Mapping dialectal variation by querying social media

Gabriel Doyle
Department of Linguistics
University of California, San Diego
La Jolla, CA, USA 92093-0108
gdoyle@ucsd.edu

Abstract

We propose a Bayesian method of estimating a conditional distribution of data given metadata (e.g., the usage of a dialectal variant given a location) based on queries from a big data/social media source, such as Twitter. This distribution is structurally equivalent to those built from traditional experimental methods, despite lacking negative examples. Tests using Twitter to investigate the geographic distribution of dialectal forms show that this method can provide distributions that are tightly correlated with existing gold-standard studies at a fraction of the time, cost, and effort.

1 Introduction

Social media provides a linguist with a new data source of unprecedented scale, opening novel avenues for research in empirically-driven areas, such as corpus and sociolinguistics. Extracting the right information from social media, though, is not as straightforward as in traditional data sources, as the size and format of big data makes it too unwieldy to observe as a whole. Researchers often must interact with big data through queries, which produce only positive results, those matching the search term. At best, this can be augmented with a set of "absences" covering results that do not match the search term, but explicit negative data (e.g., confirmation that a datapoint could never match the search term) does not exist. In addition to the lack of explicit negative data, query-derived data has a conditional distribution that reverses the dependent and independent variables compared to traditional data sources, such as sociolinguistic interviews.

This paper proposes a Bayesian method for overcoming these two difficulties, allowing query-derived data to be applied to traditional problems without requiring explicit negative data or the ability to view the entire dataset at once. The test case in this paper is dialect geography, where the positive data is the presence of a dialectal word or phrase in a tweet, and the metadata is the location of the person tweeting it. However, the method is general and applies to any queryable big data source that includes metadata about the user or setting that generated the data.

The key to this method lies in using an independent query to estimate the overall distribution of the metadata. This estimated distribution corrects for non-uniformity in the data source, enabling the reversal of the conditionality on the query-derived distribution to convert it to the distribution of interest.

Section 2 explains the mathematical core of the Bayesian analysis. Section 3 implements this analysis for Twitter and introduces an open-source program for determining the geographic distribution of tweets. Section 4 tests the method on problems in linguistic geography and shows that its results are well-correlated with those of traditional sociolinguistic research. Section 5 addresses potential concerns about noise or biases in the queries.

2 Reversing the conditionality of query data

2.1 Corpora and positive-only data

In traditional linguistic studies, the experimenter has control over the participants' metadata, but not over their data. For instance, a sociolinguist may select speakers with known ages or locations, but will not know their usages in advance. Corpus queries reverse the direction of investigation; the experimenter selects a linguistic form to search for, but then lacks control over the metadata of the participants who use the query. The direction of conditionality must be reversed to get compara-

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 98–106,
Gothenburg, Sweden, April 26-30 2014.

ble information from query-derived and traditional data.

Queries also complicate the problem by providing only positive examples. This lack of explicit negative data is common in language acquisition, as children encounter mostly grammatical statements during learning, and receive few explicitly ungrammatical examples, yet still develop a consistent grammaticality classification system as they mature. Similar positive-only problems abound in cognitive science and artificial intelligence, and a variety of proposals have been offered to overcome it in different tasks. These include biases like the Size Principle (Tenenbaum and Griffiths, 2001), heuristics like generating pseudo-negatives from unobserved data (Okanohara and Tsujii, 2007; Poon et al., 2009), or innate prespecifications like Universal Grammar in the Principles and Parameters framework.

For query-derived data, Bayesian reasoning can address both problems by inverting the conditionality of the distribution and implying negative data. The key insight is that a lack of positive examples where positive examples are otherwise expected is implicit negative evidence. This method allows a researcher to produce an estimated distribution that approximates the true conditional distribution up to a normalizing factor. This conditional distribution is that of data (e.g., a dialectal form) conditioned on metadata (e.g., a location).

This distribution can be written as $p(D|M)$, where D and M are random variables representing the data and metadata. A query for a data value d returns metadata values m distributed according to $p(M|D = d)$. All of the returned results will have the searched-for data value, but the metadata can take any value.

For most research, $p(M|D = d)$ is not the distribution of interest, as it is conflated with the overall distribution of the metadata. For instance, if the query results indicate that 60% of users of the linguistic form d live in urban areas, this seems to suggest that the linguistic form is more likely in urban areas. But if 80% of people live in urban areas, the linguistic form is actually underrepresented in these areas, and positively associated with rural areas. An example of the effect of such misanalysis is shown in Sect. 4.2.

2.2 Reversing the conditionality

Bayesian reasoning allows a researcher to move from the sampled $p(M|D)$ distribution to the desired $p(D|M)$. We invoke Bayes' Rule:

$$p(D|M) = \frac{p(M|D)p(D)}{p(M)}$$

In some situations, these underlying distributions will be easily obtainable. For small corpora, $p(D)$ and $p(M)$ can be calculated by enumeration. For data with explicit negative examples available, $p(D)$ can be estimated as the ratio of positive examples to the sum of positive and negative examples.[1] But for queries in general, neither of these approximations is possible. Instead, we estimate $p(M)$ through the querying mechanism itself.

This is done by choosing a "baseline" query term q whose distribution is approximately independent of the metadata – that is, a query q such that $p(q|m)$ is approximately constant for all metadata values $m \in M$. If $p(q|m)$ is constant, then by Bayes' Rule:

$$p(m|q) = \frac{p(q|m)p(m)}{p(q)} \approx p(m), \ \ \forall m \in M$$

Thus we can treat results from a baseline query as though they are draws directly from $p(M)$, and estimate the denominator from this distribution. The remaining unknown distribution $p(d)$ is constant for a given data value d, so combining the above equations yields the unnormalized probability $\tilde{p}(d|M)$:

$$p(d|M) \propto \tilde{p}(d|M) = \frac{p(M|d)}{p(M|q)}. \qquad (1)$$

This switch to the unnormalized distribution can improve interpretability as well. If $\tilde{p}(d|m) = 1$, then $p(m|d) = p(m|q)$, which means that the metadata m is observed for the linguistic form d just as often as it is for the baseline query. When $\tilde{p}(d|m) > 1$, the linguistic form is more common for metadata m than average, and when $\tilde{p}(d|m) < 1$, the form is less common for that metadata.[2]

[1] This can be extended to multi-class outcomes; if D has more than two outcomes, each possible outcome is an implicit negative example for the other possible outcomes.

[2] If a normalized distribution is needed, $p(d)$ may be estimable, depending on the data source. In the Twitter data presented here, tweets are sequentially numbered, so $p(d)$ could be estimated using these index numbers. This paper only uses unnormalized distributions.

2.3 Coverage and confidence

Due to the potentially non-uniform distribution of metadata, the amount of error in the estimate in Eq. 1 can vary with m. Intuitively, the confidence in the conditional probability estimates depends on the amount of data observed for each metadata value. Because queries estimate $p(M|d)$ by repeated draws from that distribution, the error in the estimate decreases as the number of draws increases. The overall error in the estimate of $\tilde{p}(d|m)$ decreases as the number of datapoints observed at m increases. This suggests estimating confidence as the square root of the count of observations of the metadata m, as the standard error of the mean decreases in proportion to the square root of the number of observations. More complex Bayesian inference can be used improve error estimates in the future.

3 Sample Implementation: SeeTweet

This section implements the method described in the previous section on a case study of the geographic distributions of linguistic forms, calculated from recent tweets. It is implemented as a suite of novel open-source Python/R programs called SeeTweet, which queries Twitter, obtains tweet locations, performs the mathematical analysis, and maps the results. The suite is available at `http://github.com/gabedoyle/seetweet`.

3.1 SeeTweet goals

Traditionally, sociolinguistic studies are highly time-intensive, and broad coverage is difficult to obtain at reasonable costs. Two data sources that we compare SeeTweet to are the *Atlas of North American English* (Labov et al., 2008, ANAE) and the *Harvard Dialect Survey* (Vaux and Golder, 2003, HDS), both of which obtained high-quality data, but over the course of years. Such studies remain the gold-standard for most purposes, but SeeTweet presents a rapid, cheap, and surprisingly effective alternative for broad coverage on some problems in dialect geography.

3.2 Querying Twitter

SeeTweet queries Twitter through its API, using Mike Verdone's Python Twitter Tools[3]. The API returns the 1000 most recent query-matching tweets or all query-matching tweets within the last week, whichever is smaller, and can be geographically limited to tweets within a certain radius of a center point. In theory, the contiguous United States are covered by a 2500km radius (Twitter's maximum) around the geographic center, approximately $39.8°$N, $98.6°$W, near the Kansas-Nebraska border. In practice, though, such a query only returns tweets from a non-circular region within the Great Plains.

Through trial-and-error, four search centers were found that span the contiguous U.S. with minimal overlap and nearly complete coverage,[4] located near Austin, Kansas City, San Diego, and San Francisco. All results presented here are based on these four search centers. Tweets located outside the U.S. or with unmappable locations are discarded.

The need for multiple queries and the API's tweet limit complicate the analysis. The four searches must be balanced against each other to avoid overrepresenting certain areas, especially in constructing the baseline $p(M)$. If any searches reach the 1000-tweet limit, only the search with the most recent 1000th tweet has all of its tweets used. All tweets before that tweet are removed, balancing the searches by having them all span the same timeframe. Due to the seven-day limit for recent tweets, many searches do not return 1000 hits; if none of the searches max out, all returned tweets are accepted.

3.3 Establishing the baseline

For the baseline query (used to estimate $p(M)$), SeeTweet needs a query with approximately uniform usage across the country. Function or stop words are reasonable candidates for this task. We use the word *I* here, which was chosen as it is common in all American English dialects but not other major languages of the U.S., and it has few obvious alternative forms. Other stop words were tested, but the specific baseline query had little impact on the learned distribution; correlations between maps with *I, of, the* or *a* baselines were all above .97 on both baseline distributions and estimated conditional distributions.

Each tweet from the target query requires its own baseline estimate, as the true distribution of metadata varies over time. For instance, there will be relatively more tweets on the East Coast in

[3] `http://mike.verdone.ca/twitter/`

[4] Northern New England has limited coverage, and the Mountain West returns little data outside the major cities.

the early morning (when much of the West Coast is still asleep). Thus, SeeTweet builds the baseline distribution by querying the baseline term *I*, and using the first 50 tweets preceding each target tweet. This query is performed for each search center for each tweet, with the centers balanced as discussed in the previous section.[5]

3.4 Determining coordinates and mapping

A tweet's geographic information can be specified in many ways. These include coordinates specified by a GPS system ("geotags"), user-specified coordinates, or user specification of a home location whose coordinates can be geocoded. Some tweets may include more that one of these, and SeeTweet uses this hierarchy: geotags are accepted first, followed by user-specified coordinates, followed by user-specified cities. This hierarchy moves from sources with the least noise to the most.

Obtaining coordinates from user-specified locations is done in two steps. First, if the user's location follows a "city, state" format, it is searched for in the US Board on Geographic Names's Geographic Names Information System[6], which matches city names to coordinates. Locations that do not fit the "city, state" format are checked against a manually compiled list of coordinates for 100 major American cities. This second step catches many cities that are sufficiently well-known that a nickname is used for the city (e.g., Philly) and/or the state is omitted.

Tweets whose coordinates cannot be determined by these methods are discarded; this is approximately half of the returned tweets in the experiments discussed here.

This process yields a database of tweet coordinates for each query. To build the probability distributions, SeeTweet uses a two-dimensional Gaussian kernel density estimator. Gaussian distributions account for local geographic dependency and uncertainty in the exact location of a tweeter as well as smoothing the distributions. The standard deviation ("bandwidth") of the kernels is a free parameter, and can be scaled to supply appropriate coverage/granularity of the map. We use 3 degrees (approximately 200 miles) of bandwidth for all maps in this paper, but found consistently high correlation (at least .79 by Hosmer-Lemeshow) to the ANAE data in Sect. 4.1 with bandwidths between 0.5 and 10 degrees.

The KDE estimates probabilities on a grid overlaid on the map; we make each grid box a square one-tenth of a degree on each side and calculate $\tilde{p}(d|m)$ for each box m. SeeTweet maps plot the value of $\tilde{p}(d|M)$ on a color gradient with approximately constant luminosity. Orange indicates high probability of the search term, and blue low probability. Constant luminosity is used so that confidence in the estimate can be represented by opacity; regions with higher confidence in the estimated probability appear more opaque.[7] Unfortunately, this means that the maps will not be informative if printed in black and white.

4 Experiments in dialect geography

Our first goal is to test the SeeTweet results against an existing gold standard in dialect geography; for this, we compare SeeTweet distributions of the *needs done* construction to those found by long-term sociolinguistic studies and show that the quick-and-dirty unsupervised SeeTweet distributions are accurate reflections of the slow-and-clean results. Our second goal is show the importance of using the correct conditional distribution, by comparing it to the unadjusted distribution. With these points established, we then use SeeTweet to create maps of previously uninvestigated problems.

4.1 Method verification on *need* + past participle

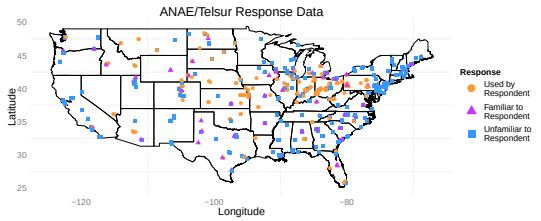
The *Atlas of North American English* (Labov et al., 2008) is the most complete linguistic atlas of American English dialect geography. It focuses on phonological variation, but also includes a small set of lexical/syntactic alternations. One is the *needs* + past participle construction, as in *The car needs (to be) washed*. This construction has a limited geographic distribution, and ANAE provides the first nationwide survey of its usage.

We compare SeeTweet's conditional probabilities for this construction to the ANAE responses to see how the relatively uncontrolled Twitter source compares to the tightly controlled telephone survey data that ANAE reports. We create a SeeTweet

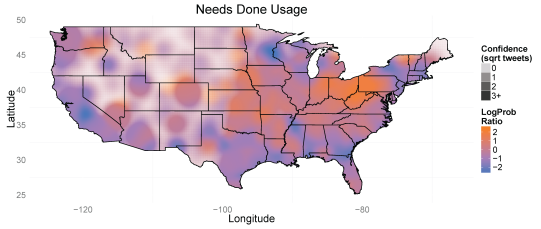
[5] An alternative baseline, perhaps even more intuitive, would be to use some number of sequential tweets preceding the target tweet. However, the Twitter API query mechanism subsamples from the overall set of tweets, so sequential tweets may not follow the same distribution as the queries and would provide an inappropriate baseline.

[6] `http://geonames.usgs.gov/domestic/download_data.htm`

[7] Confidence is given by the square root of the smoothed number of tweets in a grid box m, $p(m|d) * C(d)$.



(a) ANAE/Telsur survey responses for *need*+past participle.



(b) SeeTweet search for "needs done".

Figure 1: Comparing the SeeTweet distribution and ANAE responses for *needs done* usage. Orange indicates higher local usage, purple moderate, and blue lower. Increased opacity indicates more confidence (i.e., more tweets) in a region.

map and visually compare this to the ANAE map, along with a Hosmer-Lemeshow-style analysis. The SeeTweet map is not calibrated to the ANAE map; they are each built independently.

The ANAE map (Fig. 1a) shows the responses of 577 survey participants who were asked about *needs done*. Three possible responses were considered: they used the construction themselves, they did not use it but thought it was used in their area, or they neither used it nor believed it to be used in their area.

The SeeTweet map (Fig. 1b) is built from five searches for the phrase "needs done", yielding 480 positive tweets and 32275 baseline tweets.[8] The component distributions $p(M|d)$ and $p(M)$ are estimated by Gaussian kernels with bandwidth 3. The log of $\tilde{p}(f|M)$, calculated as in Eq. 1, determines the color of a region; orange indicates a higher value, purple a middle (approx. 1) value, and blue a low value. Confidence in the estimate is reflected by opacity; higher opacity indicates higher confidence in the estimate. Confidence values above 3 (corresponding to 9 tweets per bin) are fully opaque. This description holds for all other maps in this paper.

We start with a qualitative comparison of the maps. Both maps show the construction to be most prominent in the area between the Plains states and central Pennsylvania (the North Midland dialect region), with minimal use in New England and Northern California and limited use elsewhere. SeeTweet lacks data in the Mountain West and Great Plains, and ANAE lacks data for Minnesota and surrounding states.[9] The most notable deviation between the maps is that SeeTweet finds the construction more common in the Southeast than ANAE does.

Quantitative comparison is possible by comparing SeeTweet's estimates of the unnormalized conditional probability of *needs done* in a location with the ANAE informants' judgments there. Two such comparisons are shown in Fig. 2.

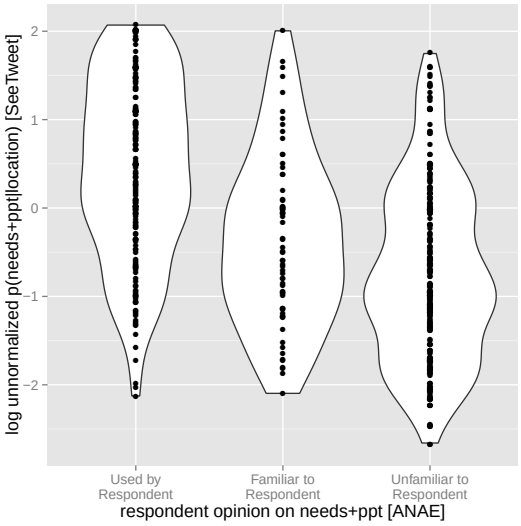
The first comparison (Fig. 2a) is a violin plot with the ANAE divided into the three response categories. The vertical axis represents the SeeTweet estimates, and the width of a violin is proportional to the likelihood of that ANAE response coming from a region of the given SeeTweet estimate. The violins' mass shifts toward regions with lower SeeTweet estimates (down in the graph) as the respondents report decreasing use/familiarity with the construction (moving left to right).

Users of the construction are most likely to come from regions with above-average conditional probability of *needs done*, as seen in the leftmost violin. Non-users, whether familiar with the construction or not, are more likely to come from regions with below-average conditional probability. Non-users who are unfamiliar with it tend to live in regions with the lowest conditional probabilities of the three groups. This shows the expected correspondence trend between the ANAE responses and the estimated prevalence of the construction in an area; the mean SeeTweet estimates for the three groups are 0.45, -0.34, and -0.61, respectively.

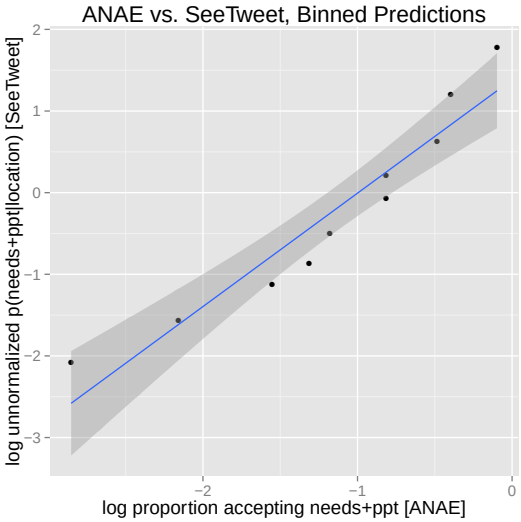
The second comparison (Fig. 2b) is a Hosmer-Lemeshow plot. The respondents are first divided into deciles based on the SeeTweet estimate at their location. Two mean values are calculated for each decile: the mean SeeTweet log-probability

[8] The verb *do* was used as it was found to be the most common verb in corpus work on *needs to be [verbed]* constructions (Doyle and Levy, 2008), appearing almost three times as often as the second-most common verb (*replace*).

[9] Murray et al. (1996)'s data suggest that these untested areas would not use the construction; the SeeTweet data suggests this as well.



(a) Violin plot of SeeTweet estimated conditional probability against ANAE response type.



(b) Hosmer-Lemeshow plot of SeeTweet distribution deciles against average probability of ANAE respondent usage.

Figure 2: Quantifying the relationship between the SeeTweet distribution and ANAE reports for *needs done*.

estimate (increasing with each decile) and the log-proportion of respondents in that decile who use the construction.[10] If SeeTweet estimates of the conditional distribution are an adequate reflection of the ANAE survey data, we should see a tight correlation between the SeeTweet and ANAE values in each decile. The correlation between the two is $R^2 = 0.90$. This is an improvement over the inappropriate conditional distribution $p(M|d)$ that is obtained by smoothing the tweet map without dividing by the overall tweet distribution $p(M)$. Its Hosmer-Lemeshow correlation is $R^2 = 0.79$

These experiments verify two important points: the SeeTweet method can generate data that is tightly correlated with gold-standard data from controlled surveys, and conditionality inversion establishes a more appropriate distribution to correct for different baseline frequencies in tweeting. This second point will be examined further with double modals in the next section.

4.2 Double modals and the importance of the baseline

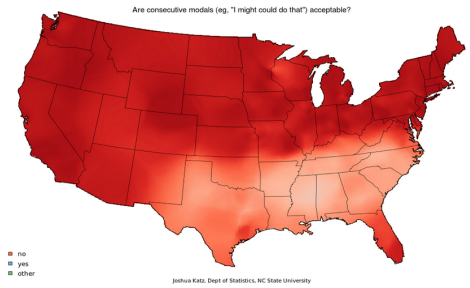
The double modal construction provides a second test case. While ungrammatical in Standard American English, forms like *I might could use your help* are grammatical and common in Southern American dialects. This construction is interesting both for its theoretical syntax implications on the nature of modals as well as the relationship between its sociolinguistic distribution and its pragmatics (Hasty, 2011).

The ANAE does not have data on double modals' distribution, but another large-scale sociolinguistic experiment does: the Harvard Dialect Survey (Vaux and Golder, 2003). This online survey obtained 30788 responses to 122 dialect questions, including the use of double modals. Katz (2013) used a nearest-neighbor model to create a $p(d|M)$ distribution over the contiguous U.S. for double modal usage, mapped in Fig. 3a.[11] Lighter colors indicate higher rates of double modal acceptance.

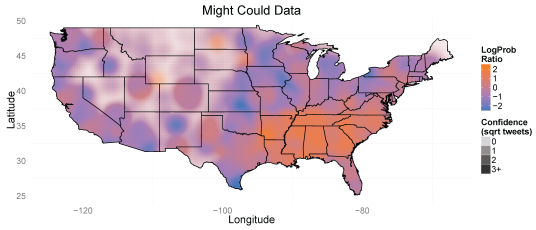
SeeTweet generates a similar map (Fig. 3b), based on three searches with 928 positive and 66272 baseline tweets. As with the ANAE test, the

[10] We remove all respondents who do not use the construction but report it in their area. Such respondents are fairly rare (slightly over 10% of the population), and removing this response converts the data to a binary classification problem appropriate to Hosmer-Lemeshow analysis.

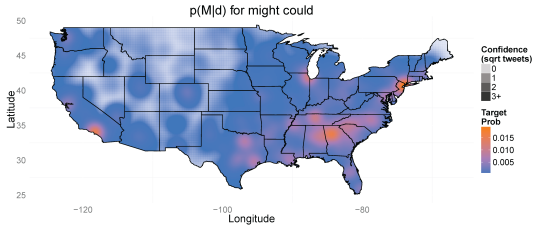
[11] `http://spark.rstudio.com/jkatz/Data/comp-53.png`



(a) Katz's nearest-neighbor estimates of the double modal's distribution in the Harvard Dialect Survey.



(b) SeeTweet distribution for *might could*.



(c) Inappropriate $p(M|d)$ distribution directly estimated from Twitter hits.

Figure 3: Maps of the double modal's distribution.

SeeTweet map is built independently of the HDS data and is not calibrated to it.

The notable difference between the maps is that SeeTweet does not localize double modals as sharply to the Southeast, with pockets in cities throughout the country. This may reflect the difference in the meaning of locations on Twitter and in the HDS; Twitter locations will be a user's current home, whereas the HDS explicitly asks for a respondent's location during their formative years. SeeTweet may partly capture the spread of dialectal features due to migration.

Double modals also provide an illustration of the importance of the Bayesian inversion in Eqn. 1, as shown in Fig. 3c. This map, based on the inappropriate distribution $p(M|d)$, which does not account for the overall distribution $p(M)$, disagrees with general knowledge of the double modal's geography and the HDS map. Although both maps find double modals to be prominent around Atlanta, the inappropriate distribution find New York City, Chicago, and Los Angeles to be the next most prominent double modal regions, with only moderate probability in the rest of the Southeast. This is not incorrect, per se, as these are the sources of many double modal tweets; but these peaks are incidental, as major cities produce more tweets than the rest of the country. This is confirmed by their absence in the HDS map as well as the appropriate SeeTweet map.

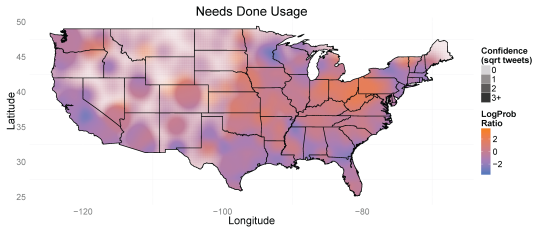
4.3 Extending SeeTweet to new problems

Given SeeTweet's success in mapping *needs done* and double modals, it can also be used to test new questions. An understudied issue in past work on the *need* + past participle construction is its relationship with alternative forms *need to be* + past participle and *need* + present participle. Murray et al. (1996) suggest that their *need* + past participle users reject both alternatives, although it is worth noting that their informants are more accepting of the *to be* alternative, calling it merely "too formal", as opposed to an "odd" or "ungrammatical" opinion about the present participle form. Their analysis of the opinions on alternative forms does not go beyond this anecdotal evidence.

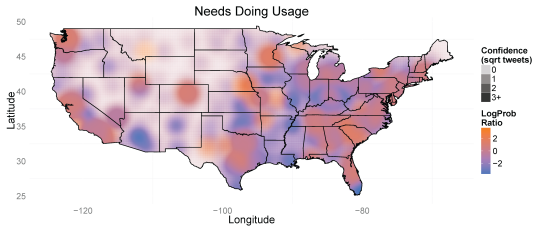
SeeTweet provides the opportunity to examine this issue, and finds that the *to be* form is persistent across the country (Fig. 4c), both in areas with and without the *need* + past participle form, whereas the present participle alternant (Fig. 4b) is strongest in areas where *need* + past participle is not used. Although further analysis is necessary to see if the same people use both the past participle forms, the current data suggests that the bare past participle and bare present participle forms are in complementary distribution, while the *to be* form is acceptable in most locations.

We also compare the alternative constructions to the ANAE data. Using Hosmer-Lemeshow analysis, we find negative correlations: $R^2 = -.65$ for *needs doing* and $R^2 = -.25$ for *needs to be done*. In addition, mean SeeTweet estimates of *needs doing* usage were lower for regions where respondents use *needs done* than for regions where they do not: $-.93$ versus $-.49$.[12] Thus, SeeTweet provides evidence that *needs done* and *needs doing* are in a geographically distinct distribution, while *needs done* and *needs to be done* are at most weakly distinct.

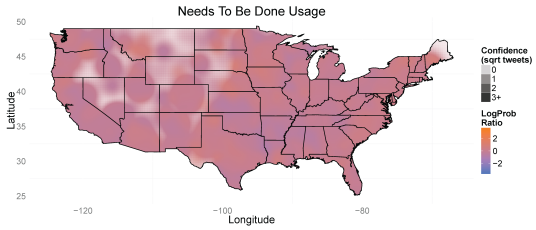
[12] SeeTweet estimates of *needs to be done* usage were comparable in both regions, $-.018$ against $.019$.



(a) "Needs done" distribution



(b) "Needs doing" distribution



(c) "Needs to be done" distribution

Figure 4: SeeTweet distributions for *needs done*, *needs to be done*, and *needs doing*.

5 The appropriateness of Twitter as a data source

A possible concern with this analysis is that Twitter could be a biased and noisy dataset, inappropriate for sociolinguistic investigation. Twitter skews toward the young and slightly toward urbanites (Duggan and Brenner, 2013). However, as young urbanites tend to drive language change (Labov et al., 2008), any such bias would make the results more useful for examining sociolinguistic changes and emergent forms. The informality of the medium also provides unedited writing data that is more reflective of non-standard usage than most corpora, and its large amounts of data in short timescales offers new abilities to track emerging linguistic change.

As for noise in the tweet data and locations, the strong correlations between the gold-standard and SeeTweet results show that, at least for these features, the noise is mitigated by the size of dataset. We examined the impact of noise on the *needs done* dataset by manually inspecting the data for false positives and re-mapping the clean data. Although the false positive rate was 12%, the conditional distribution learned with and without the false positives removed remained tightly correlated, at $R^2 = .94$. The SeeTweet method appears to be robust to false positives, although noisier queries may require manual inspection.

A final point to note is that while the datasets used in constructing these maps are relatively small, they are crucially derived from big data. Because the *needs done* and double modal constructions are quite rare, there would be very few examples in a standard-sized corpus. Only because there are so many tweets are we able to get the hundreds of examples we used in this study.

6 Conclusion

We have shown that Bayesian inversion can be used to build conditional probability distributions over data given metadata from the results of queries on social media, connecting query-derived data to traditional data sources. Tests on Twitter show that such calculations can provide dialect geographies that are well correlated with existing gold-standard sources at a fraction of the time, cost, and effort.

Acknowledgments

We wish to thank Roger Levy, Dan Michel, Emily Morgan, Mark Myslín, Bill Presant, Agatha Ventura, and the reviewers for their advice, suggestions, and testing. This work was supported in part by NSF award 0830535.

References

Gabriel Doyle and Roger Levy. 2008. Environment prototypicality in syntactic alternation. In *Proceedings of the 34th Annual Meeting of the Berkeley Linguistics Society*.

Maeve Duggan and Joanna Brenner. 2013. The demographics of social media users – 2012. Pew Internet and American Life Project.

J. Daniel Hasty. 2011. I might would not say that: A sociolinguistic study of double modal acceptance. In *University of Pennsylvania Working Papers in Linguistics*, volume 17.

Joshua Katz. 2013. Beyond "soda, pop, or coke": Regional dialect variation in the continental US. Retrieved from `http://www4.ncsu.edu/~jakatz2/project-dialect.html`.

William Labov, Sharon Ash, and Charles Boberg. 2008. *The Atlas of North American English. Phonetics, Phonology, and Sound Change*. de Gruyter Mouton.

Thomas Murray, Timothy Frazer, and Beth Lee Simon. 1996. Need + past participle in American English. *American Speech*, 71:255–271.

Daisuke Okanohara and Jun'ichi Tsujii. 2007. A discriminative language model with pseudo-negative samples. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Joshua Tenenbaum and Thomas Griffiths. 2001. Generalization, similiarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24:629–640.

Bert Vaux and Scott Golder. 2003. Harvard dialect survey. Available at `http://www4.uwm.edu/FLL/linguistics/dialect/index.html`.