palgrave communications
HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

Received 20 Nov 2015 | Accepted 24 Mar 2016 | Published 26 Apr 2016

DOI: 10.1057/palcomms.2016.10

OPEN

Race, religion and the city: twitter word frequency patterns reveal dominant demographic dimensions in the United States

Eszter Bokányi[1], Dániel Kondor[1,2], László Dobos[1], Tamás Sebők[1], József Stéger[1], István Csabai[1] and Gábor Vattay[1]

ABSTRACT Recently, numerous approaches have emerged in the social sciences to exploit the opportunities made possible by the vast amounts of data generated by online social networks (OSNs). Having access to information about users on such a scale opens up a range of possibilities—from predicting individuals' demographics and health status to their beliefs and political opinions—all without the limitations associated with often slow and expensive paper-based polls. A question that remains to be satisfactorily addressed, however, is how demography is represented in OSN content—that is, what are the relevant aspects that constitute detectable large-scale patterns in language? Here, we study language use in the United States using a corpus of text compiled from over half a billion geotagged messages from the online microblogging platform Twitter. Our intention is to reveal the most important spatial patterns in language use in an unsupervised manner and relate them to demographics. Our approach is based on Latent Semantic Analysis augmented with the Robust Principal Component Analysis methodology, which permits identification of the data's main sources of variation with an automatic filtering of noise and outliers without influencing results by *a priori* assumptions. We find spatially correlated patterns that can be interpreted based on the words associated with them. The main language features can be related to slang use, urbanization, travel, religion and ethnicity, the patterns of which are shown to correlate plausibly with traditional census data. Apart from the standard measure of linear correlation, some relations seem to be better explained by Boolean implications, suggesting a threshold-like behaviour where demographic variables influence the users' word use. Our findings validate the concept of demography being represented in OSN language use and show that the traits observed are inherently present in the word frequencies without any previous assumptions about the dataset. They therefore could form the basis of further research focusing on the evaluation of demographic data estimation from other big data sources, or on the dynamical processes that result in the patterns identified here.

[1] Department of Physics of Complex Systems, Eötvös Loránd University, Budapest, Hungary [2] SENSEable City Laboratory, Massachusetts Institute of Technology, Cambridge, USA Correspondence: (e-mail: bokanyi@complex.elte.hu)

Introduction

Geography plays an important role in many social phenomena: clearly, many aspects of life are influenced by the possibilities offered by the environment in which one lives (Quillian, 1999; Brain, 2005; Bruch and Mare, 2006; Iceland and Wilkes, 2006; Bettencourt *et al.*, 2007; Sampson, 2009). As such, uncovering the spatial structures and the dynamics of changes in them has for some time been a focus of the scientific community and policymakers. In line with this, governments and local authorities invest significant resources in creating and maintaining databases of census data, including several variables describing the local population and economic activity on the regional scale. These data-collection and monitoring activities are usually limited by the significant efforts required to obtain and process data, prompting researchers and professionals to look for alternative data sources and methods that can complement traditional data collection and that could be integrated with modelling and research efforts (Reades *et al.*, 2007; O'Connor *et al.*, 2010; Cummings *et al.*, 2012; Deville *et al.*, 2014; Frias-Martinez and Frias-Martinez, 2014; Louail *et al.*, 2014; Botta *et al.*, 2015).

In the past two decades, there has been significant growth in the amount of data collected about individuals that has been made available for research purposes. This has had a large impact on social science research where empirical studies were previously limited by the cost and effort associated with data collection. This includes studies focusing on how modern data collection methods can be used to reveal the spatial structure in society on several scales, and how quantities measured in the online or abstract environments are connected to real-world phenomena. Two common data sources are mobile phone networks, where user activity and aggregated measures of network utilization are recorded at the antenna level as part of regular operation (Blondel *et al.*, 2015), and online social networks (OSNs) (Mislove, 2009), where the content publicly shared by users in many cases includes their position (Cheng *et al.*, 2011). Some other data sources with promising application possibilities include monetary transactions (Brockmann *et al.*, 2006; Thiemann *et al.*, 2010; Sobolevsky *et al.*, 2016), GPS traces from cars (Pappalardo *et al.*, 2013, 2015), and other devices and public transportation usage as recorded by electronic payment systems (Roth *et al.*, 2011; Hasan *et al.*, 2013).

Using these data, previous research has shown that it is possible to obtain accurate and up-to-date measures of population density (Deville *et al.*, 2014) or crowd size at sports events or in airports (Botta *et al.*, 2015). Furthermore, the demographic features of a city or a country can be estimated by parsing OSN user names or user profile descriptions (Longley *et al.*, 2015; Sloan *et al.*, 2015). By focusing on the community structure instead of estimating features of individuals, networks of connections among mobile phone or social network users reveal geographic clustering on large scales (Thiemann *et al.*, 2010; Sobolevsky *et al.*, 2013; Kallus *et al.*, 2015). Twitter users' language choice reflects different cultural communities (Mocanu *et al.*, 2013), while user activity has been used on urban scales as an innovative method of land use detection (Reades *et al.*, 2007; Grauwin *et al.*, 2014; Frias-Martinez and Frias-Martinez, 2014; Louail *et al.*, 2014). In addition to land use data, commuting and mobility patterns in the city (González *et al.*, 2008; Jiang *et al.*, 2015) and larger scale travel trends on can also be investigated with the help of mobile and OSN networks (Cho *et al.*, 2011; Simini *et al.*, 2012; Hawelka *et al.*, 2014).

Apart from looking at the spatio-temporal patterns, analysing the users' content posted in OSNs can provide further insights, adapting text mining methods and results that have been previously developed and obtained on the growing corpus of digital texts (Deerwester *et al.*, 1990; Landauer and Dumais, 1997; Petersen *et al.*, 2012; Perc, 2012; Schwartz *et al.*, 2013). From predicting heart-disease rates of an area based on its language use (Eichstaedt *et al.*, 2015), connecting health measures to photo scenicness ratings (Seresinhe *et al.*, 2015) or relating unemployment to social media content (Llorente *et al.*, 2015; Pavlicek and Kristoufek, 2015) to forecasting stock market moves from search semantics (Curme *et al.*, 2014), many studies have attempted to connect online media language and metadata to real-world outcomes. Various studies have analysed spatial variation in the OSN messages' texts and its applicability to several different questions, including user localization based on the content of their posts (Backstrom *et al.*, 2010; Cheng *et al.*, 2010), empirical analysis of the geographic diffusion of novel words, phrases, trends and topics of interest (Ferrara *et al.*, 2013; Eisenstein *et al.*, 2014), measuring public mood (Mitchell *et al.*, 2013).

In these studies, either *a priori* models were used, or a model was built with a *supervised* learning method, with a focus on the specific phenomenon, meaning the exploitation of only one aspect (user name, user profile description, misspelled words, words connected to fatigue and so on), yet possibly neglecting the dataset's other features. While being effective, there remain the following questions: (1) what are main patterns in the data in general; (2) can they be discovered without making *a priori* assumptions about what to look for; (3) can we relate these patterns to relevant real social phenomena.

In this study our goal is to analyse in an *unsupervised* manner how and to what extent regional-scale demographic attributes are represented in social media posts. We approach this using geotagged short messages (*tweets*) posted on the Twitter microblogging service as a source of large-scale digital corpus. We employ a combination of Latent Semantic Analysis (LSA) (Deerwester *et al.*, 1990) and Robust Principal Component Analysis (RPCA) (Lin *et al.*, 2010; Candès *et al.*, 2011), which permits us the automated identification of the most significant topics and language use features with regional variation on Twitter. We use tweets posted in the United States over a 1-year period aggregated at the county level. This allows comparison with census data at the same level, thus allowing us to draw some hypotheses about the driving forces behind regional language dissimilarity patterns.

Methods

Twitter dataset. We use the data stream freely provided by Twitter through their Application Program Interface, which amounts to approximately 1% of all sent messages. In this study, we focus on the part of the data stream with geolocation information. These geolocated tweets originate from users who chose to allow their mobile phones to post the GPS coordinates along with a Twitter message. The total geolocated content was found to only comprise a small percentage of all tweets; therefore with data collection focusing only on these, a large fraction of all geotagged tweets can be gained (Morstatter *et al.*, 2013).

Our dataset includes a total of 335 million tweets from the contiguous United States collected between February 2012 and June 2013. These are all geotagged—that is, they have GPS coordinates associated with them. We construct a geographically indexed database of these tweets, permitting the efficient analysis of regional features (Dobos *et al.*, 2013). Using the Hierarchical Triangular Mesh scheme for practical geographic indexing (Szalay *et al.*, 2007; Kondor *et al.*, 2014), we assigned a US county to each tweet. County borders are obtained from the GAdm database (http://gadm.org).

Latent semantic indexing and RPCA. We aim to use a type of vector space model on our Twitter corpus, where documents correspond to county-level aggregated tweets. The terms we consider are raw words obtained after a tokenization process—that is, we apply a "word-bag" approach to our documents, effectively limiting any analysis to word frequencies and ignoring relations among words and longer phrases. We filter stop-words in several languages (most important being English and Spanish) to remove most common but uninformative terms from our data.

We construct a term-document matrix W_{ij} as the number of occurrences of the i-th word in the j-th cell. As the population density of the United States is very heterogeneous, the number of word occurrences in each county is also heterogeneous. To improve the quality of the dataset, we only include counties that contain at least 10,000 occurrences of at least 500 individual words. We also limit the words used to those with at least 10,000 occurrences in at least 1,000 individual counties. This way there remain 2,800 counties and 10,132 words, which

form the W_{ij} word occurrence matrix. We normalize W_{ij} so that the elements are the relative frequencies of words in each county: $X_{ij} \equiv W_{ij}/\sum_k W_{kj}$—that is, we normalize each element by the total number of words posted in that county; this is called inverse document frequency weighing in text-mining literature.

To identify all possible regional characteristics of language usage, we rely on techniques known from the field of natural language processing. There exist many feature or topic extraction methods, all of them aiming to reduce the dimensionality of the data by finding related or similar words and documents. A common approach is LSA (Deerwester *et al.*, 1990; Gotoh and Renals, 1997), which applies Singular Vector Decomposition (SVD) on a word by document matrix derived from the corpus. This method groups words together based on their semantic similarity (Landauer and Dumais, 1997), creating "feature" documents, of which the first few represent the concepts causing the most variation in the data. A notable achievement of LSA is that it is an unsupervised learning method, thus providing information about the corpus without using *a priori* assumptions or any arbitrary preselections based on the purpose of the examination.

According to the nature of our dataset, there are several users who generate automated messages like weather stations, advertisers or tornado and earthquake advisories, which are considered as noise in our investigations. Especially in sparsely inhabited areas, these outlier messages can account for a large fraction of the dataset. Also, highly localized features, such as tourist attractions, can generate outliers of significant volume. This can result in highly localized outliers dominating the results of the SVD, making identifying relevant structure challenging. Applying the Robust PCA method (Lin *et al.*, 2010; Candès *et al.*, 2011) allows us to preprocess the matrix before further analysis by separating it into a low-rank and a sparse part, whose principal components can then be computed and analysed separately. This means that the original data matrix is written as a sum of two parts:

$$X = X^S + X^{LR},$$

where X^S is a sparse matrix and X^{LR} contains the dense but low-rank part of the data. The mathematical condition for finding X^S and X^{LR} is minimizing the sum

$$\lambda\|X^S\|_1 + \|X^{LR}\|_\sigma,$$

where for a matrix X of dimensions $n_1 \times n_2$ with $n_1 \geq n_2$, $\lambda \equiv 1/\sqrt{n_1}$, and the norms are the l_1 and nuclear norms, respectively:

$$\|X\|_1 = \sum_{ij}|X_{ij}| \|X\|_\sigma = \sum_i \sigma_i(X).$$

Here, $\sigma_i(X)$ denotes the i-th singular value of X. An efficient algorithm for finding X^S and X^{LR} is the inexact augmented Lagrangian method (Lin *et al.*, 2010) (Matlab code developed by the authors of Lin *et al.*, 2010 implementing the algorithm is publicly available; http://perception.csl.illinois.edu/matrix-rank/sample_code.html). Employing this method results in the sparse part containing most of the outliers, and in true language use variations to be represented in the low-rank part. Due to the structure of our data matrix, and the employed Robust PCA method, we choose not to subtract averages from the data; of course, this will probably result in average word frequencies dominating the first principal component. We further analyse only the results of the LSA of the low-rank component.

Demographic data. To discover possible governing factors of the geographical language variation patterns and connections between topics and their geography, we correlate right singular vectors with a variety of demographic data series from the 2010 US Census (http://www2.census.gov/census_2010/, http://www.census.gov/support/USACdataDownloads.html), 2011 American Community Survey 5-year estimates concerning educational attainment by counties (http://www.census.gov/programs-surveys/acs/), county business patterns according to North American Industry Classification System classification (http://www.census.gov/econ/cbp/), and church adherence rates and congregations numbers per county provided by the Association of Religion Data Archives (http://www.thearda.com).

Boolean relationship detection. Apart from evaluating linear correlation measures with the singular vectors, we also carry out a Boolean relationship detection, using the methodology of Sahoo *et al.* (2008), which is based on calculating a test statistic based on the contingency table of the scatter plots (for example, Fig. 3f–j, see the next section for an interpretation of the results displayed) after creating the four segments of the data with a horizontal and a vertical limit. We find the most significantly sparse segment by setting the limits so that the test statistic gives a maximum for the specific segment. During the calculations, we set an error bar on both side of the limits, and points being in this error zone are not taken into consideration when testing for the sparseness.

If the contingency table is

	A low	*A high*	$\sum$
B low	m_{00}	m_{01}	b_0
B high	m_{10}	m_{11}	b_1
$\sum$	a_0	a_1	s

then the test statistic for the four segments is

$$\delta = \frac{m_{ij} - \langle m_{ij}\rangle}{\sqrt{\langle m_{ij}\rangle}},$$

where $\langle m_{ij}\rangle$ denotes the expected value in case of independent variables

$$\langle m_{ij}\rangle = \frac{a_i b_j}{s\ s}\cdot s.$$

If there are some points left in the segment, they are considered as an error, and the measure of error would be

$$\epsilon = \frac{1}{2}\left(\frac{m_{ij}}{m_{i0} - m_{i1}} + \frac{m_{ij}}{a_i}\right).$$

We consider a segment significantly sparse if $\delta > 3$, and $\epsilon < 0.2$.

Then in the whole range of variables A and B (using 100 steps in both directions and an error boundary of 1.5% for the skipping of points near the borders) we measure δ and ϵ values, and take the segmentation with the maximum δ for the sparse areas, where ϵ is still low enough.

Results

Using a corpus of over 335 million geotagged tweets posted in the United States, we compile word-frequency distributions for each US county, and then apply the automatic filtering and feature selection method described. We analyse the features found with this technique by considering the connection between geographic and semantic distances (Fig. 1), and by plotting right singular vectors on the map (Figs. 2a–e) and displaying left singular vectors as word clouds (positive weights Figs. 2f–j, negative weights Figs. 2k–o).

First, we find that the method applied successfully uncovers some coherent topics, especially in the first few singular vectors, where singular values are still great enough for the topic to give a significant variance of the dataset. As we deliberately choose not to subtract averages from the X_{ij} matrix, the first component shows no discernible pattern, and corresponds to the most common words in the sample. From the second singular vector, however, one or both ends (it can be either negative or positive, as singular vectors can arbitrarily be multiplied by a minus sign) of each of the most important semantic features on the word clouds can be related to a certain language style, concept or lifestyle.

The words giving the largest contribution to the pattern of the second left singular vector (Fig. 2f) mark a strong presence of slang in the sample. This includes forms with alternate spelling like "aint", "gotta"; swearing like "ass", "hoe", "bitch"; abbreviations of common phrases like "tryna", "imam", "kno", "yall";

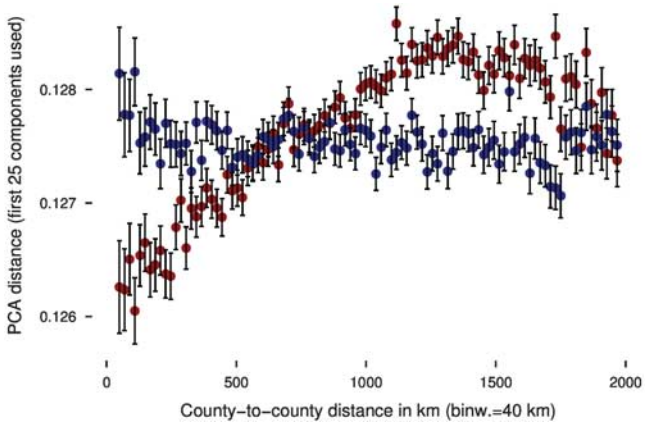

Figure 1 | Semantic versus real-world distance of counties. Euclidean distance of counties in the semantic subspace of the first 25 components obtained from LSA as a function of geographical distance (red dots). Baseline calculated from a random permutation of counties (blue dots). Error bars correspond to the error of the binwise means.

Figure 2 | Representations of the first few right (a–e) and left (f–o) singular vectors. Weights corresponding to counties are plotted on a US map, brown representing the negatively, blue the positively weighed counties. Word clouds represent left singular vectors with colouring corresponding to that of the maps, and word size representing the weight of each word in each actual singular vector.

OSN-specific slang such as "oomf", which stands for "one of my followers" (that is, on Twitter); a very specific misspelling of "goodmorning" (instead of "good morning"); and variations of the racial slurs "nigga" and "niggas". Swear words and abbreviations typical for online language also dominate this end of the component. The next most important feature, which can be found in the third vector (Fig. 2l), identifies words connected to urban lifestyle like eating out ("pizza", "grill"), drinking coffee ("coffee", "cafe", "starbucks"), education ("university", "library", "campus") or working out ("gym", "fitness").

Further dominating concepts are travel ("enjoying", "trip", "pic", "hotel") in the fourth singular vector (Fig. 2h) and religion ("lord", "prayers", "praying", "blessed") alongside with positive content ("glad", "thankful", "wonderful", "proud") in the negatively weighed words of the fifth singular vector (Fig. 2n). In this case, the opposite end can also be easily interpreted: the faith-related words in the fifth component are countered by an increased usage of profanity present among words with positive weights (Fig. 2i). This might be the consequence of people tweeting about religious topics also trying to avoid swearing; this hypothesis can also be supported with less strong swearing alternatives ("crap", "freaking", "dang") prevailing among the negatively weighed words along the religious words.

If the native language of a group is different from that of the majority, the words of this different language also stand out from the overall structure, as there is naturally a stronger correlation among words belonging to the same language. Therefore, the method applied can discover languages different from that of the bulk of the sample. In the sixth singular vector, we can observe this phenomenon with Spanish words, which form more than the third of the positively weighed word cloud (Fig. 2j). The English terms "Mexico" and "Mexican" also appear in this group, which shows that concepts related to the topic are also identified even if they do not belong to the discovered language.

Similarly to topic identification, where semantically close words form topics, analysing regional patterns reveal documents that are close to each other in the semantic space spanned by these topics. Plotting the right singular vectors on a map (Figs. 2a–e), the most striking feature is the regional proximity of documents having close weights in the singular vectors. Document-by-document (county-by-county) Euclidean distances in the PCA subspace of the first 25 component as the function of real county-by-county centroid distances (http://cta.ornl.gov/transnet/SkimTree.htm) illustrate this observation. In Fig. 1 mean PCA subspace distances (red dots) are plotted for each 40 km range of real county centroid distances. As a baseline, the same is done for a random permutation of counties (blue dots). It is remarkable that below 500 km, counties are closer in the semantic space, as could be expected from a random realization. From 700 km to 1800 km, semantic distance is greater than it would be randomly. Geographical proximity is thus a main driving force in the similarity of language patterns in Twitter-space.

Analysing these geographical patterns in each singular vectors provides insights into the regional distribution of the single topics. On a US map, the second component (Fig. 2a), which is responsible for the most variance in the Twitter data, emerges as a block in the Southeastern part of the United States. Apart from the big Southeastern block, Chicago and Detroit are also marked by this pattern of language usage. In the third component (Fig. 2b), negative weights (brown patches) mark the biggest cities and surrounding counties that belong to their agglomeration. The most positive pattern of the fourth component (Fig. 2c) reveals some important touristic attractions such as the centre of New York, Washington and San Francisco, the Craters of the Moon National Monument and Preserve in Idaho, Aspen Mountain ski area in Colorado or Hawaii. The regional pattern of the fifth component (Fig. 2d) is less obvious, though a part of the central US and the Southeastern block is discernible in the religion-related end of the component. The sixth component distinguishes the Southwestern part and the Northwestern corner of the United States (Fig. 2e), Florida and some bigger cities such as New York or Chicago.

To discover possible governing factors of the geographical language variation patterns and their relation to demography, we calculate Pearson correlation values between right singular vectors and data obtained from the US Census Bureau described in the section "Demographic data". Data series that have the greatest absolute correlation values ($P<0.0001$, Bonferroni-corrected) with each component are shown in Table 1. The large correlation (0.872) of the second component with the population proportion of African-Americans per county indicates that the observed slang words and the blockwise regional pattern are linked to the presence of this demographic group (note that, however, we have no evidence of whether the tweets causing the variation were indeed posted by African-American people). Figure 3a shows the census proportions on a US map, with the regional pattern approximately corresponding to that of the singular vector. It is worth noting that apart from the large southeastern block, Chicago and Detroit are also marked by having the characteristic slang word pattern, as well as a higher proportion of African-American population. A similarly large correlation (0.500) with ethnicity (Hispanic or Latino origin) also arises is the case of the sixth component, as expected from the observed Spanish words and the Southwestern positive weights on the map. Figure 3 shows the percentage of people with Hispanic or Latino origin in US counties, the distribution resembling that of the right singular vector.

The data series that show the largest correlation with the third component are resident total population rank (0.844) and rural–urban continuum code (http://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx) (0.630). Since neither are continuous variables, we instead show population density values in each county on the map of Fig. 3b. Densely populated areas mark the biggest cities and their surrounding agglomerations of the United States, and these areas are also discernible in the brown patches of the third singular vector in Fig. 2b. It confirms the idea of the most densely populated areas giving the negative end of the third singular vector in both the words and their regional distribution. A basic feature of the Twitter corpus is thus linked simply to city lifestyle, more generally to the associated socioeconomic status.

Correlation values show whether there exists some relation between the language patterns and demographic data (Table 1). Analysing scatter plots of the greatest correlations provides some insight into the structure of these relations. Plotting the regional weights of the second and sixth singular vector against African-American and Hispanic or Latino ethnicity percentages exhibits very similar features (Figs. 3f, j). Correlation analysis also revealed that a prevalence of evangelical religious groups (Baptists and Methodists) is related to (−0.372) the religious content of the fifth component (Fig. 3i); county-level rates of adherence of evangelical churches are plotted in Fig. 3d. The existence of a virtual "Bible Belt" is thus confirmed in the Twitter-space, corresponding to former identification of religious groups in cyberspaces (Zook and Graham, 2010; Shelton *et al.*, 2012). An opposite correlation is present with Catholic and Orthodox churches, which we speculate to be the consequence of these having a smaller attendance in counties where evangelical churches predominate.

Although almost all of the above-described correlations could be explained by an underlying function, a Boolean implication

Table 1 | Correlations with demographic data series

No. of right singular vector	*ρ*	*Dataset*
2	0.87	Population of one race—Percent Black or African American alone (2010)
	0.77	Owner-occupied housing units, African American householder, per population (2010)
	−0.75	Population of one race—Percent White alone (2010)
	0.65	Black Protestant—Rates of adherence per 1,000 population (2010)
3	0.84	Resident total population estimate—Rank (2007)
	−0.72	Population density (2010)
	−0.72	Percent of adults with a bachelor's degree or higher (2008–2012)
	0.63	Rural-urban Continuum Code (2013)
4	0.28	All other travel arrangement and reservation services total number of establishments
	0.28	Tour operators total number of establishments
	0.26	Convention and visitors bureaus total number of establishments
	0.26	Accommodation establishments with payroll per population (2007)
5	0.39	Catholic—Rates of adherence per 1,000 population (2010)
	−0.37	Evangelical Protestant—Rates of adherence per 1,000 population (2010)
	0.36	Orthodox—Total number of adherents (2010)
	−0.29	Evangelical Protestant—Total number of adherents per population (2010)
6	0.5	Percent Hispanic or Latino population (2010)
	0.5	Hispanic or Latino population—Percent Mexican (2010)
	0.38	Average household size (2010)
	0.36	Percent households with persons under 18 years (2010)

Note: Greatest Pearson-correlations ($P<0.0001$ at a Bonferroni-corrected level) in the demographic datasets with the first few singular vectors.

model description seems more plausible. Boolean implications have been used in gene expression research (Sahoo *et al.*, 2008) to uncover non-symmetric relationships where correlation analysis would only partially or not at all measure connection between two variables. In the case of ethnicities, if we take y values as a measure of how strongly slang (Fig. 3f) or Spanish (Fig. 3j) (see the word clouds of Fig. 2f and Fig. 2j) is present in the Twitter messages of the counties, we can observe that below a certain ratio of ethnicity prevalence (6.0% in the second component and 7.6% in the sixth) language patterns show different levels of non-slang or non-Spanish usage. If ethnicity prevalence is greater than the threshold value, slang or Spanish usage rises steeply with growing ethnicity proportion. Above the threshold, there are very few counties with non-slang or non-Spanish language patterns. In this terminology, the two scatter plots corresponding to ethnicity prevalence can be translated to "high ethnicity rates ⇒ missing non-slang/non-Spanish" words implication. The limits corresponding to the best implication model were the mentioned 5.99 ± 1.28 and 7.65 ± 1.43 of prevalence for the two ethnic groups, with -0.00328 ± 0.00123 and -0.00277 ± 0.00209 as a limit on the axes of the second and sixth component. The measures of sparseness for the lower right segments were $\delta = 21.941$, $\epsilon = 0.036$, and $\delta = 12.98$, $\epsilon = 0.15$, respectively.

A Boolean implication also describes the scatter plot of the fifth component measured against evangelical adherence rates (Fig. 3i). Here the y axis represents a level of swearing (cf. the words in Fig. 2i) present in the Twitter-sphere of tweets posted in a county. Thus the implication can be translated to "high evangelical prevalence rates ⇒ low swearing level". The pattern implies a stronger connection between the two variables, as could be inferred from the symmetric correlation measure. It seems as if above a certain adherence rate, a text with a high swearing level could not propagate further or could not find way to broader discussion. Here, the automatically detected limit was at a 19.67 ± 1.55 adherence level, the limit on the "swearing" axis lies at 0.02230 ± 0.00177, and the lower right corner showed a significantly large sparseness with $\delta = 8.579$ and $\epsilon = 0.013$.

Discussion

We can conclude that the applied unsupervised learning method successfully discovers topics and their regional patterns in the Twitter-sphere, with county weights in right singular vectors representing a distance in the semantic space along a topic given by word weights of the left singular vectors. It is also remarkable that geographical closeness implies closeness in the semantic space, which suggests that language usage is on a certain level bound to geographical proximity.

We also find that regional patterns in language use are driven not just by geographical proximity, but socioeconomical and cultural similarities, like degree of urbanization, religion or ethnicity. It seems that the most important factor behind the variation in the language use of different counties is the presence of Afro-American ethnicity, as confirmed by the significant correlation between the census-based share of Afro-American population and the appropriate county weights. Corresponding word weights mirror this observation with words representative of the typical slang use associated with this ethnicity. This type of slang use thus turns out to be the most distinguishing factor in everyday US Twitter conversation.

Following ethnicity, the second most important feature found in Twitter language is related to the population density of a county. The interpretation could be that beyond ethnicity, our everyday language is largely influenced by our surroundings. Therefore, living in densely populated places, which means mostly urban areas, results in words specific to urban lifestyle appearing more frequently in user messages.

The language footprints of tourism can also be captured by our method, suggesting that the effect of messages or users being on a holiday should always be considered when trying to relate online content to real-world phenomena.

Some relations are better described by a non-symmetric Boolean implication model instead of the symmetric correlation measure. We find that the presence of ethnic groups above a certain threshold implies a weight greater than a certain level along the semantic axis corresponding to the component connected to this ethnic group. We also find that counties

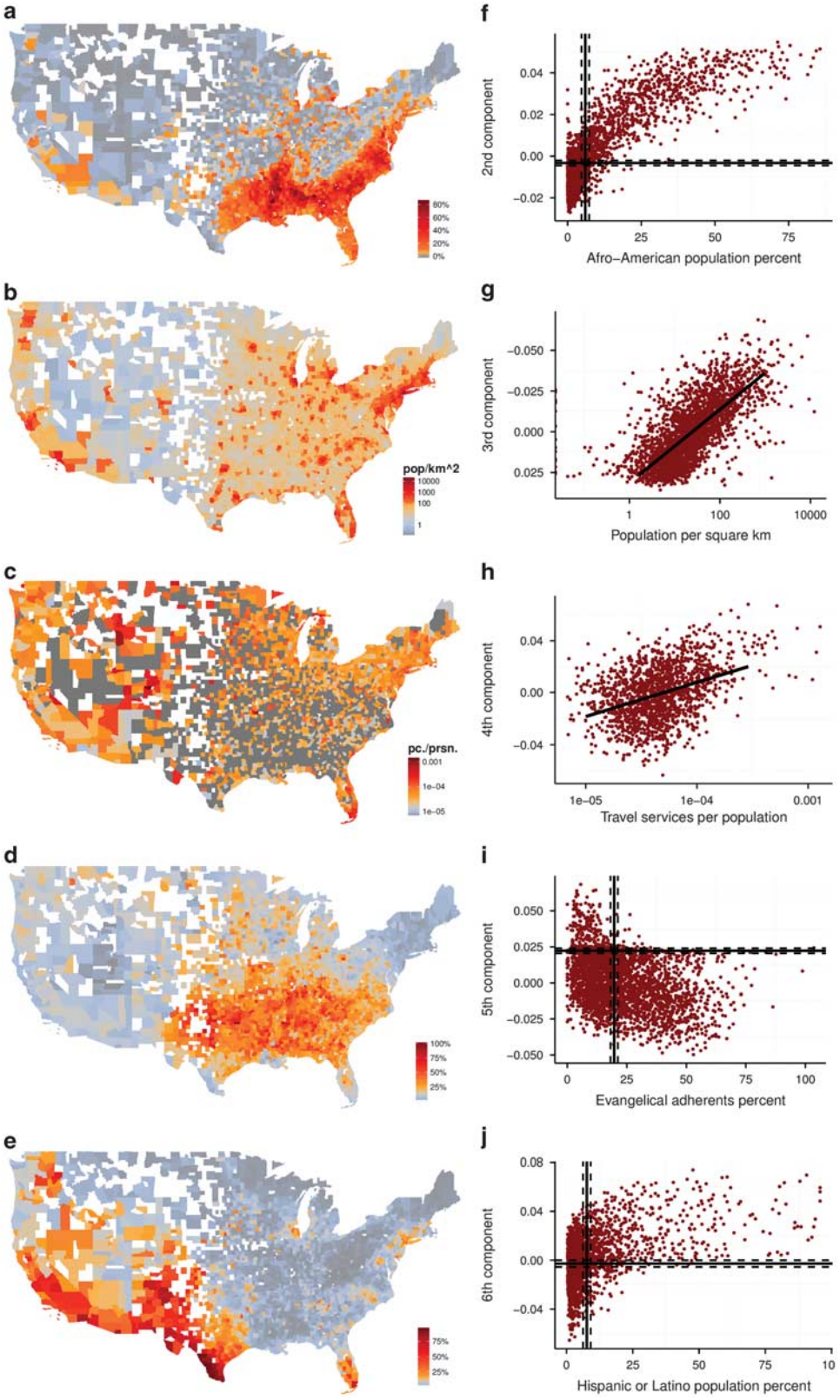

Figure 3 | Census data with most correlation for each component displayed on a map (a–e) and scatter plots of these data series with county weights from corresponding singular vectors (f–j). Lines representing a symmetric relationship (Pearson correlation) are drawn on scatter plots (g) and (h). On scatter plots (f), (i) and (j), horizontal and vertical lines correspond to the best segmentation when testing for a Boolean relationship between variables. The points between the dashed lines were not taken into account when calculating the test statistics for the sparseness of each segment.

exhibiting high evangelical adherence rates show low level on the "swearing scale" given by the corresponding component. This is interesting, as the phenomenon cannot be observed with the two other major denominations, the Catholic and Orthodox churches. It suggests that the online presence of Evangelical churches is inherently different from that of the other denominations, and its adherents have a significant effect on the word choice on the Twitter platform.

Our results suggest that OSN activity can be used effectively to monitor the spatial variation of cultural traits as represented in language use, yielding an up-to-date picture of important social phenomena. We believe our present study demonstrates an approach for measuring the importance of certain demographic attitudes when working with textual Twitter data. We suggest, therefore, that it could form the basis of further research focusing on the evaluation of demographic data estimation from other sources, or on the dynamical processes that result in the patterns found here. While our results were obtained using the Twitter microblogging platform, research could be further extended to investigate whether the incorporation of other metadata (for example, user activity, user mobility, user profile descriptions and so on) or the analysis of different text sources could refine or enhance our findings.

References

Backstrom L, Sun E and Marlow C (2010) Find me if you can: Improving geographical prediction with social and spatial proximity. *Proceedings of the 19th international conference on World wide web*. ACM, pp 61–70, http://dl.acm.org/citation.cfm?id=1772698.

Bettencourt LM, Lobo J, Helbing D, Kühnert C and West GB (2007) Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences of the United States of America*; **104** (17): 7301–7306.

Blondel VD, Decuyper A and Krings G (2015) A survey of results on mobile phone datasets analysis. *EPJ Data Science*; **4** (1): 10.

Bokányi E (2016) Replication data for: Race, religion and the city: Twitter word frequency patterns reveal dominant demographic dimensions in the United States. Dataverse, http://dx.doi.org/10.7910/DVN/EXWJRJ.

Botta F, Moat HS and Preis T (2015) Quantifying crowd size with mobile phone and Twitter data. *Royal Society Open Science*; **2** (5): 150162.

Brain D (2005) From good neighborhoods to sustainable cities: Social science and the social agenda of the new urbanism. *International Regional Science Review*; **28** (2): 217–238.

Brockmann D, Hufnagel L and Geisel T (2006) The scaling laws of human travel. *Nature*; **439** (7075): 462–465.

Bruch EE and Mare RD (2006) Neighborhood choice and neighborhood change. *American Journal of Sociology*; **112** (3): 667–709.

Candès EJ, Li X, Ma Y and Wright J (2011) Robust principal component analysis? *Journal of the ACM (JACM)*; **58** (3): 11.

Cheng Z, Caverlee J and Lee K (2010) You are where you tweet: a contentbased approach to geo-locating Twitter users. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp 759–768, http://dl.acm.org/citation.cfm?id=1871535.

Cheng Z, Caverlee J, Lee K and Sui DZ (2011) Exploring millions of footprints in location sharing services. International AAAI Conference on Web and Social Media; pp 81–88. AAAI Press.

Cho E, Myers S and Leskovec J (2011) Friendship and mobility: User movement in location-based social networks. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 1082–1090. http://dl.acm.org/citation.cfm?id = 2020579.

Cummings D, Oh H and Wang N (2012) *Who Needs Polls? Gauging Public Opinion from Twitter Data*.

Curme C, Preis T, Stanley HE and Moat HS (2014) Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences*, **111** (32): 11600–11605.

Deerwester S, Dumais S and Landauer T (1990) Indexing by latent semantic analysis. *Journal of the American Society for Information Science*; **41** (6164): 391.

Deville P, Linard C, Martin S, Gilbert M, Stevens FR and Gaughan AE (2014) Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*; **111** (45): 15888–15893.

Dobos L *et al* (2013) A multi-terabyte relational database for geo-tagged social network data. *4th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2013—Proceedings*, pp 289–294. IEEE.

Eichstaedt JC *et al* (2015) Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*; **26** (2): 159–169.

Eisenstein J, O'Connor B, Smith NA and Xing EP (2014) Diffusion of lexical change in social media. *PLoS ONE*; **9** (11): e113114.

Ferrara E, Varol O, Menczer F and Flammini A (2013) Traveling trends: Social butteries or frequent iers? *COSN '13 Proceedings of the First ACM Conference on Online Social Networks*; pp 213–222. http://dl.acm.org/citation.cfm?id=2512956.

Frias-Martinez V and Frias-Martinez E (2014) Spectral clustering for sensing urban land use using Twitter activity. *Engineering Applications of Artificial Intelligence*; **35** (10): 237–245.

González MC, Hidalgo CA and Barabási A-L (2008) Understanding individual human mobility patterns. *Nature*; **453** (7196): 779–782.

Gotoh Y and Renals S (1997) Document space models using latent semantic analysis. *Proc. Eurospeech*, pp 1443–1446, http://scholar .google.com/scholar?hl=en&btnG=Search&q=intitle:No+Title%5C#0.

Grauwin S, Sobolevsky S, Moritz S, Gódor I and Ratti C (2014) Towards a comparative science of cities: Using mobile trac records in NewYork, London and Hong Kong. In: Helbich M, Jokar Arsanjani Jl and Leitner M (eds) *Computational Approaches for Urban Environments*. Geotechnologies and the Environment, Vol. 13, pp. 363–387. Springer International Publishing. http://arxiv.org/abs/1406.4400.

Hasan S, Schneider C, Ukkusuri S and González M (2013) Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*; **151** (1/2): 304–318.

Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P and Ratti C (2014) Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*; **41** (3): 260–271.

Iceland J and Wilkes R (2006) Does socioeconomic status matter? Race, class, and residential segregation. *Social Problems*; **53** (2): 248–273.

Jiang S, Ferreira J and González MC (2015) Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. International Workshop on Urban Computing

Kallus Z, Barankai N, Szüle J and Vattay G (2015) Spatial fingerprints of community structure in human interaction network for an extensive set of large-scale regions. *PloS ONE*; **10** (5): e0126713.

Kondor D *et al* (2014) Ecient classication of billions of points into complex geographic regions using Hierarchical Triangular Mesh. *Proceedings of the 26th International Conference on Scientific and Statistical Database Management—SSDBM '14*. ACM Press: New York, USA, pp 1–4.

Landauer T and Dumais S (1997) A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*; **1** (2): 211–240.

Lin Z, Chen M and Ma Y (2010) The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices, http://arxiv.org/abs/1009.5055v3.

Llorente A, Cebrian M and Moro E (2015) "Social media ngerprints of unemployment". PLoS ONE; 10(5): e0128692.

Longley PA, Adnan M and Lansley G (2015) The geotemporal demographics of Twitter usage. *Environment and Planning A*; **47** (2): 465–484.

Louail T *et al* (2014) From mobile phone data to the spatial structure of cities. *Scientific Reports*; **4**: 5276.

Mislove A (2009) Online social networks: Measurement, analysis, and applications to distributed information systems. PhD thesis, Rice University. http://www.ccs.neu.edu/home/amislove/publications/SocialNetworks-Thesis.pdf.

Mitchell L, Frank MR, Harris KD, Dodds PS and Danforth CM (2013) The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PloS ONE*; **8** (5): e64417.

Mocanu D, Baronchelli A, Perra N, Gonçalves B, Zhang Q and Vespignani A (2013) The Twitter of Babel: Mapping world languages through microblogging platforms. *PLoS ONE*; **8** (4): e61981.

Morstatter F, Pfeffer J, Liu H and Carley K (2013) Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. International Conference on Weblogs and Social Media, pp 400–408.

O'Connor B, Balasubramanyan R, Routledge BR and Smith NA (2010) From tweets to polls: Linking text sentiment to public opinion time series. International Conference on Weblogs and Social Media. Vol. **11**: 122–129, pp. 1–2.

Pappalardo L, Rinzivillo S, Qu Z, Pedreschi D and Giannotti F (2013) Understanding the patterns of car travel. *The European Physical Journal Special Topics*; **215** (1): 61–73.

Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F and Barabási A-L (2015) Returners and explorers dichotomy in human mobility. *Nature Communications*; **6**: 8166.

Pavlicek J and Kristoufek L (2015) Nowcasting unemployment rates with Google searches: Evidence from the Visegrad Group countries. *PloS ONE*; **10** (5): e0127084.

Perc M (2012) Evolution of the most common English words and phrases over the centuries. *Journal of the Royal Society Interface*; **9** (July): 3323–3328.

Petersen AM, Tenenbaum JN, Havlin S, Stanley HE and Perc M (2012) Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientic Reports* 2: 943

Quillian L (1999) Migration patterns and the growth of high-poverty neighborhoods, 1970–1990. *American Journal of Sociology*; **105** (1): 1–37.

Reades J, Calabrese F, Sevtsuk A and Ratti C (2007) Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*; **6** (3): 30–38.

Roth C, Kang SM, Batty M and Barthélemy M (2011) Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS ONE*; **6** (1): e15923.
Sahoo D, Dill DL, Gentles AJ, Tibshirani R and Plevritis SK (2008) Boolean implication networks derived from large scale whole genome microarray datasets. *Genome Biology*; **9** (10): R157.
Sampson RJ (2009) Disparity and diversity in the contemporary city: Social (dis)order revisited. *The British Journal of Sociology*; **60** (1): 1–31.
Schwartz HA *et al* (2013) Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS ONE*; **8** (9): e73791.
Seresinhe CI, Preis T and Moat HS (2015) Quantifying the impact of scenic environments on health. *Scientic Reports* **5**: 16899.
Shelton T, Zook M and Graham M (2012) The technology of religion: Mapping religious cyberscapes. *The Professional Geographer*; **64** (4): 602–617.
Simini F, González MC, Maritan A and Barabási A-L (2012) A universal model for mobility and migration patterns. *Nature*; **484** (7392): 96–100.
Sloan L, Morgan J, Burnap P and Williams M (2015) Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS ONE*; **10** (3): e0115545.
Sobolevsky S, Szell M, Campari R, Couronné T, Smoreda Z and Ratti C (2013) Delineating geographical regions with networks of human interactions in an extensive set of countries. *PLoS ONE*; **8** (12): e81707.
Sobolevsky S, Sitko I, Tachet R, Arias JM and Ratti C (2016). Cities through the prism of people's spending behavior. PLoS ONE ; **11** (2): e0146291.
Szalay AS, Gray J, Fekete G and Kunszt PZ (2007) Indexing the sphere with the Hierarchical Triangular Mesh. http://arxiv.org/abs/cs/0701164.
Thiemann C, Theis F, Grady D, Brune R and Brockmann D (2010) The structure of borders in a small world. *PloS ONE*; **5** (11): e15422.
Zook M and Graham M (2010) Featured graphic: The virtual 'bible belt'. *Environment and Planning A*; **42** (4): 763–764.

Data Availability

Owing to Twitter's policy we cannot publicly share the original dataset used in this analysis. The county-wide word frequency matrix and the results of the LSA compiled are available in the Dataverse repository (Bokányi, 2016) at http://dx.doi.org/10.7910/DVN/EXWJRJ and also at http://www.vo.elte.hu/papers/2016/twitter-pca.

Acknowledgements

The authors would like to thank the partial support of the European Union and the European Social Fund through the FuturICT.hu project (Grant No.: TAMOP-4.2.2.C-11/1/KONV-2012-0013), the OTKA-103244, OTKA-114560, Ericsson and the MAKOG Foundation.

Additional Information

Competing interests: The authors declares no competing financial interests.

Reprints and permission information is available at http://www.palgrave-journals.com/pal/authors/rights_and_permissions.html

How to cite this article: Bokányi E *et al* (2016) Race, religion and the city: Twitter word frequency patterns reveal dominant demographic dimensions in the United States. *Palgrave Communications*. 2:16010 doi: 10.1057/palcomms.2016.10.