Computers, Environment and Urban Systems 59 (2016) 244–255

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Understanding U.S. regional linguistic variation with Twitter data analysis

Yuan Huang [a], Diansheng Guo [a,*], Alice Kasakoff [a], Jack Grieve [b]

[a] Department of Geography, University of South Carolina, United States
[b] School of Languages and Social Sciences, Aston University, United Kingdom

ARTICLE INFO

Article history:
Received 28 September 2015
Received in revised form 9 December 2015
Accepted 15 December 2015
Available online 31 December 2015

Keywords:
Social media
Linguistic
Twitter
American dialects
Regionalization
US regions
Spatial data mining

ABSTRACT

We analyze a Big Data set of geo-tagged tweets for a year (Oct. 2013–Oct. 2014) to understand the regional linguistic variation in the U.S. Prior work on regional linguistic variations usually took a long time to collect data and focused on either rural or urban areas. Geo-tagged Twitter data offers an unprecedented database with rich linguistic representation of fine spatiotemporal resolution and continuity. From the one-year Twitter corpus, we extract lexical characteristics for twitter users by summarizing the frequencies of a set of lexical alternations that each user has used. We spatially aggregate and smooth each lexical characteristic to derive county-based linguistic variables, from which orthogonal dimensions are extracted using the principal component analysis (PCA). Finally a regionalization method is used to discover hierarchical dialect regions using the PCA components. The regionalization results reveal interesting linguistic regional variations in the U.S. The discovered regions not only confirm past research findings in the literature but also provide new insights and a more detailed understanding of very recent linguistic patterns in the U.S.

1. Introduction

Dialects are forms or varieties of language that belong to a specific region or social group (Chambers & Trudgill, 1998). Research in dialectology not only seeks to understand language differences, language innovations and language variations through time and space, but also helps reveal patterns of information diffusion and cultural interpenetration (Di Nunzio, 2013). Most research on dialects relies on surveys and interviews, which may not contain enough information to identify regional linguistic variations objectively due to the small sample size and lack of computational statistical methods (Grieve, 2009). For example, the recent nationwide linguistic research, described in the *Atlas of North American English*, only contains 762 surveys (individuals) for 297 urban areas (Labov, Ash, & Boberg, 2006). Grieve (2009) introduced quantitative spatial autocorrelation statistics as well as using corpora of natural language data to dialectology. Grieve et al. (2011, 2013) also analyzed regional linguistic variation in American English based on a 26-million-word corpus of letters to editors and the data from Labov et al. (2006); however, neither data set captures linguistic variation in rural areas.

In this research, we use geo-tagged Twitter data as an alternative linguistic database, which can offer spatial and temporal continuity, granularity and up-to-date dynamics for linguistic studies. We present a linguistic study using a one-year dataset of geo-tagged tweets in the continental U.S. (48 states and Washington D.C.), from Oct. 7, 2013 to Oct. 6, 2014, which contains 6.6 million unique Twitter users, 924 million geo-tagged tweets, and 7.8 billion words.

Dialect variations can be examined by differences in lexicon, phonology, grammar, and pragmatics (Wolfram & Schilling-Estes, 2005). However, it is infeasible to attempt to study all linguistic variables that characterize dialects. Therefore, dialect studies often use representative sets of linguistic variables, which may include lexical (Grieve et al., 2011; Kurath, 1949), phonetic and phonological (Labov, Ash, & Boberg, 2006; O'Cain, 1979), and grammatical variation (Atwood, 1953). For this study, we use *lexical alternations* to examine linguistic variations and use counties in the U.S. as the unit for spatial analysis of regional linguistic variations.

In this research, we address two important questions: How do linguistic characteristics vary from place to place based on geo-tagged Twitter data and what are the linguistic regions and sub-regions in the U.S.? Twitter data not only offers spatial–temporal continuity but also allows close examination of a language in its *casual* expressions. Our data has 7.8 billion words and 6.6 million Twitter users, which is much larger than those being used in previous studies. We try to answer the above two questions based on the regional patterns generated by

* Corresponding author at: Department of Geography, University of South Carolina, 709 Bull Street, Room 127 , Columbia, SC, 29208, United States.
E-mail address: guod@mailbox.sc.edu (D. Guo).

http://dx.doi.org/10.1016/j.compenvurbsys.2015.12.003
0198-9715/

each single variable, as well as the aggregated regional patterns. Adaptive kernel smoothing is used to estimate unknown values and to reduce noise. A hierarchical regionalization method is used to discover dialect regions with the top PCA components of linguistic variables extracted from tweets. The regionalization results reveal interesting linguistic regional variations in the U.S. and each region can also have sub-regions of local linguistic characteristics.

2. Background

The traditional way to collect dialect variation was to send out fieldworkers to collect linguistic related transcriptions from selected communities and representative speakers (McDavid, McDavid, Kretzschmar, Lerud, & Ratliff, 1986). One representative survey was conducted by Hans Kurath (1949) who proposed a plan for a Linguistic Atlas of the United States and Canada, which set the foundation of the project *Linguistic Atlas of Middle and South Atlantic States* (*LAMSAS*) (Kretzschmar, 1988). *LAMSAS* included 1162 interviewed subjects and the data collection period was from 1933 to 1974 (Nerbonne & Kleiweg, 2003). Then Kretzschmar (1993) spent several years making the data in LAMSAS accessible for reanalysis. Another work that has had a profound influence on North America English dialect research is the Atlas of *North American English* (*ANAE*) (Labov et al., 2006). It indicated that dialect diversity is increasing and several dialect regions display homogeneity across great distances (Labov, 2011). However, the interviewed subjects in both *LAMSAS* and *ANAE* are rather few people compared to the population and it took a long time to collect the data. *ANAE* even does not include rural areas. Grieve (2009) put forward a corpus-based regional dialect survey based on letters to editors and presented a statistical analysis of lexical variations in American English (Grieve et al., 2011). Their approach includes three steps: (1) identify significant regional variation patterns with spatial autocorrelation measures; (2) apply factor analysis to identify common dialect patterns; and (3) conduct cluster analysis to identify dialect regions. However, the data set focuses on formal written English.

Previous linguistic studies that use Twitter data have mainly focused on natural language processing and parts-of-speech tagging. Hong, Convertino, and Chi (2011) conducted a systematic analysis on the cross-language differences in tweets. Petrovic, Osborne, and Lavrenko (2010) built a Twitter corpus to help researchers work on natural language processing. Gimpel et al. (2011) used Twitter data to address the problem of part-of-speech tagging. Recently, more research has begun to use Twitter to study linguistic variations. Gonçalves and Sánchez (2014) used two years of Twitter data to study Spanish varieties at a global scale. Eisenstein, O'Connor, Smith, and Xing (2014) applied a latent vector autoregressive model on 107 million Twitter messages to study the diffusion of linguistic change over the United States. Criticisms of using Twitter data are mainly based on the uncertainty of its data quality and its socio-demographic representativeness (Crampton et al., 2013). Longley, Adnan, and Lansley (2015) attempted to profile Twitter users in terms of age, gender, and ethnicity based on user names. They point out that Twitter data may have an over representation of males and young adults. Goodchild (2013) argued that although big data may lack a normal process for quality control and rigorous sampling, big data can still be of high quality with its detailed, timely and original information (Kitchin, 2013).

Traditional dialectology research is generally qualitative. Séguy (1971) was the first to introduce statistical analysis of aggregated regional linguistic variation, an approach to dialectology known as *dialectometry*, which has been expanded on by various researchers who use multivariate and spatial methods to identify common patterns of regional linguistic variation (Goebl, 2006; Grieve et al., 2011; Heeringa, 2004; Kretzschmar, 1996; Lee & Kretzschmar, 1993; Nerbonne, 2006, 2009; Nerbonne & Kretzschmar, 2003; Nerbonne et al., 1996; Szmrecsanyi, 2013; Wieling & Nerbonne, 2011). Multivariate analysis usually involves examination of the joint relationship of variables and dimension reduction (James & McCulloch, 1990). Nerbonne (2006) introduced factor analysis to aggregate linguistic analysis. Thill, Kretzschmar, Casas, and Yao (2008) adopted Kohonen's (2001) self-organizing map to analyze the variations of word usage and pronunciation using the *LAMSAS* dataset. Principal component analysis (PCA) is another popular method used for multivariate analysis, which reduces variable dimensions with fewer measurements while retaining data variability in the original data (Rao, 1964). In spatial analysis, regionalization is the process of constructing homogeneous regions, e.g., climate zones or dialect regions, by optimizing a homogeneity function during the partition of space (Goodchild, 1979; Guo, 2008; Haining, Wise, & Blake, 1994, Handcock & Csillag, 2004; Masser & Scheurwater, 1980; Spence, 1968). Guo (2008) proposed a family of regionalization methods for constrained hierarchical clustering and partitioning (REDCAP) with multivariate information and a homogeneity measure, which has been applied in different domains such as forestry (Kupfer, Gao, & Guo, 2012) and health studies (Wang, Guo, & McLafferty, 2012). In this research, we use PCA to extract variables for describing linguistic characteristics and use REDCAP to discover dialect regions with the top PCA components.

Table 1
Content word lexical alternations.

Alternation		Alternation		Alternation	
Variant A	Other variant(s)	Variant A	Other variant(s)	Variant A	Other variant(s)
Bag	Sack	Mom	Mother	Absurd	Ridiculous
Clearly	Obviously	Whilst	While	Chuckle	Laugh
Grandfather	Grandpa	Center	Middle	Disturb	Bother
Couch	Sofa	Clothing	Clothes	Humiliating	Embarrassing
Automobile	Car	Best	Greatest	Job	Employment
Pupil	Student	Loyal	Faithful	Joy	Pleasure
Maybe	Perhaps	Real	Genuine	Likely	Probable
Especially	Particularly	Sad	Unhappy	Normal	Usual
Alley	Lane	Smart	Intelligent	Starting	Beginning
Holiday	Vacation	Baby	Infant	Start	Begin
Big	Large	Bet	Wager	Stupid	Dumb
Little	Small	Bought	Purchased	Unclothed	Naked
Supper	Dinner	Careful	Cautious	Bathroom	Restroom/washroom
Wrong	Incorrect	Comprehend	Understand	Envious	Jealous/covetous
Anywhere	Anyplace	Rude	Impolite	Quick	Fast/rapid
Required	Needed	Drowsy	Sleepy	Stomach	Tummy/belly
Each other	One another	Honest	Truthful	Trash	Garbage/rubbish
Afore	Before	Hug	Embrace	Grandma	Grandmother/granny/nana
Dad	Father	Hurry	Rush	All you	Y'all/you all/you guys
Ill	Sick	Band	Aid		

3. Methodology

3.1. Tweets and derived linguistic measures

The Twitter data used in this study includes geo-tagged tweets from Oct. 7, 2013 to Oct. 6, 2014 within the continental U.S. (48 states and Washington D.C.), which had 6.6 million Twitter users, 924 million geo-tagged tweets, and 7.8 billion words. We use lexical alternations to examine linguistic variations. A lexical alternation consists of two or more different words with the same referential meaning, referred to as variants, e.g. "*dad*/*father*". The set of 211 lexical alternations that we adopt in this study was first introduced by Grieve et al. (2013). The variants (words) of each alternation are generally interchangeable across contexts (i.e., independent of context) so that they can be directly extracted from Twitter messages. For each alternation we find the number of unique Twitter users in a specific county that have used any variant

Table 2
Descriptive statistics and global Moran's I test for 59 lexical alternations, out of which 38 alternations exhibit significant spatial autocorrelation (p-value < 0.001).

Alternation	Number of users in a county who has used any variant of the alternation		Number of counties where the alternation appeared	Spatial autocorrelation testing of mean-variant-preference (MVP) values for the alternation		
	Max	Mean		Moran's I	Z score	p-value
Best/Greatest	173,471	1541	3070	0.0426	9.5085	<0.0001
Little/Small	131,406	1132	3070	0.086	18.9327	<0.0001
Mom/Mother	103,840	919	3068	0.1561	34.3434	<0.0001
Big/Large	100,548	889	3068	0.0675	15.2916	<0.0001
Start/Begin	98,333	876	3064	0.0527	11.696	<0.0001
All You/Y'all/You All/You Guys	88,065	706	3064	0.045	9.9444	<0.0001
Ill/Sick	75,619	664	3059	0.1251	27.4097	<0.0001
Stupid/Dumb	75,131	642	3052	0.1177	25.7857	<0.0001
Dad/Father	67,119	590	3064	0.3287	72.0897	<0.0001
Maybe/Perhaps	62,650	572	3056	0.0364	8.1059	<0.0001
Quick/Fast/Rapid	60,483	523	3053	0.1866	40.7747	<0.0001
Center/Middle	70,947	498	3051	0.0815	17.8459	<0.0001
Starting/Beginning	47,242	415	3034	0.0244	5.3975	<0.0001
Supper/Dinner	54,070	410	3013	0.3486	75.4429	<0.0001
Bought/Purchased	34,586	260	3009	0.0187	4.1847	<0.0001
Clothing/Clothes	28,107	243	2985	0.0324	7.1045	<0.0001
Stomach/Tummy/Belly	27,392	238	2973	0.0636	13.7612	<0.0001
Clearly/Obviously	23,427	228	2997	0.0751	16.2802	<0.0001
Hurry/Rush	24,244	212	2985	0.0992	21.4767	<0.0001
Holiday/Vacation	24,494	210	3003	0.0552	11.9983	<0.0001
Grandma/Grandmother/Granny/Nana	21,560	195	3026	0.6216	134.916	<0.0001
Normal/Usual	22,151	189	2972	0.0906	19.5519	<0.0001
Bag/Sack	22,371	187	2978	0.0477	10.3641	<0.0001
Band/Aid	22,502	171	2956	0.0869	18.7065	<0.0001
Bathroom/Restroom/Washroom	18,526	148	2952	0.126	29.0561	<0.0001
Absurd/Ridiculous	13,845	137	2945	0.0374	8.1637	<0.0001
Loyal/Faithful	13,138	135	2870	0.0564	11.9408	<0.0001
Trash/Garbage/Rubbish	14,054	129	2910	0.183	43.613	<0.0001
Hug/Embrace	15,785	118	2935	0.0182	3.959	<0.0001
Joy/Pleasure	18,025	118	2893	0.0193	4.1471	<0.0001
Couch/Sofa	10,755	109	2911	0.1519	32.5271	<0.0001
Disturb/Bother	10,681	95	2863	0.0635	13.4411	<0.0001
Alley/Lane	9695	82	2848	0.0397	8.4611	<0.0001
Grandfather/Grandpa	9603	79	2964	0.3229	69.054	<0.0001
Job/Employment	58,169	540	3051	0.0148	3.6218	0.0002
Each Other/One Another	24,577	216	2999	0.0159	3.5518	0.0003
Chuckle/Laugh	35,388	287	3008	0.015	3.3801	0.0007
Drowsy/Sleepy	23,654	162	2854	0.0151	3.3593	0.0007
Wrong/Incorrect	61,950	605	3053	0.0146	3.2789	0.001
Real/Genuine	109,293	983	3065	0.013	3.0088	0.0026
Baby/Infant	95,805	844	3060	0.0119	2.8042	0.005
Rude/Impolite	16,336	140	2920	0.0103	2.6679	0.0076
Smart/Intelligent	19,605	173	2970	0.0107	2.3922	0.0167
Required/Needed	31,002	270	3014	0.0102	2.3256	0.02
Bet/Wager	25,044	278	3005	0.0098	2.2794	0.0226
Sad/Unhappy	64,025	505	3055	0.0088	2.0074	0.0447
Careful/Cautious	7058	63	2792	0.0064	1.4297	0.1527
Honest/Truthful	13,087	128	2917	0.0059	1.3461	0.1782
Especially/Particularly	22,880	186	2987	0.0053	1.2485	0.2118
Whilst/While	88,459	728	3059	0.0025	1.1269	0.2597
Likely/Probable	9329	78	2839	0.0031	0.834	0.4042
Anywhere/Anyplace	12,429	115	2938	0.0022	0.5754	0.565
Automobile/Car	78,790	695	3060	0.0017	0.4905	0.6237
Humiliating/Embarrassing	8753	76	2833	−0.0022	−0.4647	0.6421
Pupil/Student	9751	103	2818	0.0016	0.4451	0.6562
Envious/Jealous/Covetous	23,991	219	3001	−0.002	−0.3857	0.6996
Unclothed/Naked	14,082	120	2905	−0.0017	−0.3325	0.7394
Afore/Before	105,723	916	3066	0.0002	0.1606	0.8723
Comprehend/Understand	43,617	438	3043	−0.0004	−0.0343	0.9725

(word) of the alternation. For example, 9256 unique users in Richland County (SC) used the word "dad" or "father" in their tweets during the one-year period. Based on the user-county frequencies, we eliminate 152 infrequently used alternations. The elimination rule has two parts: (1) an alternation is considered present in a county if either variant has at least five users in the county; and (2) an alternation will be eliminated if it is not present in more than 1000 (out of 3111) counties. The remaining 59 alternations are listed in Table 1, which we use to define and analyze regional linguistic variations in the U.S.

We calculate a *variant-preference* (*VP*) value for each alternation and user in each county. Let $A(w, v)$ be an alternation with two variants w and v, e.g., "w = father, v = dad". If A has more than two variants, the one with the highest overall frequency in the corpus will be designated as w and all other variants is combined as v. Let $T(u, c, w)$ be the total number of tweets sent by user u in county c that contain variant w. Similarly $T(u, c, v)$ is the total number of tweets containing variant v that were sent by user u in county c. Then $VP(u,c,A) = T(u,c,w)/(T(u,c,w) + T(u,c,v))$, which is a ratio value of range [0, 1]. For example, for the alternation "w = mom, v = mother", if a user tend to use "mom" (15 tweets) more than "mother" (10 tweets), then his/her *VP* score for this alternation is 15 / (15 + 10) = 0.6. If both $T(u, c, w)$ and $T(u, c, v)$ are zero, then $VP(u, c, A)$ is assigned 0. Since mobility and migration have strong influence on the formation of linguistic characteristics of a place, we use the location of each tweet instead of finding a home county for each Twitter user. If a Twitter user has tweeted in two or more counties, he/she will be treated as a unique user in each county with his/her tweets in that county.

Next a *mean-variant-preference* (*MVP*) value is calculated for each county c and alternation $A(w, v)$, which is the average of non-zero variant-preference (*VP*) values for all users in c. Let $U(c, A)$ be the total number of unique users in a county c who has used alternation $A(w, v)$. Then $MVP(c,A) = \sum_u VP(u,c,A)/U(c,A)$, which represents the aggregated preference score of a county for the variants of A. Note that both measures, MVP is a normalized score that gives each user equal weight, regardless of the number of tweets (involving the alternation) sent by the users. After calculating an MVP value for each county/alternation combination, we have a newly derived dataset, which is a table of 3111 rows (counties) and 59 columns (alternations), with each cell being the MVP value of an alternation in a county.

3.2. Selection of alternations with spatial autocorrelation testing

As the goal is to extract regional linguistic variation based on the usage (choice) of alternation variants, it is necessary to focus primarily on those lexical alternations that exhibit significant spatial autocorrelation. Therefore, for each alternation we calculate a Global Moran's I value and its associated p-value, which are shown in Table 2. Among the 59 alternations, 38 exhibit highly significant spatial autocorrelation (p-value <0.001), 8 alternations are in the range of [0.001, 0.05], and the remaining 13 alternations have p-value >0.05. We use the 38 alternations with p-value <0.001 for further analysis to detect regional linguistic patterns. We also check the correlation coefficients for all pairs of the 38 alternations, which are all less than 0.9, indicating that there are no alternations that carry duplicate (identical) information.

3.3. Spatial variation of alternations

We can map and examine the spatial variation of each alternation based on MVP values. For example, Fig. 2a shows the map for alternation "Mom vs. Mother", where people in red counties tend to use

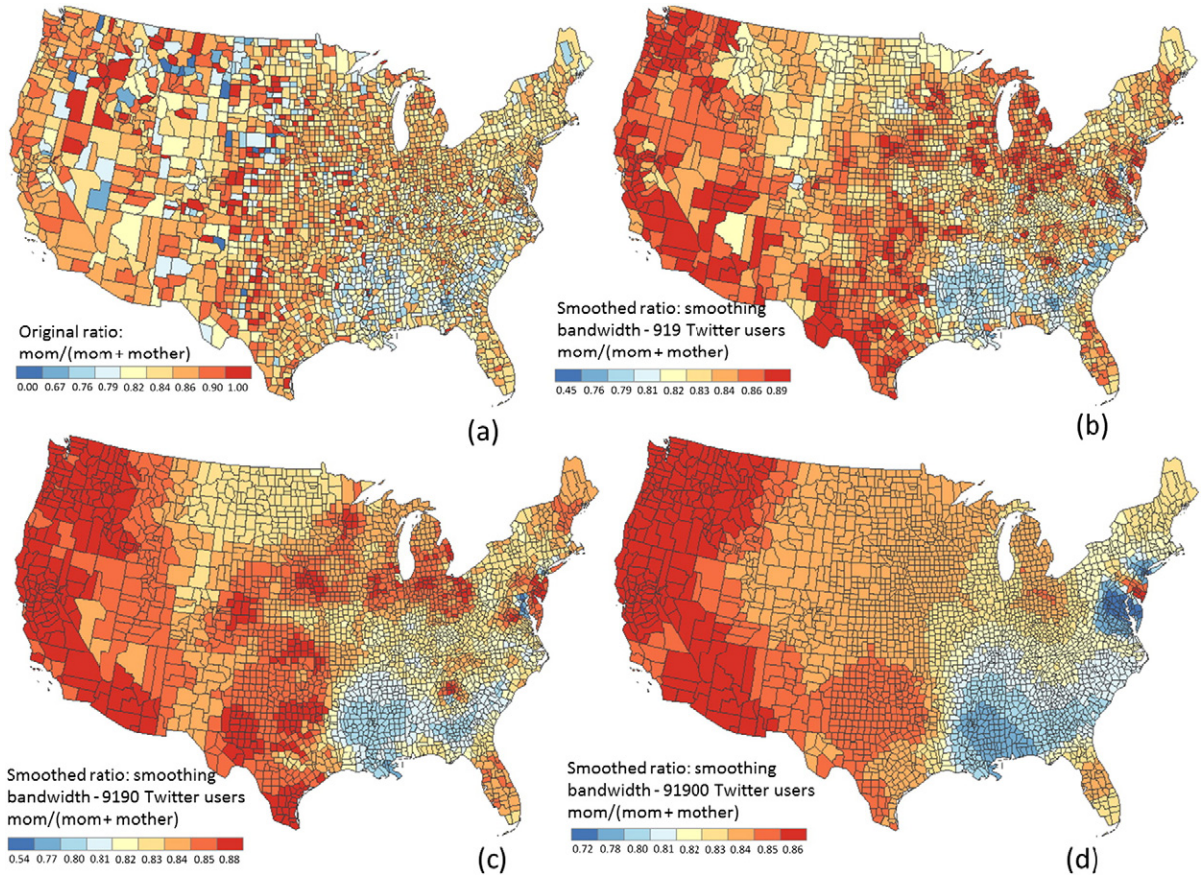

Fig. 1. Smoothed mean-variant-preference (MVP) values for the alternation "Mom/Mother", with different bandwidths. On average, there are 919 users of alternation "Mom/Mother" per county. (a) Original MVP values; (b) smoothed values with a bandwidth of 919 Twitter users; (c) smoothing with a bandwidth of 9190 users; and (d) smoothing with a bandwidth of 91,900 users.

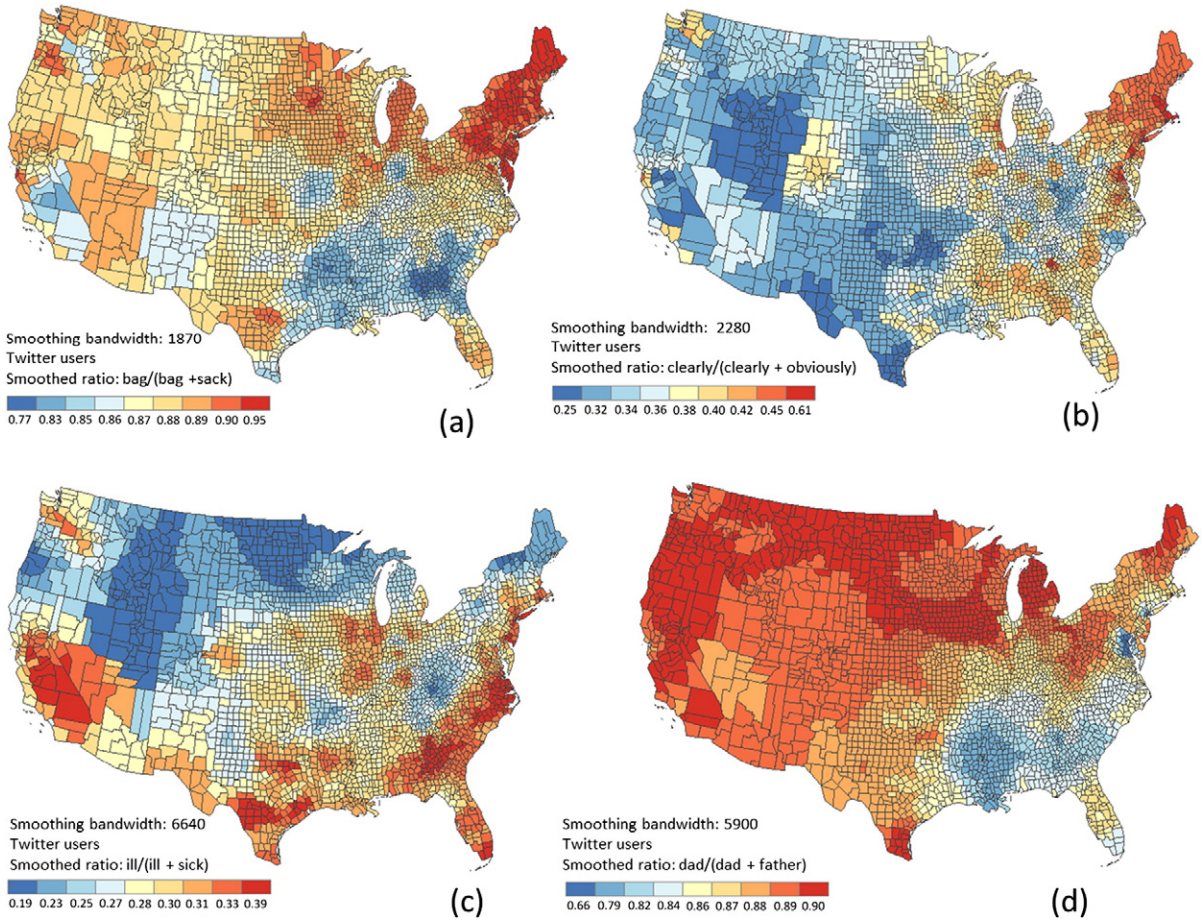

Fig. 2. Smoothed maps of four alternations: (a) "*bag/sack*", (b) "*clearly/obviously*", (c) "*ill/sick*", and (d) "*dad/father*".

"Mom" while people in blue counties prefer "Mother". However, due to the uneven spatial distribution of Twitter users and the dramatic size difference among counties, the MVP values may not be reliable for small counties that have too few users. Some counties do not have any user for a specific alternation. Furthermore, there are various other sources of linguistic variation that have not been directly controlled (including social, situational, topical and temporal variation), which can obscure underlying regional patterns. To address these problems, spatial smoothing can be applied to reduce spurious data variation, estimate values for counties that have no data available, and ultimately accentuate spatial patterns that are otherwise difficult to discern (Borruso & Schoier, 2004; Carlos, Shi, Sargent, Tanski, & Berke, 2010; Kafadar, 1996; Koylu & Guo, 2013).

We perform an adaptive kernel smoothing for each alternation, where the bandwidth d for a county is the minimum number of twitter users of the alternation in its neighborhood. Let $U(c,A)$ be the number of unique users of alternation A in county c. Then the smoothing neighborhood for c is $N(c,A)=\{b_i|\sum_i U(b_i,A)>d\}$,, i.e., the minimum set of nearest neighbors of c (inclusive) that contains at least d users of the alternation. With a kernel (e.g., the Gaussian kernel) each neighbor b_i is assigned a weight l_i and the smoothed value is the weighted average of neighbors' values: $MVP'(c,A)=\sum_i MVP(b_i,A)l_i$,where $\sum_i l_i=1$.

Table 3
Top PCA components.

	Component 1	Component 2	Component 3	Component 4	Component 5
Standard deviation	3.316	2.761	1.675	1.551	1.436
Proportion of variance	0.289	0.201	0.074	0.063	0.054
Cumulative proportion	0.289	0.490	0.564	0.627	0.681
	Component 6	Component 7	Component 8	Component 9	Component 10
Standard deviation	1.223	0.998	0.899	0.883	0.825
Proportion of variance	0.039	0.026	0.021	0.021	0.018
Cumulative proportion	0.721	0.747	0.768	0.789	0.807
	Component 11	Component 12	Component 13	Component 14	Component 15
Standard deviation	0.786	0.770	0.743	0.717	0.694
Proportion of variance	0.016	0.016	0.015	0.014	0.013
Cumulative proportion	0.823	0.839	**0.853**	0.867	0.879

Table 4
The loadings of the top five principal components (with top three loadings shaded).

Alternation	Component 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Bag/Sack	–0.217		–0.138		–0.202
Clearly/Obviously		0.301	–0.108	0.124	–0.157
Grandfather/Grandpa	0.118	0.203	0.122		–0.346
Couch/Sofa	–0.137		0.126		–0.271
Maybe/Perhaps	0.169	–0.218	–0.103	–0.196	–0.163
Alley/Lane	–0.171				0.208
Holiday/Vacation				–0.130	0.458
Big/Large	0.252	–0.108			0.166
Little/Small	–0.113	–0.173	0.116		–0.323
Supper/Dinner		–0.120		–0.443	0.124
Each Other/One Another	–0.114	–0.124	–0.358		
Dad/Father	–0.233	–0.178		0.122	
Ill/Sick	0.190		–0.268	0.159	
Mom/Mother	–0.154	–0.187	–0.117	0.286	
Center/Middle		0.223		0.181	0.246
Clothing/Clothes	–0.224	0.157		0.174	
Best/Greatest	0.178			0.262	
Loyal/Faithful		0.134	–0.352		
Bought/ Purchased	0.169	–0.199	–0.200	0.104	
Drowsy/Sleepy	–0.249				
Hug/Embrace		–0.212	–0.247		
Hurry/Rush	0.222	–0.174			
Band/Aid	–0.166		0.192		
Absurd/Ridiculous	–0.125	0.255			–0.194
Chuckle/Laugh	–0.221				
Disturb/Bother	0.197	0.187			0.125
Job/Employment	0.135	–0.209			
Joy/Pleasure	0.106	–0.210	0.211	–0.140	
Normal/Usual	–0.173	–0.240			
Starting/Beginning	–0.193		–0.245	–0.241	
Start/Begin	0.160	–0.140	–0.339		
Stupid/Dumb		–0.176	0.235	0.317	
Bathroom/Restroom + Washroom		0.131		–0.355	–0.183
Quick/Fast + Rapid	0.149	0.248	–0.136	0.142	
Stomach/Tummy + Belly	0.239		–0.106	–0.130	–0.104
Trash/Garbage + Rubbish	0.189			0.281	
Grandma/Grandmother + Granny + Nana	–0.194	–0.132	–0.216		0.268
All You/Yall + You All + You Guys		–0.203	0.190		–0.183

To configure the bandwidth value d, we first find the average user count (k) of an alternation per county and then set $d = ak$, where a is a positive integer. For example, on average there are 919 users per county for the alternation "Mom/Mother". Fig. 1(b), (c), (d) shows the smoothing results for bandwidth $d = 919, 919*10$, and $919*100$, respectively. We empirically set $d = 10\,k$ in our analysis, where a neighborhood consists of about 15 nearest counties on average. Note that, while k varies for different alternations, the neighborhood size in terms of the number of counties involved remains relatively stable with the above bandwidth setting. This leads to smoothing results of similar spatial resolution (detail) and meanwhile adaptive to the spatial distribution of users for a given alternation.

The geographical distribution of the smoothed MVP values of each alternation reveals interesting patterns of regional linguistic characteristics. Fig. 2 shows the smoothed patterns (with $d = 10\,k$) for four alternations: "*bag*/*sack*", "*clearly*/*obviously*", "*ill*/*sick*", and "*dad*/*father*". While the spatial pattern of "*dad*/*father*" is similar to that of "mom/mother" (in Fig. 1b), the other three exhibit very different patterns. For example, people in the Northeast region clearly favor "bag" over "sack" while it is much less so in the South. The alternation "clearly/obviously" shows a different divide of the country, where "clearly" is preferred in the East and users in the West uses "obviously" more. The alternation of "ill" and "sick" seems to reveal a general difference between the North and the South.

While the spatial variation of each alternation reveals interesting regional linguistic patterns, a more important and challenging question remains unanswered: What is the overall regional linguistic pattern manifested by ALL alternations (maps) collectively? We need to synthesize the patterns in 38 maps and present a holistic view of the linguistic characteristics at each place (i.e., county in our case) and the regional linguistic structure (dialect regions) of the U.S. Towards this goal, we perform a principal component analysis (PCA) of the 38 linguistic variables (alternations) to extract orthogonal dimensions (Section 3.4) and then use a multivariate regionalization method to detect the natural hierarchy of dialect regions in the U.S. (Section 3.5).

3.4. Principal components analysis

With the 38 county-based alternation MVP variables as input, a PCA is carried out to derive a smaller set of linearly uncorrelated variables

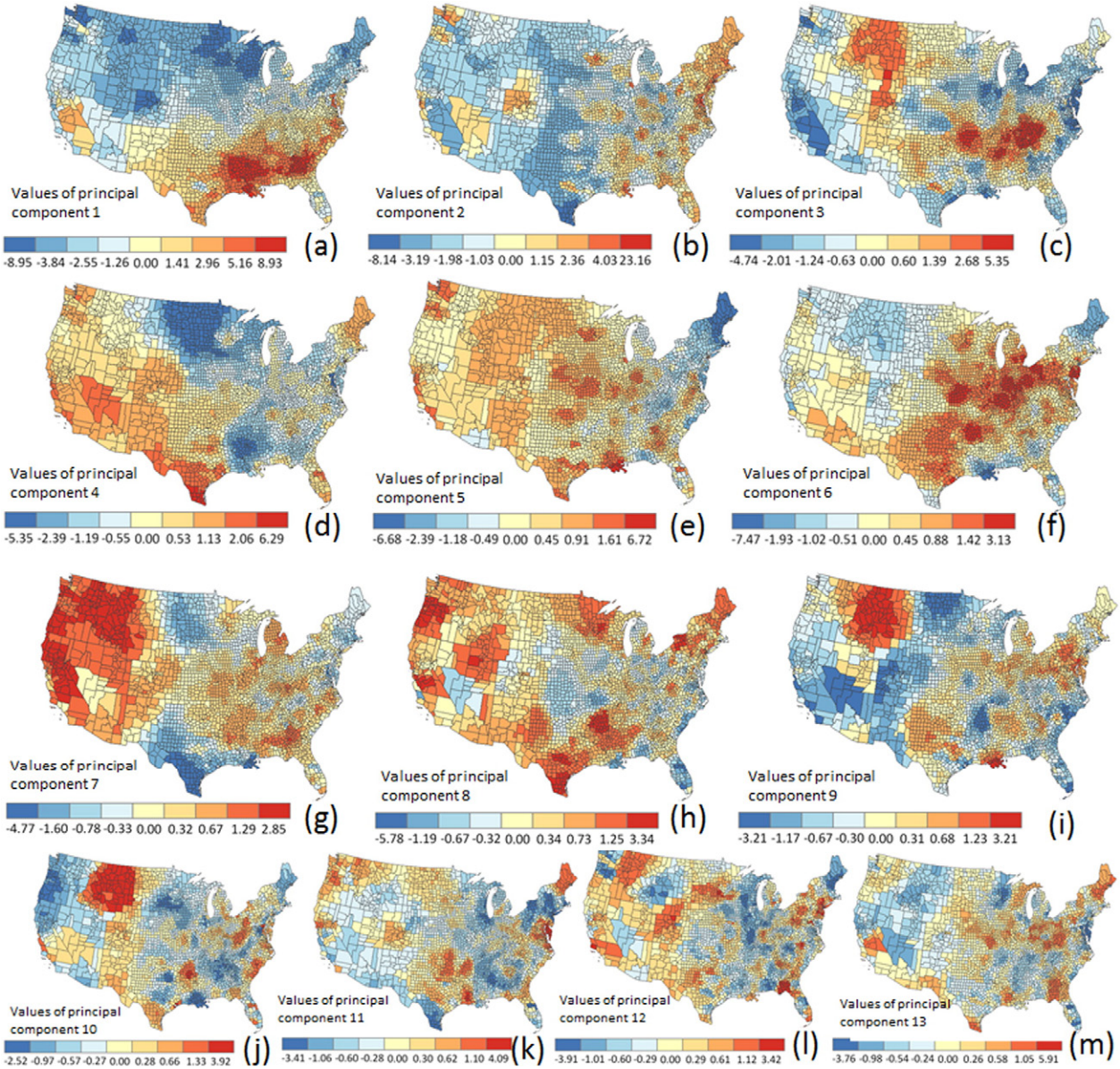

Fig. 3. The spatial distribution patterns of the top thirteen principal components.

(i.e., principal components) that explain that majority of data variance. Table 3 shows the top 15 PCA components. The selection of principal components to represent the original data may follow three different approaches (Bro & Smilde, 2014): (1) choose the top principal components whose eigenvalues are greater than 1; (2) choose principal components based on a scree test; or (3) choose the top principal components that explain a majority of data variation (Bro & Smilde, 2014). In our study, we adopt the third approach and choose the top 13 principal components together explain more than 85% of the original data variance.

Table 4 shows the loadings of the top five principal components, each of which explains more than 5% of the original data variance. The top three loadings for Component 1 are: *big* (vs. *large*), *sleepy* (vs. *drowsy*), and *stomach* (vs. *tummy* or *belly*). The top three loadings for Component 2 are the following: *clearly* (vs. *obviously*), *absurd* (vs. *ridiculous*), and *quick* (vs. *fast* + *rapid*). The top three loadings for Component 3 are: *one another* (vs. *each other*), *faithful* (vs. *loyal*), and *begin* (vs. *start*). The top three loadings for Component 4 are: *dinner* (vs. *supper*), *bathroom* (vs. *restroom* or *washroom*), and *stupid* (vs. *dumb*). The top three loadings for Component 5 are: *holiday* (vs. *vacation*), *grandfather* (vs. *grandpa*), and *little* (vs. *small*). As can be seen, the top three loadings for the top five components do not overlap each other.

Fig. 3 shows the spatial distribution patterns of the chosen thirteen principal components. Each component represents a set of alternations, which together exhibit a unique spatial pattern. Different components show uniquely different regional linguistic patterns. For example, Fig. 3(a) highlights the north/south distinctions, Fig. 3(b) shows the east/west divide; and the map in Fig. 3(c) highlights the coast/central difference.

3.5. Multivariate mapping

We use the SOMVIS multivariate mapping approach (Guo, Gahegan, MacEachren, & Zhou, 2005; Kohonen, 2001) to produce one map that synthesizes all thirteen components (Fig. 4). The approach groups counties into clusters with the thirteen input variables using the self-organizing map clustering method, which also arranges the clusters on a 2D layout (Fig. 4 top right) so that similar clusters are nearby each other. Then a 2D color scheme is imposed onto the layout to assign a color to each cluster (i.e., a node in the layout, represented by a circle) and make sure similar clusters have similar colors (Guo et al., 2005). Since each cluster represents a set of counties, each county is also assigned a color, same as that of its containing cluster. As such, a multivariate map is produced (Fig. 4 top left), where the color of each county indicates its cluster membership and each cluster represents a set of counties of similar linguistic characteristics (defined with the thirteen input variables, i.e., PCA components). The parallel coordinate plot (Fig. 4 bottom) shows the "meaning" of each cluster, with by a string of line segments of the same color connecting the value on each vertical axis, which represents a variable (i.e., PCA component in this case).

The multivariate map in Fig. 4 represents the holistic patterns manifested by the thirteen PCA components (and hence the 38 lexical alternations). From the map, one can visually understand the spatial variation of linguistic characteristics in the U.S. For example, it is evident that the northeast region (in red) is rather different from the rest of the country. From the parallel coordinate plot, we can tell that the red cluster has a very high value on Component 2 and relatively high on Components 4 and 8. The variable loadings of these components can be looked up in Table 4 (and its complete version that includes all 13 components).

3.6. Discovering hierarchical linguistic regions

While the multivariate map presents the overall regional linguistic patterns, the regional boundary and hierarchical structure is not explicit and their visual interpretation can be subjective. Therefore, we apply the REDCAP regionalization method to explicitly discover and define linguistic regions based on the multivariate data (i.e., the 13 PCA components that represent the 38 lexical alternations). Note that REDCAP is a different method from the SOMVIS (in the previous section), although they work with the same input data. REDCAP is a family of regionalization methods based on contiguity constrained hierarchical clustering, e.g., average-linkage or complete-linkage clustering (Guo,

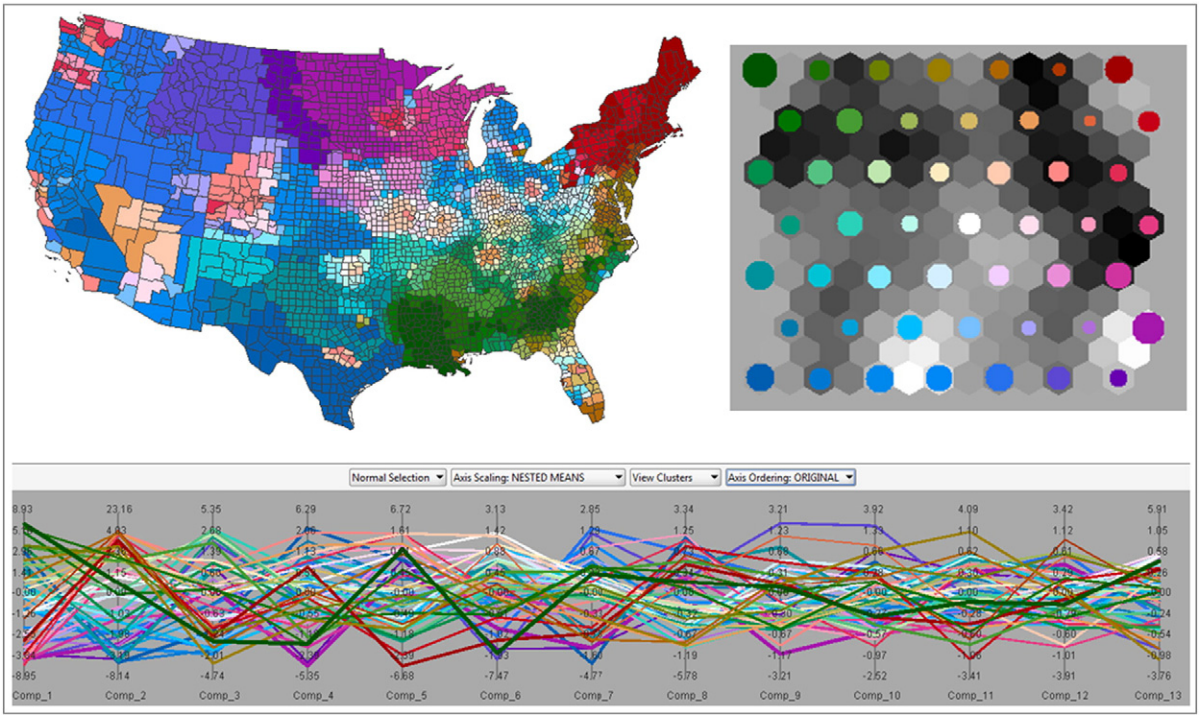

Fig. 4. Multivariate mapping of the thirteen PCA components.

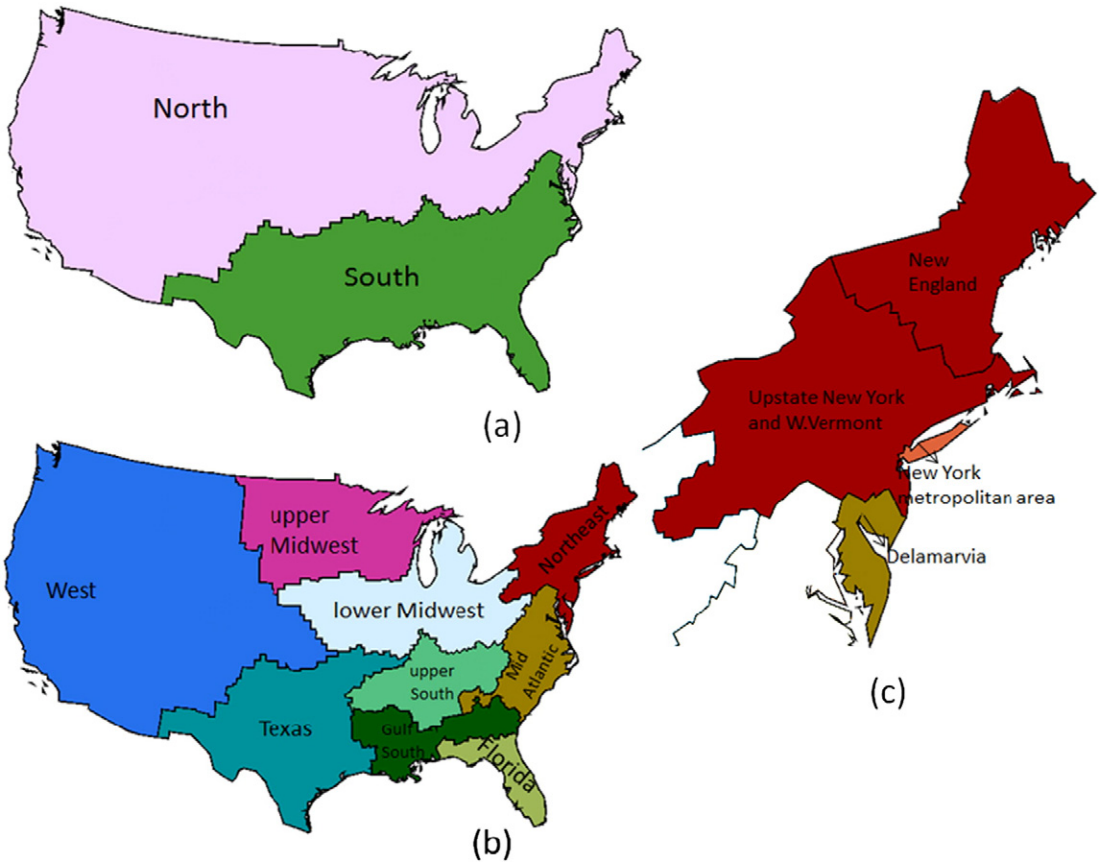

Fig. 5. Regionalization results at three hierarchical levels: (a) two regions; (b) nine regions; (c) sub-linguistic regions within the Northeast region.

2008). Here we use the full-order average-linkage method in REDCAP. Two counties are considered contiguous to each other if they share a segment of boundary. The "distance" or "dissimilarity" between two counties is the Euclidean distance between their multivariate linguistic vectors, each with values for the 13 PCA components. Note that spatial distance is not used in the dissimilarity definition. The method produces a hierarchy of clusters, the same as that of a traditional hierarchical clustering method, except that each cluster is a geographically contiguous region with internal homogeneity in terms of the 13 PCA component values.

Fig. 5 shows the discovered regions at three different hierarchical levels: two regions, nine regions, and the sub-regions within the northeast region. At the top level the country is divided into two primary linguistic regions: the North and the South (Fig. 5a). Interestingly this two-region boundary closely matches the cultural and lexical dialect boundaries (shown in Fig. 6) resulted from two broad streams of migration during the westward expansion and the cultural division between the North and South in the U.S. (Carver, 1987; Gastil & Glazer, 1975). Given that linguistic variation is a complicated phenomenon whose main processes include settlement history and migration, the matching between our two-region boundary and the cultural geography boundary indicates that (1) the regionalization with the 38 lexical alternations produces highly meaningful results, and (2) migration and settlement history still have great influence on regional linguistic characteristics, even in social media used by the younger segment of the population.

Going further down the hierarchy, more local dialect regions emerge. Fig. 5b presents the nine-region result and Fig. 7 shows its comparison with Labov et al.'s (2006) work, which is a study of the regional dialects of English spoken in the U.S. Their work was based on interviews of 762 people sampled from major urban areas in the U.S. between 1992 and 1999. Given the relatively small sample size and the focus on urban areas only, their dialect regions only have approximate boundaries. Nevertheless, we can see a strong similarity between our regions and theirs: (1) the West and Florida regions exist in both; and

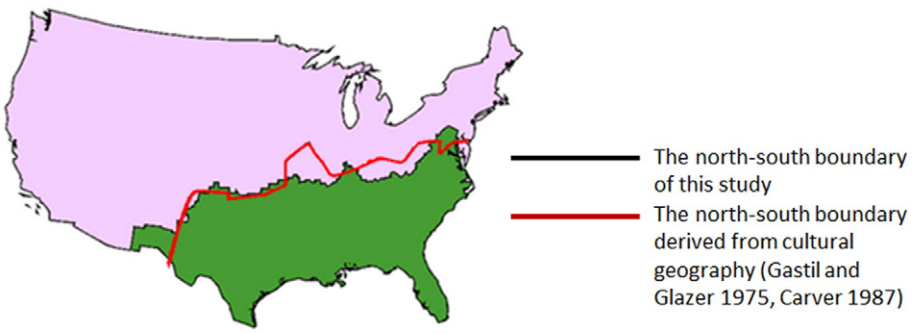

Fig. 6. The North–South boundary derived from our study and the cultural geography boundary of the North and South in the literature (Carver, 1987; Gastil & Glazer, 1975).

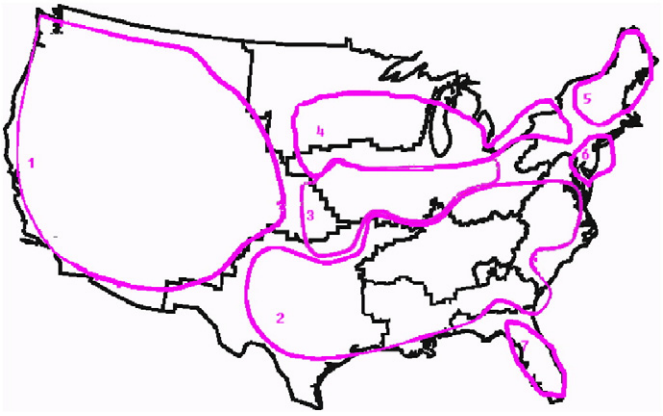

Fig. 7. Comparison of our nine-region result (dark boundaries) with the dialect regions from Labov et al.'s (2006) study (purple boundaries).

(2) the South and Northeast regions in our results are similar to their regions 2 and 5 (we further partitioned the South into several local regions). Our result is also very similar to that of Grieve et al. (2013), which is a reanalysis of Labov et al.'s (2006) data.

On the other hand, the difference between our results and that of Labov et al. (2006) may be attributed to several factors. First, our data set is much larger and covers all counties in United States, including both urban and rural areas, which may lead to more detailed and hopefully more reliable results. Second, the two studies use different sets of linguistic variables: we use lexical alternations while Labov et al. (2006) used phonetic characteristics. Third, Labov and colleagues derive the dialect regions with a manual approach (which may be affected by prior assumptions) while our regionalization is based on a more objective computational approach. This is likely why our results are more similar to Grieve et al.'s (2013) statistical reanalysis of Labov et al.'s (2006) phonetic data.

We can further divide regions into sub-regions along the hierarchy by our result. For example, within the Northeast region we can derive four sub-linguistic regions (Fig. 5c), including two small but distinct dialect regions: the New York City metropolitan area and the Delamarvia region, both of which agree well with the literature (Kurath, 1949; Labov et al., 2006). It is commonly accepted that there is a New York City dialect and the finding of the New York City dialect region indicates to certain degree that our results are convincing. To help readers understand the regional hierarchy better, Fig. 8 presents the maps for 3–8 regions, which form the hierarchy between the 2-region level and the 9-region level in Fig. 5.

4. Discussion and conclusion

There are several issues worthy of further discussion and investigation. First, as is inherent in most big data, the quality of Twitter data needs to be examined. Other than the potential demographic representation bias in Twitter data, one important issue is to deal with abbreviations and spelling-mistakes in such casual and short messages. However, our study focuses upon a set of commonly used and simple words (lexical alternations), for which we believe misspelling and abbreviation are not major concerns. Second, geo-tagged tweets only represent a small portion (2–3%) of all tweets—if more location information can be extracted through text-mining (Xu, Wong, & Yang, 2013), it would lead to even larger data sets and possibly more reliable outcomes. Third, twitter data contains spam messages, including non-personal and organization-initiated messages such as weather alerts, news feeds, etc. Guo and Chen (2014) find that 2–3% of Twitter users are spam users that send geo-tagged tweets and these users often send more tweets than regular human users. Ideally spam messages should be excluded but since we create a MVP score for each user in each county, its effect should be lessened.

The smoothing bandwidth is chosen through a visual comparison of different smoothing bandwidths to avoid under-smoothing and over-smoothing. It should be interesting to design a more objective selection procedure to help select an "optimal" bandwidth automatically. We select thirteen principal components that explain more than 85% of total data variance, with the assumption that there is a certain level of noise variance that should not be included. It is unknown, however, how much of the total variance is noise.

Linguistic variation is a gradual and fuzzy process. It may not be surprising to have different results based on different criterion, methodology and data sources. It can also be interesting to further examine the difference between linguistic regions defined with lexical information and those with acoustic information. In this paper we have shown that the lexical regions we produced match rather well with the phonetic regions in previous studies. Regarding lexical information, our choice of lexical alternations is based on the most recent research in linguistic studies. In future more alternations or better choices may emerge. An even more challenging but also interesting direction would be to extract and select lexical alternations automatically from the Twitter data.

To summarize, this study derives linguistic variation (dialect) regions based on lexical alternations with one year of Twitter data.

Fig. 8. Region hierarchy, from three regions to eight regions.

Principal component analysis and regionalization methods are used to automatically discover hierarchical dialect regions, which reveal interesting and up-to-date regional variation patterns of linguistic characteristics. Compared to prior studies, our results show both convincing similarity and difference. While the difference may need further validation, the advantages of our approach and results are clear. *First*, geo-tagged tweets provide unprecedented rich information for linguistic analysis which has spatial and temporal continuity, a large sample size, and is very recent. This is quite different from traditional linguistic studies that often take years to collect a small sample. *Second*, with automatic computational methodologies, more objective outcomes can be achieved in an efficient way. With both advantages, it becomes possible to examine the dynamics of linguistic characteristics and their spread at finer spatial–temporal resolutions.

To the best of our knowledge, this paper is among the very few papers that use social media data to study nation-wide linguistic variations. Although Twitter data is sometimes criticized due to its bias and uncertainty, and its demographic representation of the language community (Eisenstein & Nerbonne, 2015), the regionalization results indicate that it makes sense to use Twitter data in linguistic studies. The spatially and temporally continuous attributes of Twitter data could not only reflect regional linguistic characteristics, but also could contribute greatly to the study of other types of spatial–temporal variation in linguistics. The geo-tagged tweets also represent an opportunity for cultural geographers to get involved in research using big data. Such studies could reinvigorate cultural geography by examining how present-day spatial patterns may reflect deep historical processes, for example, Cheshire and Longley's study of surnames (2012).

Acknowledgments

This paper is partly based upon work funded by NSF Grant No. 0748813 and Digging Into Data Award LG-00-14-0030-14 by the Institute of Museum and Library Services (IMLS).

References

Atwood, E. B. (1953). *A survey of verb forms in the Eastern United States*. University of Michigan Press.
Borruso, G., & Schoier, G. (2004). Density analysis on large geographical databases: Search for an index of centrality of services at urban scale. *Computational science and its applications–ICCSA 2004* (pp. 1009–1015). Springer.
Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, *6*(9), 2812–2831.
Carlos, H. A., Shi, X., Sargent, J., Tanski, S., & Berke, E. M. (2010). Density estimation and adaptive bandwidths: A primer for public health practitioners. *International Journal of Health Geographics*, 39.
Carver, C. M. (1987). *American regional dialects: A word geography*. University of Michigan Press.
Chambers, J. K., & Trudgill, P. (1998). *Dialectology*. Cambridge University Press.
Cheshire, J. A., & Longley, P. A. (2012). Identifying spatial concentrations of surnames. *International Journal of Geographical Information Science*, *26*(2), 309–325.
Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: Situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, *40*(2), 130–139.
Di Nunzio, G. M. (2013). Digital geolinguistics: On the use of linked open data for data-level interoperability between geolinguistic resources. Paper read at *SDA*.
Eisenstein, J., & Nerbonne, J. (2015). Identifying regional dialects in online social media. In C. Boberg, & D. Watt (Eds.), *Handbook of dialectology*. Wiley-Blackwell Press.
Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PloS One*, *9*.
Gastil, R. D., & Glazer, N. (1975). *Cultural regions of the United States*. Seattle: University of Washington Press.
Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., ... Smith, N. A. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. Paper read at *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: short papers-volume 2*.
Goebl, H. (2006). Recent Advances is Salzburg Dialectometry. *Literary and Linguistic Computing*, *21*, 411–435.
Gonçalves, B., & Sánchez, D. (2014). Crowdsourcing dialect characterization through Twitter. *PloS One*, *9*(11), e112074.
Goodchild, M. F. (1979). The aggregation problem in location allocation. *Geographical Analysis*, *11*, 240–255.
Goodchild, M. F. (2013). The quality of big (geo) data. *Dialogues in Human Geography*, *3*(3), 280–284.
Grieve, J. (2009). *A corpus-based regional dialect survey of grammatical variation in written standard American English*. Northern Arizona University.
Grieve, J., Asnaghi, C., & Ruette, T. (2013). Site-restricted web searches for data collection in regional dialectology. *American Speech*, *88*, 413–440.
Grieve, J., Speelman, D., & Geeraerts, D. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, *23*(02), 193–221.
Grieve, J., Speelman, D., & Geeraerts, D. (2013). A multivariate spatial analysis of vowel formants in American English. *Journal of Linguistic Geography*, *1*(01), 31–51.
Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, *22*(7), 801–823.
Guo, D., & Chen, C. (2014). Detecting non-personal and spam users on geo-tagged Twitter network. *Transactions in GIS*, *18*(3), 370–384.
Guo, D., Gahegan, M., MacEachren, A. M., & Zhou, B. (2005). Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach. *Cartography and Geographic Information Science*, *32*(2), 113–132.
Haining, R. P., Wise, S. M., & Blake, M. (1994). Constructing regions for small area analysis: Material deprivation and colorectal cancer. *Journal of Public Health Medicine*, *16*, 429–438.
Handcock, R., & Csillag, F. (2004). Spatio-temporal analysis using a multiscale hierarchical ecoregionalization. *Photogrammetric Engineering and Remote Sensing*, *70*, 101–110.
Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance. Ph.D. Dissertation. Manuscript*. University of Groningen.
Hong, L., Convertino, G., & Chi, E. H. (2011). Language matters In Twitter: A large scale study. Paper read at *ICWSM*.
James, F. C., & McCulloch, C. E. (1990). Multivariate analysis in ecology and systematics: Panacea or Pandora's box? *Annual Review of Ecology and Systematics*, *21*, 129–166. http://dx.doi.org/10.2307/2097021.
Kafadar, K. (1996). Smoothing geographical data, particularly rates of disease. *Statistics in Medicine*, *15*(23), 2539–2560.
Kitchin, R. (2013). Big data and human geography opportunities, challenges and risks. *Dialogues in Human Geography*, *3*(3), 262–267.
Kohonen, T. (2001). *Self-organizing maps*, *Vol. 30*, Springer Science & Business Media.
Koylu, C., & Guo, D. (2013). Smoothing locational measures in spatial interaction networks. *Computers, Environment and Urban Systems*, *41*, 12–25. http://dx.doi.org/10.1016/j.compenvurbsys.2013.03.001.
Kretzschmar, W. A., Jr. (1988). Computers and the American linguistic atlas. Paper read at *Methods in dialectology: proceedings of the sixth international conference on methods in dialectology*.
Kretzschmar, W. A. (1993). *Handbook of the linguistic atlas of the middle and South Atlantic states*. University of Chicago Press.
Kretzschmar, W. A. (1996). Quantitative areal analysis of dialect features. *Language Variation and Change*, *8*(01), 13–39.
Kupfer, J. A., Gao, P., & Guo, D. (2012). Regionalization of forest pattern metrics for the continental United States using contiguity constrained clustering and partitioning. *Ecological Informatics*, *9*, 11–18.
Kurath, H. (1949). *A word geography of the Eastern United States*. University of Michigan Press.
Labov, W. (2011). *Principles of linguistic change, cognitive and cultural factors*, *Vol. 3*, John Wiley & Sons.
Labov, W., Ash, S., & Boberg, C. (2006). *The atlas of North American English: Phonetics, phonology, and sound change: A multimedia reference tool*, *Vol. 1*, Walter de Gruyter.
Lee, J., & Kretzschmar, W. (1993). Spatial analysis of linguistic data with GIS functions. *International Journal of Geographical Information Science*, *7*(6), 541–560.
Longley, P. A., Adnan, M., & Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment and Planning A*, *47*(2), 465–484.
Masser, I., & Scheurwater, J. (1980). Functional regionalization of spatial interaction data — An evaluation of some suggested strategies. *Environment and Planning A*, *12*(12), 1357–1382.
McDavid Raven I, Virginia G McDavid, William A Kretzschmar, Theodore K Lerud, and Martha Ratliff. 1986. "Inside a linguistic atlas." Proceedings of the American Philosophical Society:390–405.
Nerbonne, J. (2006). Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing*, *21*(4), 463–475.
Nerbonne, J. (2009). Data-driven dialectology. *Language and Linguistics Compass*, *3*, 175–198.
Nerbonne, J., & Kleiweg, P. (2003). Lexical distance in LAMSAS. *Computers and the Humanities*, *37*(3), 339–357.
Nerbonne, J., & Kretzschmar, W. (2003). Introducing Computational Techniques in Dialectometry. *Language Resources and Evaluation*, *3*, 245–255.
Nerbonne, J., Wilbert, H., Eric, H., Peter, K., Simone, O., & Willem, V. (1996). Phonetic distance between Dutch dialects. Gert Durieux, Walter Daelemans, & Steven Gillis (Eds.), *Proceedings of the Sixth CLIN Meeting* (pp. 185–202). Antwerp: Centre for Dutch Language and Speech.
O'Cain, R. K. (1979). Linguistic atlas of New England. *American Speech*, 243–278.
Petrovic, S., Osborne, M., & Lavrenko, V. (2010). The Edinburgh Twitter corpus. Paper read at *Proceedings of the NAACL HLT 2010 workshop on computational linguistics in a world of Social Media*.
Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002)*, *26*(4), 329–358. http://dx.doi.org/10.2307/25049339.
Séguy, J. (1971). In G. Straka (Ed.), *La relation entre la distance spatiale et la distance lexicale*. Palais de l'université.

Spence, N. A. (1968). A multivariate uniform regionalization of British counties on the basis of employment data for 1961. *Region Studies*, *2*, 87–104.
Szmrecsanyi, B. (2013). *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*. Cambridge: Cambridge University Press.
Thill, J. -C., Kretzschmar, W. A., Casas, I., & Yao, X. (2008). Detecting geographic associations in English dialect features in North America within a visual data mining environment integrating self-organizing maps. *Self-organising maps: applications in geographic information science* (pp. 87–106).
Wang, F., Guo, D., & McLafferty, S. (2012). Constructing geographic areas for cancer data analysis: A case study on late-stage breast cancer risk in Illinois. *Applied Geography*, *35*(1), 1–11.
Wieling, M., & Nerbonne, J. (2011). Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language*, *25*, 700–715.
Wolfram, W., & Schilling-Estes, N. (2005). *American English: Dialects and variation*. Wiley.
Xu, C., Wong, D. W., & Yang, C. (2013). Evaluating the "geographical awareness" of individuals: An exploratory analysis of Twitter data. *Cartography and Geographic Information Science*, *40*(2), 103–115.