Helena Littman, Katerina Alvarez and Remy Wang
Professor Evan Ray
STAT 340: Applied Regression Methods
17 December 2019

# Dialects in America

## Summary

In this project, we investigated the effect of location on distinct American dialects. In a country as vast as the United States, there is hardly ever a consensus on how to pronounce certain words. Linguists Bert Vaux and Scott Golder surveyed more than 30,000 people from all 50 states in the early 2000s to compile some of the starkest regional divisions in American English, from vocabulary to pronunciation. Graphic artist Josh Katz eventually turned the results into a series of maps, and updated them for his 2016 book "Speaking American." The surprising data illuminate the linguistic quirks that make American English such a fascinating dialect.

The Dialect Survey uses a series of questions, including rhyming word pairs and vocabulary words, to explore the distribution of dialects in America. In our project, we were able to use longitude and latitude to predict an individual's pronunciation of the word lawyer, mayonnaise and pajamas from Question 14, 16 and 20. Our explanatory variable is latitude and longitude and response is one of the three questions at any given time. In regards to our in-class method, we used this dataset to fit a KNN classification model showing predicted class membership at each location in the US based on dialect. Our second analysis was our out-of-class method, and we chose neural networks to help us cluster, classify and recognize patterns in the data.

## Data

We used a subset of “The Harvard Dialect Survey,” a 2003 survey of 47,472 people that collected their city, state, zip code, and answers to 122 dialect questions. Data was collected from each of the 50 states, and from a variety of age groups, ranging from respondents 13 and younger to 70 and older. The majority of participants were from the east coast, and approximately a third of the participants were in the 20-29 age range (Vaux & Golder 2003). We merged in latitude and longitude data based on zip codes to use in our data set as well. This procedure uses a smaller subset of the data that includes zip code, latitude, longitude and the responses to a given question from the dialect survey.

*Question 14*

Question 14 asks respondents about their pronunciation of the word "Lawyer." As seen in Figure 1, the options and percentage of individuals who responded with that answer were: (a) pronounced like "boy" ("loyer") (72.84%), (b) pronounced like "saw" ("law-yer") (21.96%), (c) I use both interchangeably (4.86%), and (d) other (0.34%) (Vaux & Golder 2003). 11,421 of the participants responded to this question, and their answers were coded 1, 2, 3, or 4 in the data, respectively. Participants who did not respond to this question were coded as 0 in the dataset. Figure 1 below shows the distribution of answers across the United States.

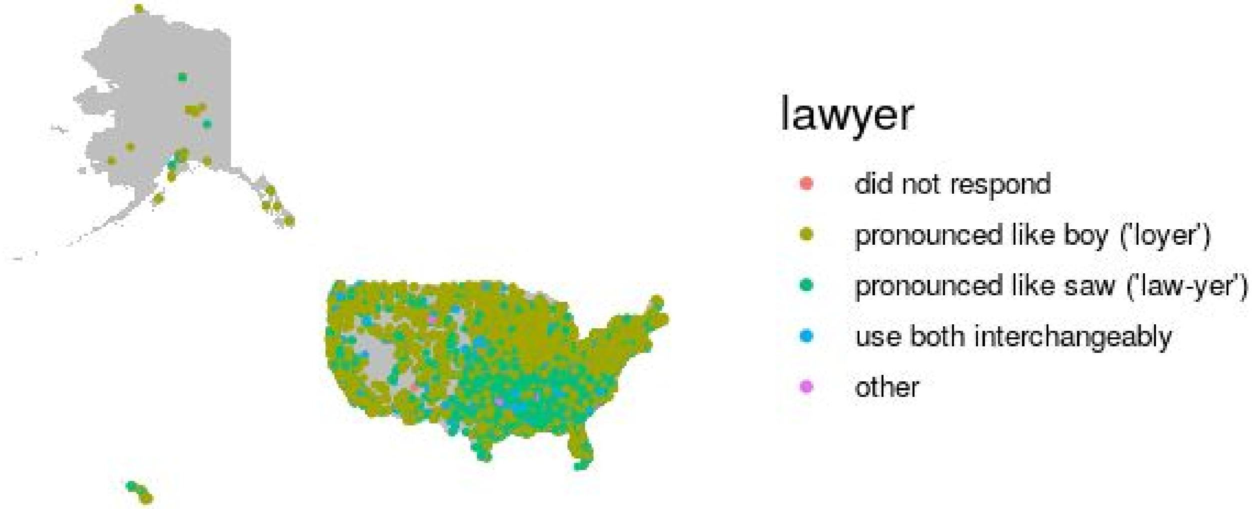

lawyer
did not respond
pronounced like boy ('loyer')
pronounced like saw ('law-yer')
use both interchangeably
other

Figure 1

*Question 16*

Question 16 asks respondents about their pronunciation of the word "mayonnaise." As displayed in figure 2, the options and percentage of individuals who responded with that answer were: (a) pronounced "man" (2 syllables--"man-aze") (41.65%), (b) with 3 syllables--"may-uh-naze" (45.83%), (c) I use both interchangeably (8.81%), (d) other (3.71%) (Vaux & Golder 2003). 11,372 of the participants responded to this question, and their answers were coded 1, 2, 3, or 4 in the data, respectively. Participants who did not respond to this question were coded as 0 in the dataset.

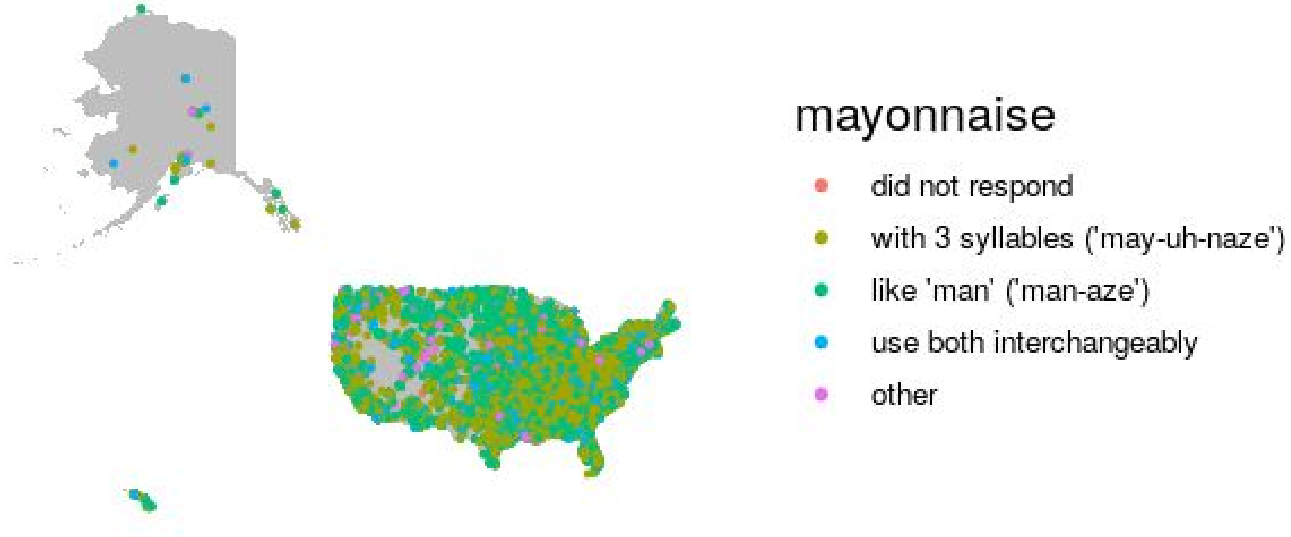

mayonnaise
● did not respond
● with 3 syllables ('may-uh-naze')
● like 'man' ('man-aze')
● use both interchangeably
● other

Figure 2

*Question 20*

Question 20 asks respondents about their pronunciation of the second vowel in the word "pajamas." With 11, 277 responses recorded, the possible answers and percentages of each were: (a) pronounced as in "jam" (45.92%), (b) pronounced as in "father" (51.86%), and (c)other (2.23%) (Vaux & Golder 2003). Answers were coded 1, 2, 3, respectively, in the dataset, and no response was recorded as 0. The distribution of these responses across the United States is shown in Figure 3 below.

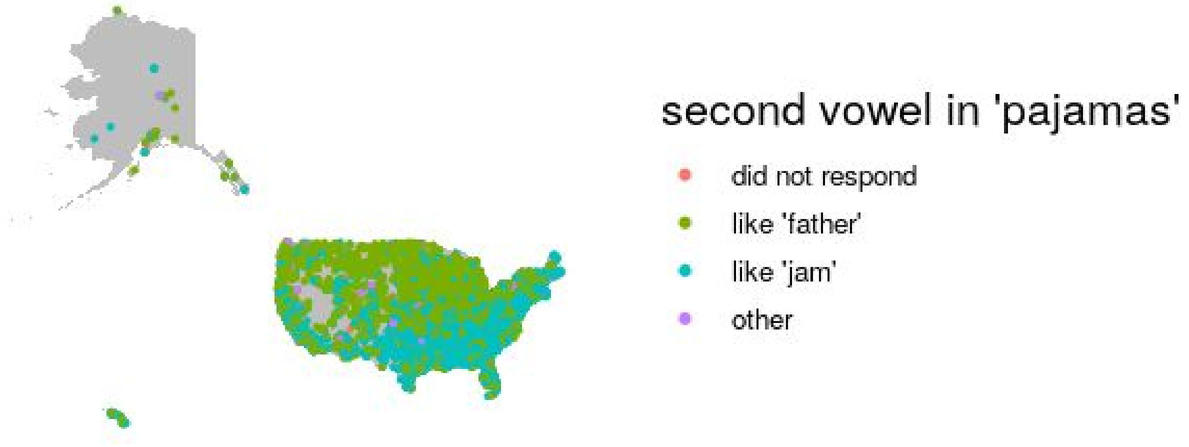


Figure 3

**Methods**

We used a subset of 10,000 observations for our analyses because the dataset was so large. Then, we chose K-Nearest Neighbors (KNN) for classification because of the nature of the data set. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure. K refers to the number of neighbors which you look at to make your estimation. Therefore, we knew that using a method that looked at the neighbors would be helpful because it would map a dialect group. Moving forward, we performed cross-validation and looked at values of K from 1 to 10, 25, 50, 75, 100, 150, 200, 250, and 300. We then calculated error rates for each K value to determine which number of neighbors would result in the best performing model.

In order to select the appropriate value of K to use, we perform cross-validation. We begin by splitting our subset of the data into "estimation" and test sets. Within our "estimation" set data, we split this subset into ten folds; 9 folds were the training set data, and 1 fold was the validation set data. After using the training set data to fit the model, we use that model to predict the responses of the observations in the validation set, and get a validation error rate. The validation set allows us to compare a set of candidate models, for example using different explanatory variables, and pick the one that works best for this dataset. We then use the test set from earlier to see how well our model works on a new portion of the data. Since we are doing a classification problem, we find the classification error rate to determine how well our model worked.

The second method we used was neural networks (NN) or a set of algorithms that are designed to recognize patterns since they interpret sensory data through a kind of machine perception, labeling and clustering of raw input. Most importantly, a NN is "a highly parameterized model, inspired by the architecture of the human brain, that was widely promoted as a universal approximator—a machine that with enough data could learn any smooth predictive relationship" (Bradley, 351). In addition, while a neural network is just a nonlinear model, NNs use a procedure called supervised learning to discover new features from the data.

For example, NN's can be scaled up and generalized in a variety of ways: "many hidden units in a layer, multiple hidden layers, weight sharing, a variety of colorful forms of regularization, and innovative learning algorithms for massive data sets" (352). However, it's important to note that without the nonlinearity in the hidden layer, the neural network would reduce to a generalized linear model. Also important to note, typically, neural networks are fit by maximum likelihood, usually with a variety of forms of regularization, which is what we see in our own models, especially for questions 14 where there is a high disparity in chosen dialect for the word "lawyer".

## Results

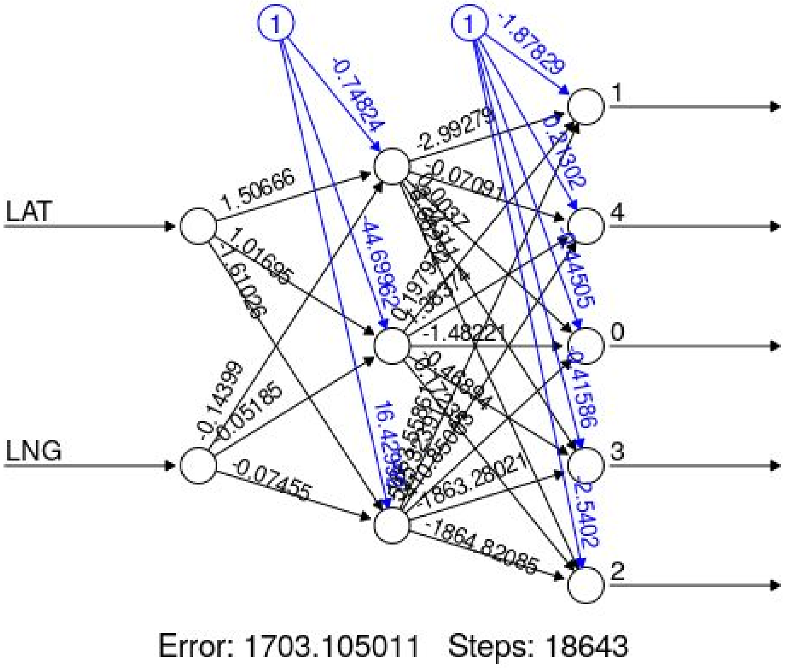
### *Question 14*

For KNN, we used cross-validation and selected a K value of 50 because it had the lowest classification error rate.

| **K** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 25 | **50** | 75 | 100 | 150 | 200 | 250 | 300 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CER** | 0.380 | 0.367 | 0.327 | 0.317 | 0.311 | 0.299 | 0.296 | 0.289 | 0.290 | 0.287 | 0.277 | **0.274** | 0.277 | 0.278 | 0.278 | 0.280 | 0.279 | 0.280 |

For our neural network, our map predicted that each response would be response 1 (pronounced like “loyer”). However, the network cannot do any better than predicting the majority value for the whole data set, in this case “loyer” at 1,408 responses. For all other responses, the neural network was marked incorrect at value “0.”

N = 10,000, Threshold = 0.1



| | Did not respond | Pronounced like “loyer” | Pronounced like “saw” | Use both interchangeably | Other |
|---|---|---|---|---|---|
| **Did not respond** | 0 | 17 | 0 | 0 | 0 |
| **Pronounced like “loyer”** | 0 | **1408** | 0 | 0 | 0 |
| **Pronounced like “saw”** | 0 | 465 | 0 | 0 | 0 |
| **Use both interchangeably** | 0 | 103 | 0 | 0 | 0 |
| **Other** | 0 | 5 | 0 | 0 | 0 |

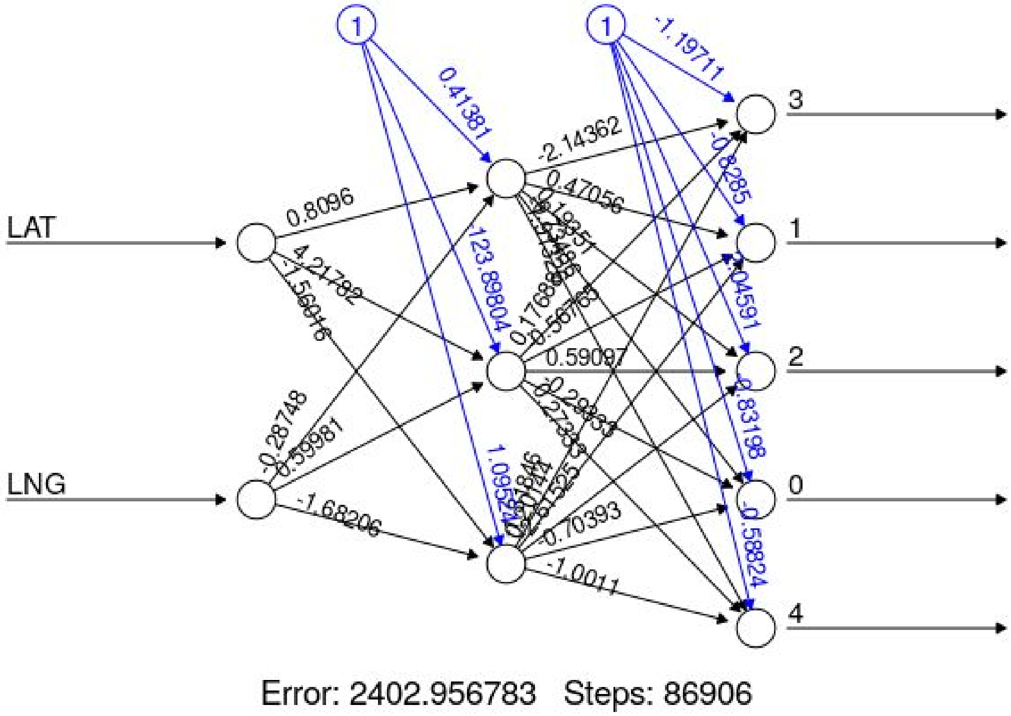
CER = 590/1998 = **0.295**

*Question 16*

For KNN, we used cross-validation and selected a K value of 250 because it had the lowest classification error rate.

| **K** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 25 | 50 | 75 | 100 | 150 | 200 | **250** | 300 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CER** | 0.583 | 0.574 | 0.563 | 0.550 | 0.545 | 0.538 | 0.534 | 0.527 | 0.529 | 0.524 | 0.515 | 0.507 | 0.505 | 0.503 | 0.501 | 0.497 | **0.496** | 0.497 |

N = 10,000, Threshold = 0.1



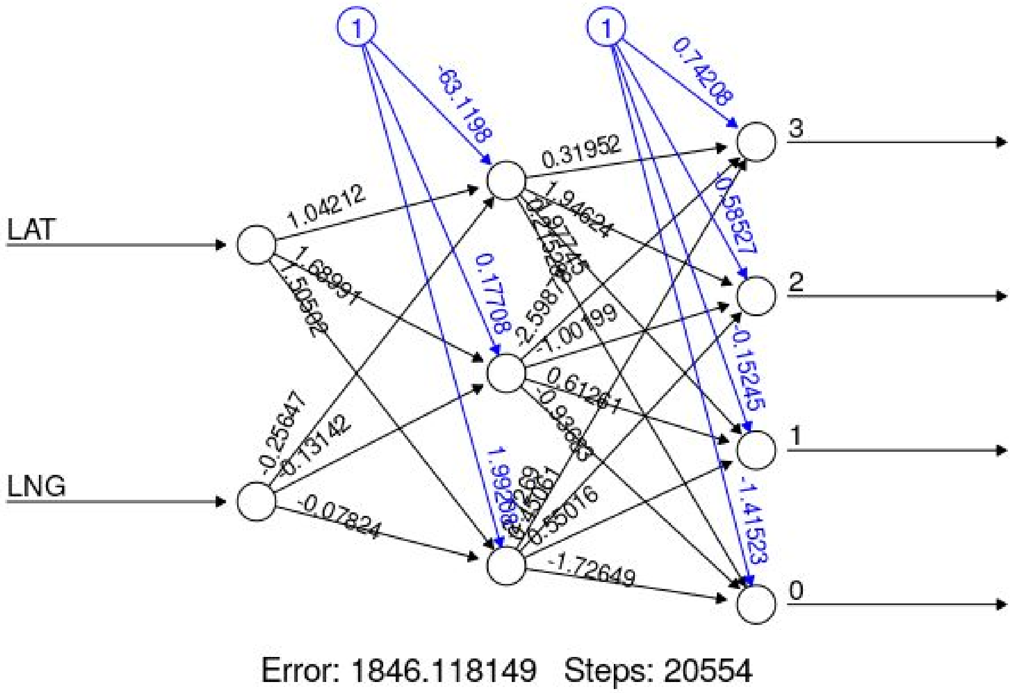
| | | Predicted Response | | | | |
|---|---|---|---|---|---|---|
| | | Did not respond | 3 syllables | 2 syllables | use both interchangeably | other |
| Actual Response | Did not respond | 0 | 11 | 8 | 0 | 0 |
| | 3 syllables | 0 | **525** | 313 | 0 | 0 |
| | 2 syllables | 0 | 413 | **486** | 0 | 0 |
| | use both interchangeably | 0 | 98 | 80 | 0 | 0 |
| | other | 0 | 31 | 33 | 0 | 0 |

CER = (11+413+98+31+8+313+80+33)/1998 **= 0.494**

*Question 20*

For KNN, we used cross-validation and selected a K value of 200 because it had the lowest classification error rate.

| **K** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 25 | 50 | 75 | 100 | 150 | **200** | 250 | 300 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CER** | 0.439 | 0.420 | 0.391 | 0.387 | 0.375 | 0.369 | 0.369 | 0.365 | 0.360 | 0.356 | 0.348 | 0.348 | 0.341 | 0.339 | 0.338 | **0.336** | 0.337 | 0.338 |



| | | Predicted Response | | | |
|---|---|---|---|---|---|
| Actual Response | | Did not respond | as in "jam" | as in "father" | other |
| | Did not respond | 0 | 16 | 12 | 0 |
| | as in "jam" | 0 | **673** | 297 | 0 |
| | as in "father" | 0 | 316 | **654** | 0 |
| | other | 0 | 19 | 11 | 0 |

CER = 671/1998= **0.336**

## Discussion

Our methods utilizing KNN and Neural Networks both attempted to predict the dialects of people by shaping networks and connections based on their locations. In our project, we subset the data to a smaller set because it was more important to explore our methods versus using the full data set that interrupted our code running time on R. Therefore, we used a random sample of 10,000 observations and applied this to question 14, 16 and 20.

For question 14 (lawyer), the classification error rate (CER) for KNN was 0.274, while the CER for the neural network was 0.295. KNN was a better indicator with a slightly lower classification error rate, because the threshold we used in our neural network did not create a model that allowed us to predict the distinctions in the groups of responses. This question had the majority of responses in one category making it easier for models to predict correctly than our other questions. However, the CER for KNN was still relatively high, even with 50 neighbors, so the accuracy of this method is not very high either.

For question 16 (mayonnaise), we chose a much higher value for K with our KNN method, with K = 250. The CERs for both methods were relatively high, but the neural network provided a slightly more accurate model. Still, with CERs of 0.496 and 0.494, neither model was very good at predicting dialect based on longitude and latitude. For question 20 (pajamas), the lowest CER for KNN was 0.336 for a K value of 200. The neural network for this question produced a CER of 0.336 as well, making both methods equally accurate.

Overall, in order to compile some of the starkest regional divisions in American English, from vocabulary to pronunciation, we analyze our US map diagram, depicted in Figure 1, 2 and 3. As predicted, we see significant regional differences with how people in the North and South speak. In a country as vast as the United States, you're hardly ever going to find a consensus on how to say something, so the answers really vary depending on where you ask the question. The surprising data illuminate the linguistic quirks that make American English such a fascinating dialect. We mostly see this striking contrast in question 14 and 20, and less so in question 16.

**References**

Abadi, Mark. “27 Fascinating Maps That Show How Americans Speak English Differently across the US.” Business Insider, Business Insider, 3 Jan. 2018, https://www.businessinsider.com/american-english-dialects-maps-2018-1.

Dialect Survey Results, https://www4.uwm.edu/FLL/linguistics/dialect/staticmaps/q_14.html.

Efron, Bradley, et al. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, Cambridge University Press, 2019, pp. 351–371. https://web.stanford.edu/~hastie/CASI_files/PDF/casi.pdf

Ellis, Laura. “Map Plots Created With R And Ggmap.” Little Miss Data, Little Miss Data, 15 Apr. 2018, https://www.littlemissdata.com/blog/maps.

“How to Increase the Size of Points in Legend of ggplot2?” *Stack Overflow*, 6 Dec. 2013, stackoverflow.com/a/20416049.

Kassambara. “How To Easily Customize GGPlot Legend for Great Graphics.” *Datanovia*, 18 Nov. 2019, www.datanovia.com/en/blog/ggplot-legend-title-position-and-labels/#change-legend-title

Lorenzo, Paolo Di. Usmap: Mapping the US, 12 Sept. 2019, https://cran.r-project.org/web/packages/usmap/vignettes/mapping.html.

“Neural Network Models in R.” DataCamp Community, https://www.datacamp.com/community/tutorials/neural-network-models-r.

Vaux, Bert, and Scott Golder. “The Harvard Dialect Survey.” Cambridge, MA: Harvard University Linguistics Department (2003).