Spatial Linguistics Variations in Northeastern U.S.

Based on Geo-Tagged Tweets

Andi Liao
Advisor: Luc Anselin
2019/04/18

Research Question

How do **linguistic features** derived from

lexical alternation pairs

local words

average score for state

geo-tagged Tweets vary within the **Northeast** part

latitude + longitude + text info

to overcome data sparsity

of U.S.?

Boundary Clarification

Content

- **No NLP!!!**
- **Spatial Linguistics**

Method: Spatial Data Science

Concept: Geo-Linguistics

Task

- **Explore spatial linguistics variations at state level**
- **Predict geo-locations using text information**

Relevant Literature: Geo-Linguistics Patterns

Spatial

- **Spatial distribution of lexical alternation pairs**
  - Multivariate mapping approach using 13 principal components
  - Regionalization methods for constrained hierarchical clustering and partitioning

Mom-Mother

- **Spatial variations of African American Vernacular English**
  - Mapping around 30 common nonstandard spellings on Twitter
  - Subregions align with movement patterns during the Great Migrations
- Huang, Guo, Kasakoff & Grieve (2016); Jones(2015)

Temporal

- **Diffusion of lexical changes**
  - An autoregressive model of word frequencies to demonstrate the linguistic influence between American cities
  - The network is helpful in identifying geographical and demographical factor that drives the spread of lexical innovation
- **Linguistics evolvement in urban areas using frequently used terms**
  - A logistics regression model consisting of geographical and demographical predictors
  - Absolute difference of the percentage of African Americans was the most powerful indicator of linguistics transmit
- Eisenstein, O'Connor, Smith, and Xing (2012, 2014)

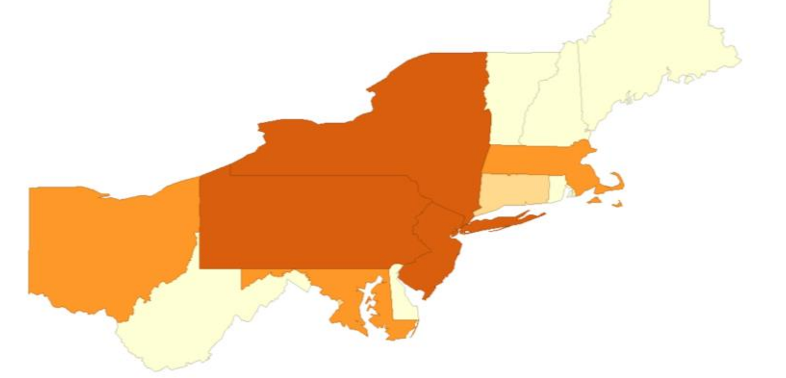
Relevant Literature: Geo-Prediction Models

- **A probabilistic framework via Tweet contents**
  - Trained a local word classifier
  - Constructed a lattice-based neighborhood smoothing model to balance cities and words of various distributions
  - Both local word filtering and smoothing have positive impact on prediction accuracy, and with location estimators, 51% of Twitter users can be placed within 100 miles of their actual locations at the city level
- **Decomposing lexical variation as regional and topical variation**
  - Constructed a prediction model with the assumption that regions and topics interact to shape observed lexical frequencies
  - The model can identify words with high regional affinity as well as geographically-coherent linguistic regions
- **A multi-elemental location inference method**
  - Combing text contents, profile location and place labelling
  - The model can successfully predict 87% of Tweets locations at the average distance error of 12.2 km
- Cheng, Caverlee and Lee (2010) ; Eisenstein, O'Connor, Smith and Xing (2010) ; Laylavi, Rajabifard and Kalantari (2016)

Data

Overcome Data Sparsity

Figure2: Natural Break Maps for Northeast Region



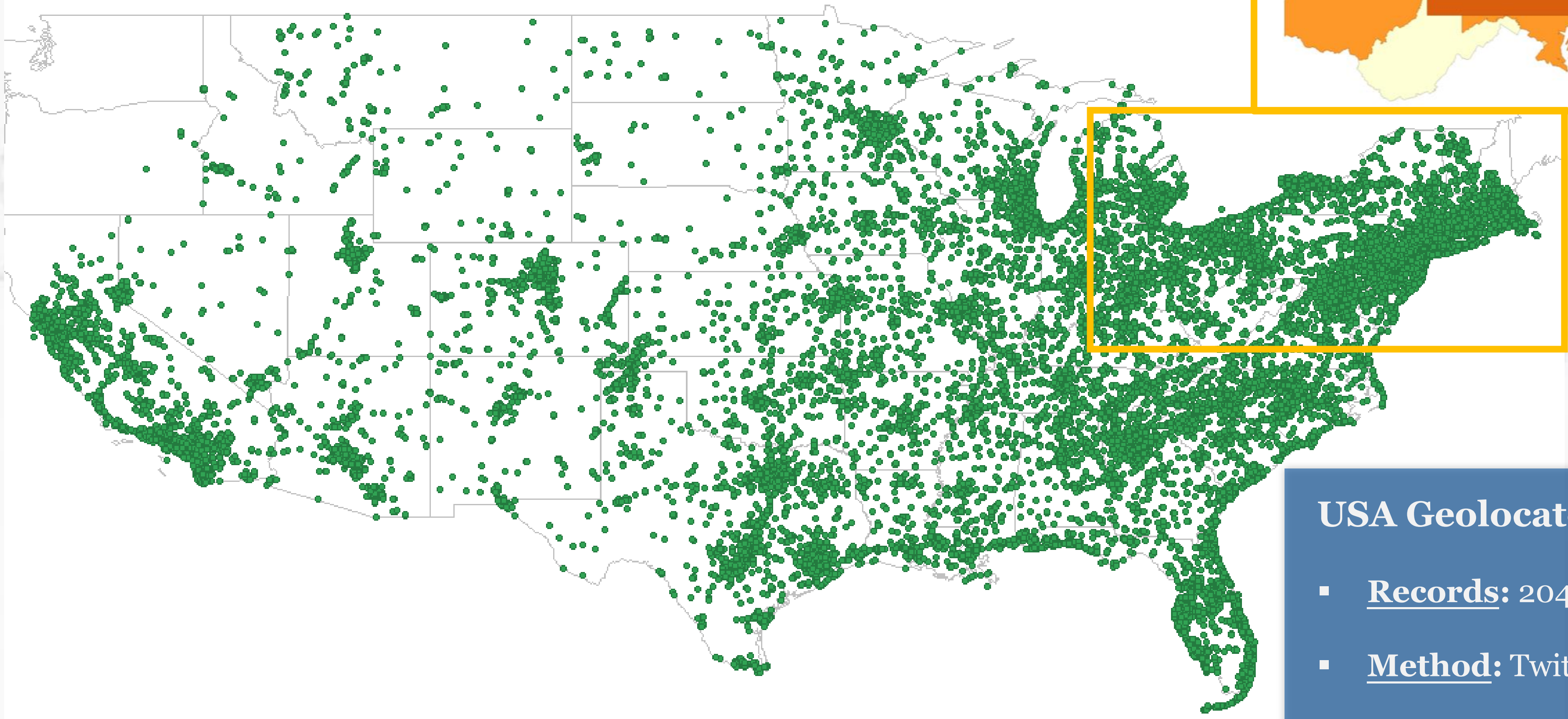
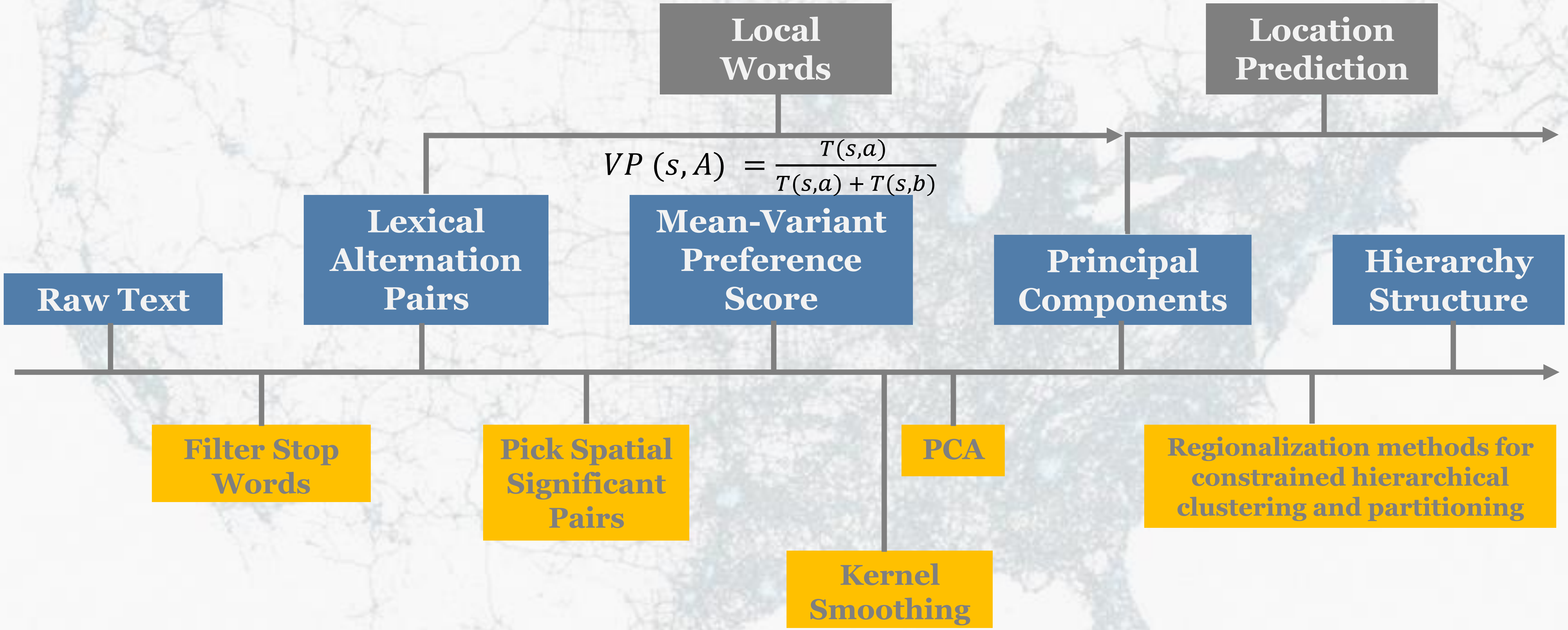


Figure 1: Spatial Distribution of Tweets Dataset

USA Geolocated Twitter Dataset

- **Records**: 204,820 observations
- **Method**: Twitter API
- **Time**: 2016/04/14-16
- **Source**: http://followthehashtag.com/

Method



Local Words
Location Prediction
$VP\ (s, A) = \frac{T(s,a)}{T(s,a) + T(s,b)}$
Raw Text
Lexical Alternation Pairs
Mean-Variant Preference Score
Principal Components
Hierarchy Structure
Filter Stop Words
Pick Spatial Significant Pairs
Kernel Smoothing
PCA
Regionalization methods for constrained hierarchical clustering and partitioning

Results

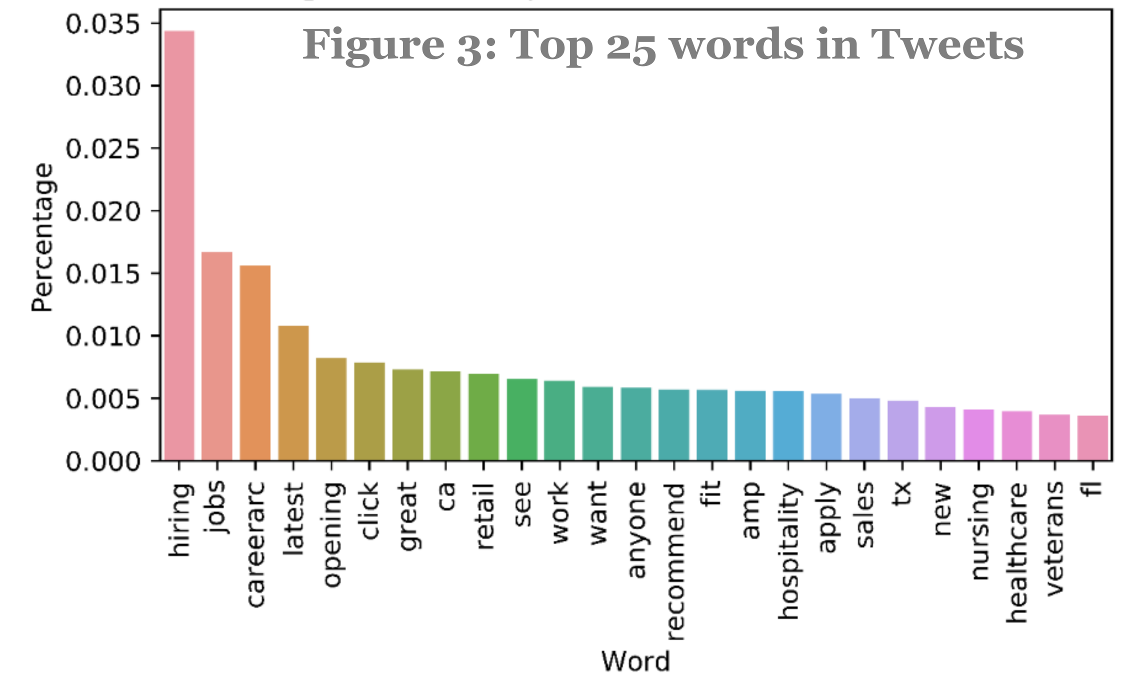
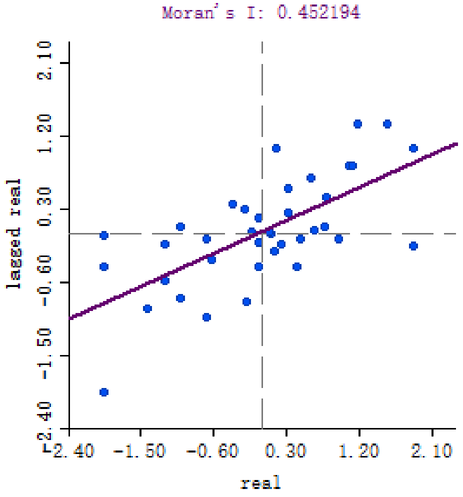
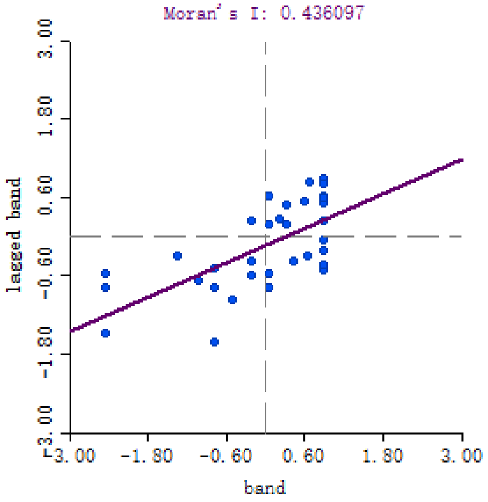
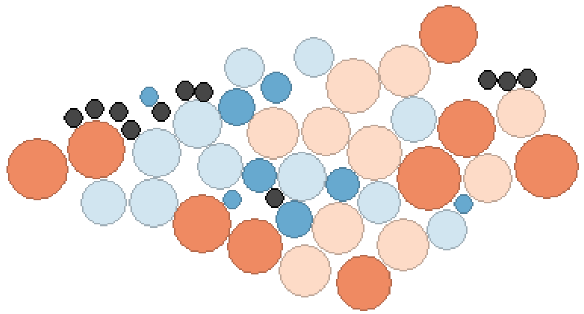


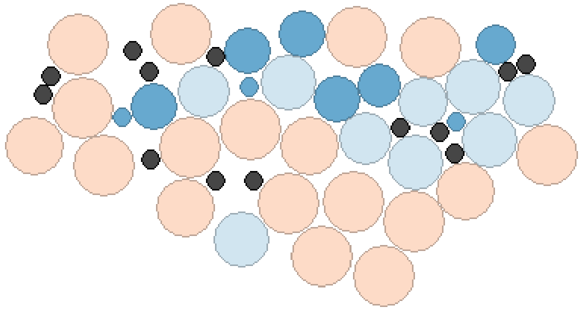
Figure 3: Top 25 words in Tweets

	Total Variances Explained	Singular Values
PC1	0.97181456	99.11797571
PC2	0.00711633	8.48181549
PC3	0.00201042	4.50821371
PC4	0.00178227	4.2447043
PC5	0.00168624	4.12876605






Real - Genuine


Band - Aid

Discussion

Issue

- Data Sparsity
  - Not enough data for each user or county
  - Might overlook existing patterns
- What to include in local words
  - Too similar for each state
  - Prediction model failed
- We are where we tweet?
  - Geo-locations can cheat
  - Reply on self-report

Contribution

- Used spatial methods to study linguistics topics
- The Northeast region does have prefer words compared to the country

Spatial Linguistics Variations in Northeastern U.S.

Based on Geo-Tagged Tweets

Thank you!