sufficiently large to show a sizeable amount of tweets and sufficiently small to allow for a careful delimitation of cultural areas (states would be too big units whereas towns would be too small).

From the remaining tweets, we extract and count the tokens in their text, and assign them to counties. Counties that accumulate fewer than 50,000 tokens are not taken into account, leaving us with $N_c = 2576$ counties which defines our sub-corpora. We thus keep 83% of the total number of counties. After this filtering, the full dataset contains 9.1 billion tokens (see Table S1 for a summary description of the dataset). We subsequently convert the remaining word forms to lowercase and aggregate the token counts on these forms. We then remove all function words (like *the, and*) and interjections (like *um, oh*) (see Data availability for access to the full list of exclusions), and consider the 10,000 most common remaining word forms. Note that this list of word forms emerges from the data, and is not imposed by any previous topical or dialect classification.

Measuring regional variation. We then measure and map the relative frequencies $f_{c,w}$ for every word w in every county c. We illustrate our raw results by plotting in Fig. 2 the relative frequency in each county of four representative words: (a) *today*, (c) *mountain*, (e) *traffic,* and (g) *bruh* (cells that appear grayed out do not reach a minimum number of tweets as explained in the paragraph above). In the first case, *today* appears at relatively stable rates in most of the counties, as expected. Alternatively, *mountain* is a regionally dependent word as clearly seen. The item *traffic* appears more frequently in urban areas. Finally, *bruh* is an African–American English variant that appears to be especially common in southern counties, where there are large African–American populations.

A word of caution is now required. A relative frequency map alone is not able to fully reveal regional variations due to the wide range of different factors besides regional variation that affect word use and add noise to the signal. To extract the underlying regional signal from each word map, we conduct a multivariate spatial analysis (Grieve, 2016; Grieve et al., 2011) of the relative frequencies of our 10,000-word forms. In order to identify geographical hotspots in the usage of each word (Fig. 2), we compute Getis–Ord's z-scores (G_i^* (Ord and Getis, 1995)) for each county c and word w, which are defined as

$$G^*_{c,w} = \frac{\sum_{c'} W_{c,c'}(f_{c',w} - \bar{f_w})}{\sigma_w \sqrt{\frac{N_c \sum_{c'} W^2_{c,c'} - \left(\sum_{c'} W_{c,c'}\right)^2}{N_c - 1}}}, \quad (1)$$

with $\bar{f_w}$ the average frequency of w over the whole dataset, σ_w the standard deviation in w's frequencies, and $W_{c,c'}$ are the elements of a proximity matrix, which we take as equal to 1 if $c' = c$ or c' belonging to c's 10 nearest neighbors, and equal to 0 otherwise.

The metric given by Eq. (1) ultimately diminishes spurious data variation and smooths spatial patterns, allowing us to discern a regional pattern in a word's usage. In Fig. 2b, d, f and h we show, respectively, the G_i^* z-scores for the previous words *today*, *mountain*, *traffic*, and *bruh*. White, light blue, or light red counties do not depart significantly from average utilization, whereas a bright red or blue, respectively, means that the word is relatively frequently or infrequently used in that region. Since *today* is a rather generic word, we do not find any strong regional pattern, whereas the others do. The usage hotspots of *mountain* display the main mountain ranges of the country. While the map for *traffic* is correlated with large urban areas (and can be interpreted as a topical word), the dialect word *bruh* seems to be significantly more used in counties pertaining to the Deep South. We see here that different attributes that define a culture (interests, behavior, dialect) are captured within our scheme and, notably, are treated on equal footing.

Obtaining the principal dimensions of regional variation. The G_i^* distributions for all 10,000 top words by usage thus hold

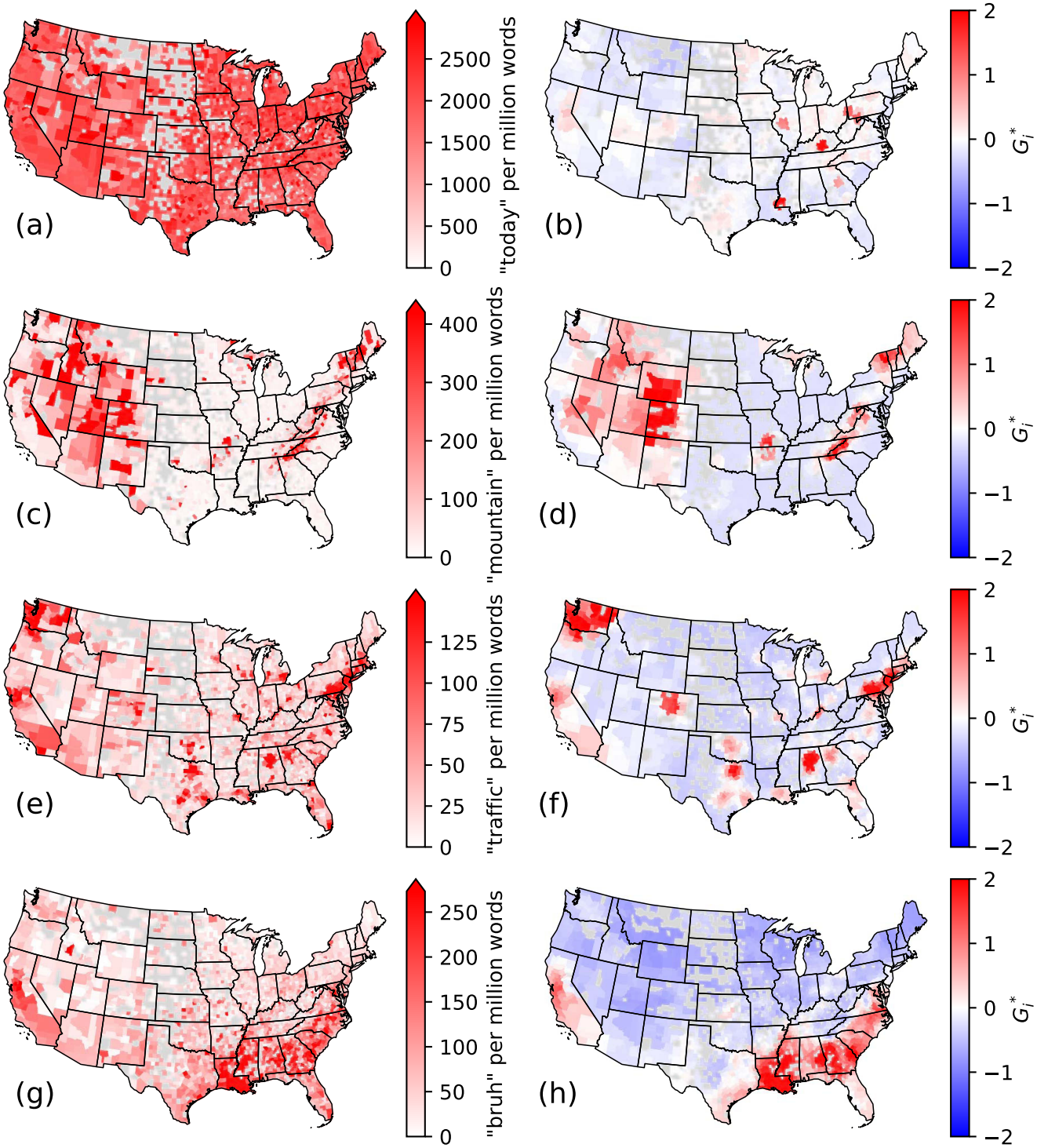

Fig. 2 Hotspots in the usage of characteristic words. Maps showing the (**a**, **c**, **e**, **g**) relative frequency and (**b**, **d**, **f**, **h**) Getis-Ord G_i^* *z*-score for the words *today*, *mountain*, *traffic*, and *bruh*, respectively. One can note how the latter metric enables to reveal of word usage hotspots, smoothing out the raw noisy signal from the data.

valuable information. However, a considerable part of this information can be analyzed more efficiently, since some words may belong to the same semantic field (*mountain* and *peak*) or characterize the same particular dialect (*bruh* and *aight*). Furthermore, a few variations may simply be uninformative noise, intrinsic to real individuals' behaviors, but also potentially resulting from imperfect filtering of Twitter data, as aforementioned. The most important dimensions of regional lexical variation are then found by subjecting the hotspot maps for the complete set of words to a principal component analysis (Lieske, 1993; Wold et al., 1987). Another possible approach would have consisted in performing topic modeling, for instance by ways of a Latent Dirichlet Allocation on the word frequency matrix, to then infer a distribution of topics for every county. It is,

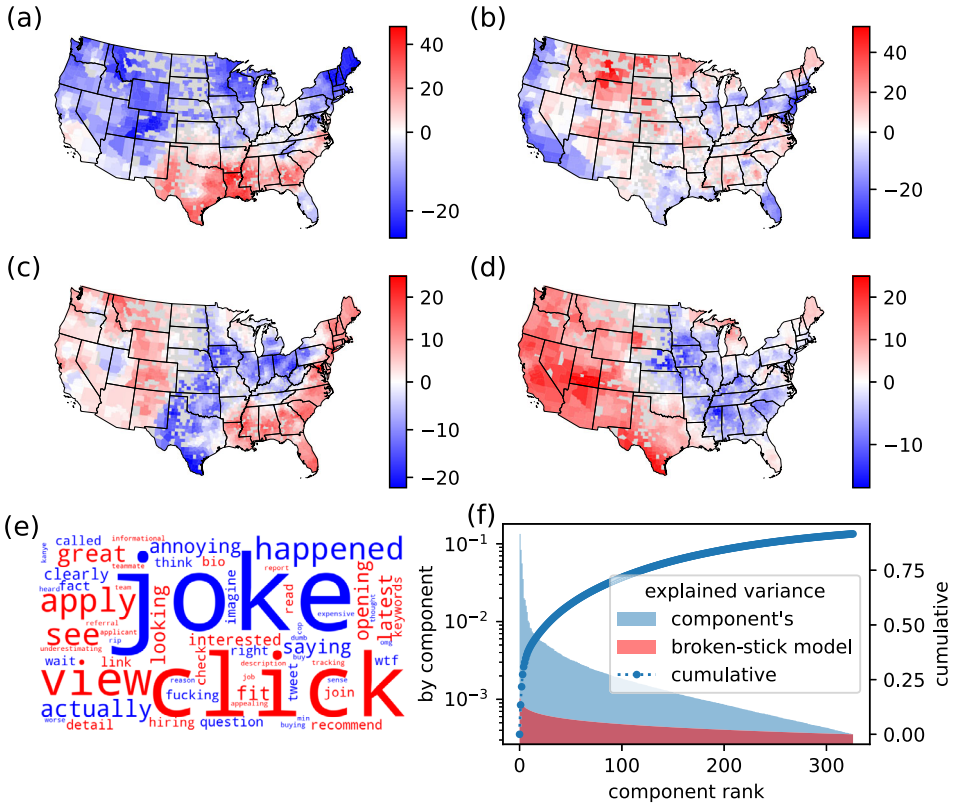

Fig. 3 Result of the principal component analysis carried out on our whole dataset. a–d Four maps show the projection of the data along the first four components, highlighting regional lexical variations. Note that the scale on the divergent color scales is not symmetrical around zero in order to utilize the full range of colors of the color map. **e** Word cloud showing the words with the strongest positive (red) and negative (blue) loadings for the second component, with each word's font size depending on its loading's absolute value. **f** Explained variance of the principal components compared to the broken-stick model on a logarithmic scale, which shows how the number of components to keep is selected at the first intersection of the two curves. The cumulative proportion of the variance explained by the components is also plotted, showing that our dimension reduction explains around 92% of the observed variance. The first four components shown in panels **a–d** capture alone 31% of the variance.

however, more computationally intensive, and poses questions about the selection of the number of topics, their interpretability, and their internal coherence (Arun et al., 2010; Hasan et al., 2021). In a case like ours where documents are so large (aggregating all tweets in a county), it is far from obvious to select a number of topics such that there is little overlap between them and to know that these topics are actually representative of the dataset as a whole. This is much more clear when selecting components in PCA, as we show below.

From the $N_w = 10{,}000$ dimensions of our dataset, we thus project to a principal component (PC) space of $N_{\mathrm{PC}} = 326$ dimensions. It turns out that these 326 components explain 92% of the observed variance (see Fig. 3f). We do not set this number of components arbitrarily, by choosing one directly or by setting a percentage of variance we wish to explain using these components. Instead, we use the broken-stick rule to fix the number of components (Frontier, 1976; Jackson, 1993). This heuristic compares the decrease of the variance explained by each successive component to the one expected from a random partition of the whole variance in N_w parts. Components, sorted by decreasing explained variance, are kept until they do not explain more variance than their corresponding random part would. With this method, we do not make any assumption about the amount of variance in our data that is simply due to random fluctuations.

We show the projected data along the first four PCs in Fig. 3a–d, which displays a neat visualization of the spatial patterns. The map for each dimension shows two opposing regions (red and blue) which can be linked to their characteristic words, the ones with the highest (positive, in red) and lowest (negative, in blue) loading. For an illustration, in Fig. 3e we show in a word cloud the most characteristic words for each of the two regions in Fig. 3b, which corresponds to the second component. In Figs. S2 and S3 we plot, respectively, the projected data for the proximity matrix $W_{c,c'}$ of Eq. (1) defined based on the 5 and 15 nearest neighbors. The results show that the components are not significantly altered by a slight change in the proximity matrix. Figure S4 shows the results when $W_{c,c'}$ is alternatively defined in terms of a fixed distance. In this case, the modifications are stronger because the size of the counties is not uniform. This demonstrates that one should take proper neighbor couplings when dealing with heterogeneous geographical units.

Inferring cultural regions. We are now in a position to generate a single overall taxonomy of American cultural regions by clustering together counties with similar lexical signatures. To do so, we subject the previous PC maps to hierarchical clustering, using the Euclidean distance and the Ward variance minimization algorithm (Everitt et al., 2011). This is how we define the cultural regions from our corpus, as depicted in Fig. 4. From the dendrogram and the evolution of the average silhouette score for different levels of clustering, we select a meaningful number of clusters n_{clusters} (Rousseeuw, 1987). The hierarchical nature of the clustering is useful to see how regions are grouped together at different levels of clustering, indicating which regions are closer

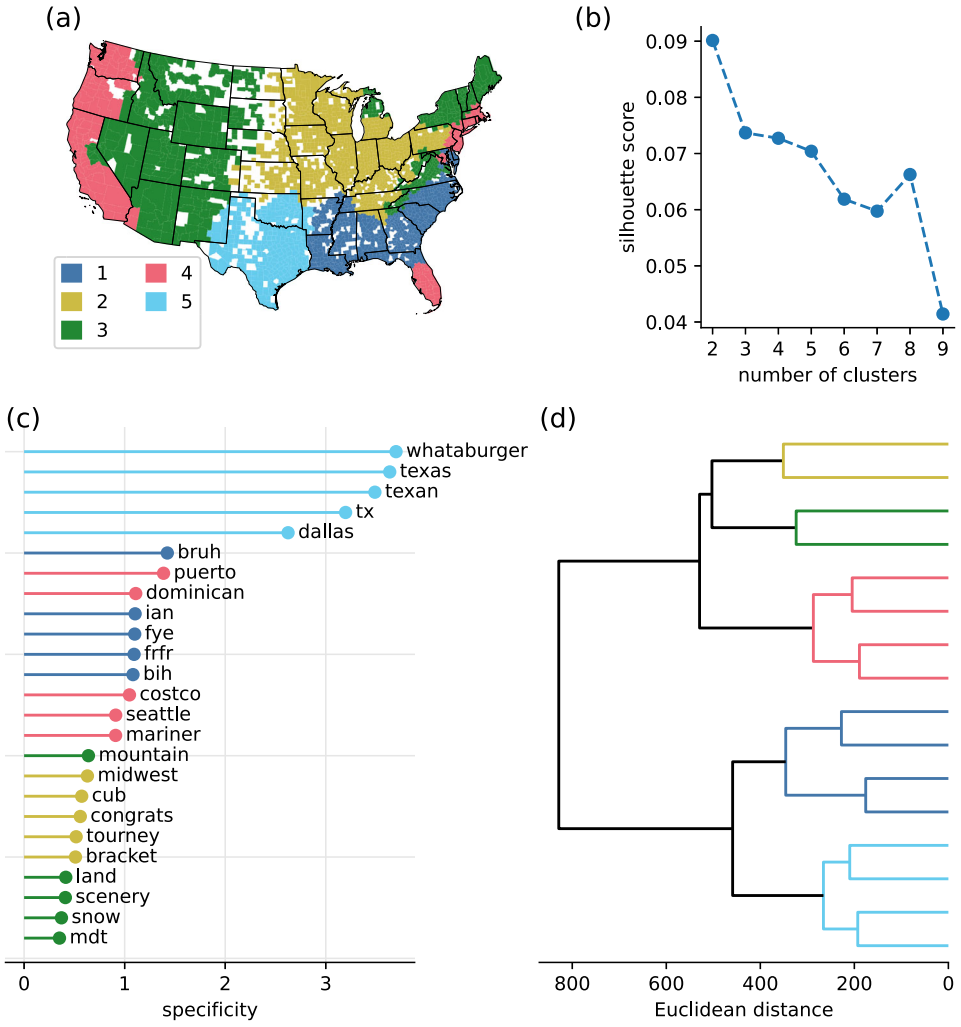

Fig. 4 Cultural regions obtained from our whole dataset. a Map of the five clusters obtained through hierarchical clustering, selected from a high value of **b** the mean Silhouette score. A significant drop in the Silhouette score after 5 levels indicates that further splitting counties into more regions do not yield coherent cultural regions. **c** Five most specific words for the five clusters shown in **a**, along with their specificity values. **d** The dendrogram allows seeing which clusters are first joined if going to a higher level of clustering, and thus which ones are closer together. It clearly shows that the strongest division is the one between the North and the Southeast (excluding Florida) with further splittings as the cluster distance increases.

together. Importantly, applying hierarchical clustering to the principal dimensions of variation of the data obtained through PCA allows us to focus on the main regional patterns of variation. Applying the algorithm directly to the 10,000-word distance matrix would yield highly noisy results.

We plot the main divisions in Fig. 4a. This is the main result of our paper. In the map, we present the division into five clusters since it is one of the two best options as characterized by the Silhouette score analysis in Fig. 4b, and at a clear-cut on the dendrogram in Fig. 4d. The optimal choices correspond to the two significant drops in the score: the first (second) corresponds to a cluster number equal to 2 (5).

Indeed, the dendrogram in Fig. 4d shows that the counties can be initially classified into two large-scale subgroups representing a North vs. South divide. The North is then further fragmented into clusters 2–4 shown in Fig. 4a, whereas the South group splits into clusters 1 and 5. For the most part, our map in Fig. 4a is consistent with standard theories of American cultural regions, with all five of our regions finding analogs in existing systems. Yet, taken as a whole our clusters do not match any previous system and reveal non-contiguous culture regions such as clusters 3 and 4. Moreover, in contrast to previous proposals our results have the advantage of being data-driven, based on variation in the topics people care to discuss as opposed to factors selected by hand by the researcher (and consequently subjected to many more, uncontrolled biases than our Twitter data).

Further, to be able to better interpret the obtained regions, it is insightful to know which words characterize each cluster the most. To infer them, we start by taking the center of each cluster in words—G_i^*-space. Hence, for each cluster, we take the average G_i^* score over its counties for all words. From these n_{clusters} vectors of N_w elements, we calculate the minimum absolute difference between each cluster center's word's score and the ones of all other clusters, i.e., we take the distance to the closest cluster's center along the word's dimension. More formally, we define the specificity $S_{C,w}$ of word w for cluster C as:

$$S_{C,w} = \min_{C' \in \mathcal{C}\setminus C} \left(\frac{1}{N_C} \sum_{c \in C} G^*_{c,w} - \frac{1}{N_{C'}} \sum_{c \in C'} G^*_{c,w} \right)^2, \qquad (2)$$

where $\mathcal{C}$ denotes the set of clusters, N_C the number of counties belonging to cluster C, and $G^*_{c,w}$ the G_i^* score of word w in county c. For each cluster C, we thus define the most characteristic words as the ones with highest $S_{C,w}$ values. In the case of the division

into five clusters, the top 5 most characteristic words per cluster are shown in Fig. 4c, according to the specificity metric defined in Eq. (2). In all cases, the five cultural regions are linked to clear and distinct topical patterns (see the Supplementary Information for a more exhaustive list). We stress that these characteristic words are automatically identified based on the quantitative analysis presented above. Notably, for each cluster, we see three basic types of lexical patterns.

First, we see words associated directly with those locations, most commonly the names of cities, states, and sports teams. This is basic evidence that the method works: we would expect these words to be associated with the cultural regions that contain them. However, these results also reflect how often people from different cities and states refer to each other. For example, the fifth cluster which is centered on Texas also includes Oklahoma, which contributes various place names to the list of words most strongly associated with this region. This means not only that people in Texas and Oklahoma talk more about place names in their own states, as would be expected, but that they talk more about place names in each other's states. This is one type of regional topical pattern that our approach draws to identify cultural regions.

Second, we observe words connected with non-regional topics, which nonetheless show regional differences. In this case, our approach can be seen as discovering topical patterns and by extension cultural patterns that distinguish between different regions of the US. For example, cluster 2 is strongly associated with the discussion of a range of American sports, as well as the names of the states that fall within this region. Although we would expect that a region centered around the Midwest would be associated with the names of Midwestern states, their preoccupation with the discussion of sports on Twitter is not so easy to predict.

Third, we find words that are dialect items, i.e., alternative ways of referring to a given concept. This type pattern is especially apparent for cluster 1, which aligns closely with the region of African American population density and is therefore associated with numerous lexical items from African-American English (e.g. *bruh, lawd, turnt*). Although dialectologists do not usually focus on the frequencies of individual words, this result is to be expected: dialect regions, which can be seen as a type of cultural region, have been found to generally align with broader cultural regions (Grieve, 2016).

We can now examine each of the five cultural regions we have identified in turn and consider what the words that are most strongly associated with each tells us about the culture of that region, as well as the factors that drive cultural variation in the US more generally.

The first cluster [blue in Fig. 4a], which identifies a southeastern region, largely reflects African American culture, as can be predicted based on the close correlation between our map and the distribution of counties with relatively large African American populations (see Fig. S1). Most notably, tweets from the South are more likely to contain words related to African American culture, including, for example, cuisine (e.g. *grits, cookout*), fashion (e.g. *braids, dreads*), and music (e.g. *rappers, rapping*). As noted above, this cluster is also strongly characterized by many vocabulary items associated with African American English, especially for referring to people (e.g. *bruh, dawg*), as well as many acronyms (e.g. *frfr, stg*). Place names associated most strongly with this cluster primarily include southern states (e.g. *Georgia, Carolina*), despite the fact that, in general, references to place names are relatively rare compared to other clusters.

The second cluster [yellow in Fig. 4a] has its core in the Midwest and is clearly characterized by more frequent references to sports. American team sports especially stand out, with 40 words of the top 50 most strongly associated with this cluster being directly linked to this topic. In particular, these are words associated with basketball (e.g. *basketball, rebound*) and baseball (e.g. *baseball, innings*), although football, wrestling, and cheering are also referenced, as well as various more generic sporting terms (e.g. *teams, tourney*). Similarly, many place names are associated with local sports teams (e.g. *Cubs, Chiefs*), although various state names are also strongly associated with this cluster (e.g. *Ohio, Illinois*), as well as the word *Midwest* itself. A smaller number of lexical items are also associated with school (e.g. *locker, choir*). Overall, this cluster, therefore, shows that sports are a central part of this region.

The third cluster [green in Fig. 4a] can be identified with a discontinuous region that mostly aligns with rural areas of the US, as well as areas that focus on outdoor activities, especially in mountainous regions (e.g., the Rocky or Appalachian Mountains). This cluster is relatively hard to interpret topically, in part because, unlike the other regions, it is characterized by the relatively infrequent use of a number of words. In terms of words that are relatively common in this region, the clearest pattern is a relatively large number of words associated with nature (e.g. *mountains, tree*), weather (e.g. *snow, seasonal*), and outdoor activities (e.g. *adventures, trail*). Clearly, people in this region tend to focus more on their natural surroundings. In addition, there are a number of words related to work (e.g. *hiring, jobs*), as well as numerous place names (e.g. *Colorado, Montana*) that are strongly associated with this region. In terms of words that are uncommon within the cluster, there exist many verbs, especially verbs associated with human actions like communication (e.g. *said, told*), thought (e.g. *understand, confused*), and physical actions (e.g. *put, hit*), which implies overall less focus on the individual. This region is also associated with relatively infrequent use of a wide range of negative words (e.g. *wrong, bad*), which largely hints at a more positive outlook.

The fourth cluster [red in Fig. 4a] also identifies a discontinuous region that primarily encompasses large urban areas on the coasts (Northeast and West). Unsurprisingly, this region is characterized by a wide range of words associated with more urban life (e.g. *homeless, traffic*), especially terms related to different nationalities and immigration (e.g. *Latino, Asian*). We also find a relatively large number of place names (e.g. *California, NYC*). Strikingly, this cluster is associated with a very large number of words with negative connotations, including relating to violence (e.g. *violence, attack*), danger (e.g. *dangerous, crime*), cursing (e.g. *asshole, fucking*), political unrest (e.g. *protests, indicted*), racism (e.g. *Nazi, supremacist*), and general negative adjectives (e.g. *disgusting, abusive*). Quite generally, people from this cluster are more likely to discuss negative topics than other parts of the US, at least on social media. Taken together, the third and fourth clusters suggest an opposition in the culture of more rural and urban areas in the US, which appear to engage in more positive and negative discourse respectively (Vanderbeck and Dunkley, 2003).

Finally, the fifth cluster [cyan in Fig. 4a], which is centered around the South Central States, especially Texas and Oklahoma, is characterized by frequent reference to place names, relative to the other clusters, especially in these two states, as has already been noted. For example, the first five most strongly associated words are *Whataburger* (a fast food chain from Texas), followed by *Texas, TX, Texan, and Dallas*. This not only shows that people in this region tend to discuss places more on Twitter but implies that this cultural region is characterized by a relatively high amount of local pride. Correspondingly, this region is also associated with a relatively large number of dialect terms, both of Anglo (e.g. *yalls, fixing*) and Hispanic (e.g. *queso, taco*) origins, reflecting the diverse makeup of this region.

The analysis yielding Fig. 4 was repeated, adding a stemming step at the very beginning of our pipeline. We obtain a very similar result, shown in Fig. S5, indicating little sensitivity of our results to stemming.

Given the lack of consensus in previous research, our results can help resolve long-standing debates relating to the distribution of American cultural regions. We find that the division between the Southeast and the rest of the US is the strongest. This result attests to the importance of the cultural divide between White and Black America and between the North and the South. Although all previous major theories of American cultural regions have identified a distinction between the North and the South, our southern region is especially similar to relatively recent theories, which identify a southern region that closely aligns with the part of the south with an especially high proportion of African Americans (Lieske, 1993; Woodard, 2012). Another key finding that emerged from our analysis is a broad opposition between coastal and internal areas, which has not previously been identified as important sources of distinction of American cultural regions (Elazar, 1970; Fischer, 1989; Garreau, 1996; Gastil, 1975; Lieske, 1993; Odum, 1936; Woodard, 2012; Zelinsky, 1973) but reflects a modern political trend of undeniable significance (Gelman, 2009) that is currently reconfiguring the nation. The discontinuous nature of these regions, which is not required by our definition of a cultural region, is also notable. It demonstrates how patterns in American culture can be distributed across very wide areas, reflecting complex patterns in physical and human geography, and the underlying complexity and dynamic nature of American society. This result is broadly in line with other recent theories of American cultural regions which have also identified discontinuous cultural regions (Lieske, 1993; Woodard, 2012).

Our analysis is further useful for understanding the relationship between these regions. It divides the South into two regions, splitting Texas off from the rest of the Southeast, and splits the Midwest off from the rest of the North, divided into discontinuous countryside/coastal regions, rather than contiguous cultural regions. However, on the question of the number of primary American cultural regions, we can only safely say that with our data and methodology, at least five distinct regions can be discerned. We do not see it here, but we still cannot discard recent theories that claim that America is fundamentally far more culturally fragmented (Garreau, 1996; Lieske, 1993; Woodard, 2012).

Temporal aspect of the results. Given the success of our analysis, it would be interesting to see how the cultural regions found in Fig. 4 change with time, as has been done in other research analyzing diachronic corpora (Alshaabi et al., 2021; Bentley et al., 2014; Bochkarev et al., 2015; Karjus et al., 2020; Momeni et al., 2018). Although we would not expect significant changes due to the short timescale imposed by our Twitter dataset, we can still carry out a diachronic study to validate the very existence and meaningfulness of the cultural regions. To do so, we split our corpus into three datasets corresponding to different year ranges: 2015–2016, 2017–2018, and 2019–2021. These periods have a similar amount of tokens and can be then subjected to comparison (see Supplementary Table 1). We show their maps in Fig. 5a–c. We obtain similar patterns, despite the variety of topics and forms employed on Twitter over the years and the heuristic nature of the clustering method that introduces a small amount of noise in the results. The North–South division is stable over time with small variations that can be due to either fluctuations or incipient structural changes. The latter cannot be conclusive due to the short time period considered in this work.

Next, we take the hierarchical clustering in Fig. 4d and select the county-to-cluster assignment corresponding to the highest level of the hierarchy. This is represented by the two-way division between North and South. For each year in our dataset, we then measure the pairwise distances between counties belonging to both clusters. The distances are calculated as Euclidean distances between rows of the matrix $G^*_{c,w}$ (see Eq. (1)). We thus obtain the evolution with time of the inter-cluster distances distribution as shown in Fig. 5d. The box plots demonstrate that (i) the median distance is roughly constant over the years, and (ii) the distance distribution shows little variation. Both findings suggest that the

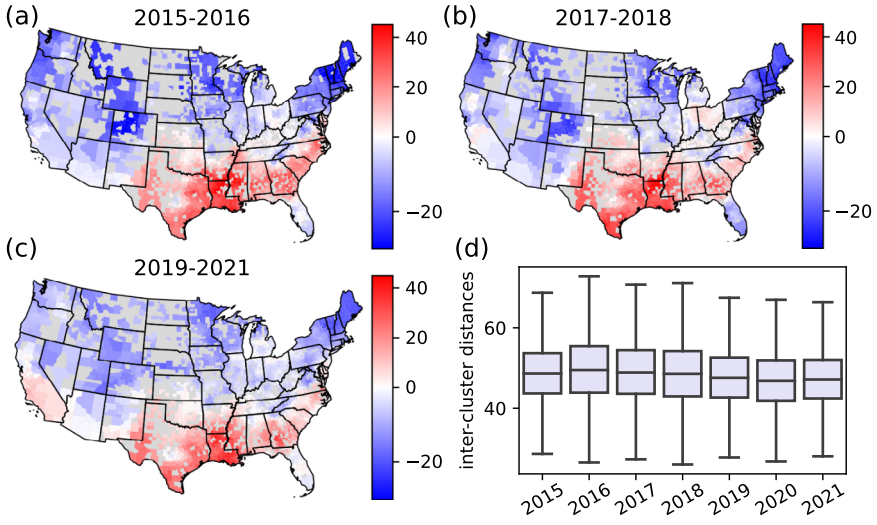

Fig. 5 Effect of temporal segmentation for the data on the obtained divisions. a–c Maps of the data projected along the first PC obtained for the years 2015–2016, 2017–2018, and 2019–2021, respectively. Apart from slight variations in California and Florida, the first component translates the same division between the Southeast and the rest of the US. Note that the scale on the divergent color scales is not symmetrical around zero in order to utilize the full range of colors of the color map. **d** Evolution of the distributions of inter-cluster distances along the years spanned by our dataset. The box plots show the median, first and third quartile, and the boundaries of the whiskers are within the 1.5 interquartile range value. We use the cluster assignment obtained with the whole dataset and measure the Euclidean distance in G^*_i space between counties belonging to different counties. The distribution is thus shown to vary little from year to year, which demonstrates the stability of the two-way division we found.

detected cultural regions are no artifact of the method, but a genuine data structure that exists within our corpus.

Discussion

Overall, our analysis has therefore identified regional patterns of lexical variation of clear cultural importance. Furthermore, the themes associated with each of these patterns provide a new perspective on American cultural geography. For example, although our analysis has confirmed that factors such as ethnicity and religion are important for defining American cultural regions, we found substantial variation in the relevance of these factors across the US. Our analysis has also identified other subtler cultural patterns—such as a focus on social interaction, the outdoors, family, and leisure—which have been overlooked in previous research, in part because they cannot be easily studied through the analysis of traditional sources of secondary data. Our method has therefore not only allowed us to map cultural regions, but it has also allowed us to identify cultural factors that are important for defining these regions, at least in this communicative context, providing a foundation for a more complete picture of the American cultural landscape.

Clearly, our study has only analyzed one genre of American English. The specific topical patterns on Twitter would not be exactly replicated in other genres, especially given the communicative purpose and user base associated with microblogging platforms. Nevertheless, assuming that American cultural regions are important and pervasive forces, similar regional patterns should be reflected across all genres. This issue could be further clarified when more richly annotated natural language data becomes available in a near future. Our methods, however, will remain valid for any such dataset. Crucially, we expect that our main idea of inferring cultural regions and the topics defining them from people's speech will be applicable to any big data resource with linguistic value.

Data availability

All aggregated data generated by our Twitter data analyses as well as our list of excluded words are available for download from a figshare repository (Louf, 2023a). County and state boundary shapefiles from the US census of 2018 that we used to draw our maps are freely available for download at https://www.census.gov/geographies/mapping-files/2018/geo/carto-boundary-file.html.

Code availability

The data processing and plotting of results were carried out in Python with the help of open-source libraries. All code used for this work is hosted on GitHub (Louf, 2023b).

Received: 15 August 2022; Accepted: 6 March 2023;
Published online: 30 March 2023

References

Abitbol JL, Karsai M, Magué JP, Chevrot JP, Fleury E (2018) Socioeconomic dependencies of linguistic patterns in Twitter: a multivariate analysis. In: The Web conference 2018—Proceedings of the world wide web conference, WWW 2018. pp. International World Wide Web Conferences Steering Committee, 1125–1134

Al-Rfou R, Solomon B (2014) Python bindings for the compact language detector 2. https://github.com/aboSamoor/pycld2

Alshaabi T et al. (2021) Storywrangler: a massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter. Sci Adv 7:eabe6534. https://doi.org/10.1126/sciadv.abe6534

Arun R, Suresh V, Veni Madhavan CE, Narasimha Murthy MN (2010) On finding the natural number of topics with latent Dirichlet allocation: some observations. In: Proceedings of the 14th Pacific-Asia conference on advances in knowledge discovery and data mining—volume Part I, PAKDD'10. Springer-Verlag, Berlin, Heidelberg, pp. 391–402

Auxier B, Anderson M (2021) Social media use in 2021. Technical Report, Pew Research Center. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Bentley RA, Acerbi A, Ormerod P, Lampos V (2014) Books average previous decade of economic misery. PLoS ONE 9:e83147. https://doi.org/10.1371/journal.pone.0083147

Bochkarev VV, Shevlyakova AV, Solovyev VD (2015) The average word length dynamics as an indicator of cultural changes in society. Soc Evol Hist 14:153–175

Broek JOM, Webb JW, Hsu M-L (1973) A geography of mankind. McGraw-Hill, New York

Diaz F, Gamon M, Hofman JM, Kıcıman E, Rothschild D (2016) Online and social media data as an imperfect continuous panel survey. PLoS ONE 11:e0145406

Donoso G, Sánchez D (2017) Dialectometric analysis of language variation in Twitter. In: Proceedings of the fourth workshop on NLP for similar languages, Varieties and Dialects (VarDial). Association for Computational Linguistics (ACL), pp. 16–25

Eisenstein J, O'Connor B, Smith NA, Xing EP (2014) Diffusion of lexical change in social media. PLoS ONE 9:e113114. https://doi.org/10.1371/journal.pone.0113114

Elazar DJ (1970) Cities of the Prairie: the metropolitan frontier and American politics. Basic Books, New York

Everitt BS, Landau S, Leese M, Stahl D (2011) Cluster analysis. John Wiley & Sons, Wiley, Chichester, UK

Fischer DH (1989) Albion's seed. Oxford University Press, Oxford, UK

Frontier S(1976) étude de la décroissance des valeurs propres dans une analyse en composantes principales: Comparaison avec le modèle du bâton brisé. J Exp Mar Biol Ecol 25:67–75

Funkner AA et al. (2021) Geographical topic modelling on spatial social network data. Procedia Comput Sci 193:22–31. https://www.sciencedirect.com/science/article/pii/S1877050921020445

Garreau J (1996) The Nine Nations of North America. Houghton Mifflin Company, Boston

Gastil RD (1975) Cultural Regions of the United States. University of Washington Press, Seattle

Gelman A (2009) Red state, blue state, rich state, poor state: why Americans vote the way they do. Princeton University Press, Princeton

Gonçalves B, Loureiro-Porto L, Ramasco JJ, Sánchez D (2018) Mapping the americanization of English in space and time. PLoS ONE 13:e0197741. https://doi.org/10.1371/journal.pone.0197741

Gonçalves B, Sanchez D (2014) Crowdsourcing dialect characterization through Twitter. PLoS ONE 9:e112074. https://doi.org/10.1371/journal.pone.0112074

Grieve J (2016) Regional variation in written American English. Cambridge University Press

Grieve J, Montgomery C, Nini A, Murakami A, Guo D (2019) Mapping lexical dialect variation in British English using Twitter. Front Artif Intell 2:11. https://doi.org/10.3389/frai.2019.00011/full

Grieve J, Speelman D, Geeraerts D (2011) A statistical method for the identification and aggregation of regional linguistic variation. Language Variation Change 23:193–221

Hasan M, Rahman A, Karim MR, Khan MSI, Islam MJ (2021) Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA). In: Kaiser MS, Bandyopadhyay A, Mahmud M, Ray K (eds) Proceedings of international conference on trends in computational and cognitive engineering, advances in intelligent systems and computing. Springer, Singapore, pp. 341–354

Huang Y, Guo D, Kasakoff A, Grieve J (2016) Understanding U.S. regional linguistic variation with Twitter data analysis. Comput Environ Urban Syst 59:244–255. https://doi.org/10.1016/j.compenvurbsys.2015.12.003

Jackson DA (1993) Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology 74:2204–2214

Karjus A, Blythe RA, Kirby S, Smith K (2020) Quantifying the dynamics of topical fluctuations in language. Language Dyn Change 10:86–125. https://brill.com/view/journals/ldc/10/1/article-p86_5.xml

Koylu C (2018) Uncovering geo-social semantics from the Twitter Mention Network: an integrated approach using spatial network smoothing and topic modeling. In: Shaw S-L, Sui D (eds) Human dynamics research in smart and connected communities, human dynamics in smart cities. Springer International Publishing, Cham, pp. 163–179

Kramsch C (2014) Language and culture. AILA Rev 27:30–55

Lane J-E, Ersson S (2016) Culture and politics: a comparative approach, 2nd edn. Routledge, London

Lieske J (1993) Regional subcultures of the united states. J Politics 55:888–913. https://doi.org/10.2307/2131941

Louf T (2023a) Word counts per US county in geo-tagged Tweets posted between 2015 and 2021. https://figshare.com/articles/dataset/Word_counts_per_US_county_in_geo-tagged_Tweets_posted_between_2015_and_2021/20630919
Louf T (2023b) Words-use. https://github.com/TLouf/words-use
Mislove A, Lehmann S, Ahn Y-Y, Onnela J-P, Rosenquist JN (2011) Understanding the demographics of Twitter users. In: Proceedings of the international AAAI conference on web and social media, vol 5. AAAI Press, Barcelona, pp. 554–557
Momeni E, Karunasekera S, Goyal P, Lerma, K (2018) Modeling evolution of topics in large-scale temporal text corpora. In: Proceedings of the 12th international AAAI conference on web and social media. Association for the Advancement of Artificial Intelligence, pp. 656–659
Nguyen D, Doğruöz AS, Rosé CP, de Jong F (2016) Computational sociolinguistics: a survey. Comput Linguist 42:537–593. https://doi.org/10.1162/COLI_a_00258
Odum HW (1936) Southern regions of the United States. University of North Carolina Press, Chapel Hill, NC
Ord JK, Getis A (1995) Local spatial autocorrelation statistics: distributional issues and an application. Geogr Anal 27, 286–306
Pavalanathan U, Eisenstein J (2015) Confounds and consequences in geotagged Twitter data. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics (ACL), Lisbon, pp. 2138–2148
Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65
Steiger E, De Albuquerque JP, Zipf A (2015) An advanced systematic literature review on spatiotemporal analyses of Twitter data. Trans GIS 19:809–834
Vanderbeck RM, Dunkley CM (2003) Young people's narratives of rural–urban difference. Child Geogr 1:241–259
Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemometr Intell Lab Syst 2:37–52
Woodard C (2012) American Nations: a history of the eleven rival regional cultures of North America. Penguin Books, New York, NY
Zelinsky W (1973) The cultural geography of the United States. Prentice-Hall, Englewood Cliffs, 1st. ed

Acknowledgements

This work was partially supported by the Spanish State Research Agency (MCIN/AEI/10.13039/501100011033) and FEDER (UE) under project APASOS (PID2021-122256NB-C21 and PID2021-122256NB-C22) and the María de Maeztu project CEX2021-001164-M, by the Comunitat Autonoma de les Illes Balears through the Direcció General de Política Universitària i Recerca with funds from the Tourist Stay Tax Law ITS 2017-006 (PDR2020/51), and by the Arts and Humanities Research Council (UK), the Economic and Social Research Council (UK), Jisc (UK) (Jisc grant reference number 3154), and the Institute of Museum and Library Services (US), as part of the Digging into Data Challenge (Round 3).

Author contributions

All authors designed research and performed the data analyses. TL built the corpus and wrote the code that produced the results. JG and TL drafted the initial manuscript, and all authors were involved in subsequent revisions. JJR, DS, and JG acquired funding.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-023-01611-3.

Correspondence and requests for materials should be addressed to Thomas Louf or David Sánchez.