Predicting Local Socioeconomic Status using Twitter Content Mining and Machine Learning

ABSTRACT

Employing machine learning algorithms, vast quantities of geotagged tweet records can be mined and classified to predict economic, demographic and even cultural traits specific to the place where the messages are originated. Once the effectiveness of these algorithms has been established, they become a valuable tool for the detection and quantification of dynamic urban processes such as gentrification. Algorithm based predictions can flag the presence of gentrification by detecting a turnover in the socioeconomic composition of a place. Social media production occurs continuously. Prediction algorithms can be applied to captured data at frequencies approaching real time, thus offering social analysts and policymakers access to insights with an immediacy that can not be matched by traditional means like Census surveys.

The current project intends to use machine learning classification algorithms to predict the socioeconomic status of the local population based on the attributes of geocoded tweets emitted from the area.

1. INTRODUCTION

The combined effect of social media, big data and mobile telecommunications has opened up a fertile field of investigation. The increasing availability and variety of large-scale digitized records of human interaction has made them a staple of current scientific research. The particular interest posed by software-mediated social interaction records comes from its inherent freshness, frequency and ubiquity. The finer spatial and temporal scale of these data sets permits asking different kinds of questions. Social media applications track the location, activity and interaction of users in real or near-real time. In contrast, Census and other traditional sources of organized data are often several years old by the time of their release, oftentimes tracking only places of residence, and aggregated to arbitrary spatial units [Shelton et al. 2015].

Despite the aforementioned advantages, researchers mining social media datasets face particular challenges. The adoption rate and intensity of use of online mobile applications has not been uniform, resulting in an uneven distribution of content and nuanced digital divides that overlap previous inequalities. Digital divides tend to follow the fault-line of socioeconomic gaps [Stephens 2013]. Uneven geographies of information are the consequence of a wide range of social, economic, and political patterns, practices, and processes. Graham [Graham 2014] describes the persistence of dense clusters of information in many places, and the under-representation of others, caused by the self-perpetuating inequality of knowledge production and transfer.

While it would be normally considered a shortcoming, the uneven representation of people and communities in social media traces can be an advantage for our specific research interest. If patterns of social media production are inherently distinct between advantaged and disadvantaged populations, we can measure and categorize this differential to predict the socioeconomic traits of a place. Twitter is one of the most widely used social networks. It offers micro-blogging service that invites users to express short, 140-character messages. It is functionally a free high-speed global text-messaging service. It reached 320 million active users by the end of 2015 [Ingram 2015]. Because of its popularity and adoption rates all over the world, it functions as a global infrastructure for rapid and easy communication [Russell n.d.].

While users have the option to post messages (or “tweets”) than can only be seen by another specified user, the vast majority of tweets are public. What makes Twitter particularly interesting for socio-spatial research is that tweets that users publish from their mobile phones are geo-tagged: latitude and longitude coordinates are attached to the message. While only a fraction of tweets is geo-tagged, the relative number is growing. In 2014, the Wall Street Journal reported that Twitter users spent 86% of their time with the service accessing through their smartphones. Geo-tagging enables place-specific analysis of Twitter activity, by selecting only the content whose coordinates indicate it was produced inside a given geographical area.

Employing machine learning algorithms, vast quantities of geo-tagged tweet records can be mined and classified to predict economic, demographic and even cultural traits specific to the place where the messages are originated. Once the effectiveness of these algorithms has been established, they become a valuable tool for the detection and quantification of dynamic urban processes such as gentrification. Algorithm based predictions can flag the presence of gentrification by detecting a turnover in the socioeconomic composition of a place. Social media production occurs continuously. Prediction algorithms can be applied to captured data at frequencies approaching real time, thus offering social analysts and policymakers access to insights with an immediacy that can not be matched by traditional means like Census surveys.

Our study is motivated by the following question: “Is it possible to mine publicly available tweet content produced in real time to infer the socioeconomic composition of an urban area?

Our findings indicate that, for the city of Boston, machine learning algorithms fed with social media messages can identify areas of high socioeconomic status with an accuracy rate above 90%, precision > 92% and recall > 96%.

With the goal of facilitating the implementation of such techniques by a wider user base, including planners and social researchers, we advance a computationally efficient methodology that can reproduce our results with consumer-grade hardware.

2. BACKGROUND

Social network data analysis has been successfully employed to characterize neighborhood change based on the weak and strong link social networks of residents and visitors [Hristova et al. 2012]. It has also been shown how social media can reveal changes in the character of neighborhoods. Cranshaw et al. [Cranshaw et al. 2012] studied the large scale social dynamics of a city based on the social media its residents generate. They collected 18 million place check-ins from users of a location-based online social network (Foursquare), and then applied a cluster model to the identify the dynamic areas that comprise the city. In many cases the detected boundaries matched municipal borders, which shows the capacity of neighborhood bounds to influence behavior. However, a few clusters of associated activity spilled across the borders between neighborhoods, indicating a shift in people’s behaviors and perceptions of that area. For example, after conducting interviews with local residents, one of these shifts was attributed to a concerted effort of developers to blur the lines between what were once two very different neighborhoods. By contrasting their findings with the opinion of local experts, the researchers found their methodology able to reveal “subtle changes in local social patterns, and the effects they have on the character of the city.”

3. DATA SOURCES

3.1 Georeferenced tweet database

It consists of 3.6 million tweets, geo-referenced by longitude-latitude coordinate pairs establishing the location where they were sent. The tweets were produced from January to December 2014, and posted from the Boston metropolitan area or its surroundings. The tweet database is available for researchers at Boston Data Library, a collection of administrative data curated by the Boston Area Research Initiative [BARI n.d.].

3.2 American Community Survey

For demographic indices, our research relies on the American Community Survey, or ACS, the ongoing survey conducted by the US Census Bureau. It provides information on a yearly basis about the United States and its population [US Census Bureau n.d.]. The yearly frequency makes ACS information extremely valuable for finer-grained temporal analysis, that would not be feasible using decennial Census information. ACS reports provides information down to the Census tract level for the entire state of Massachusetts, including the City of Boston. ACS data includes 48 demographic indicators, which include and extend those recorded by the decennial Census. Variables of particular interest are those characterizing the socio economic status associated with the population of a given tract.

4. METHODS

4.1 Measuring Socioeconomic Status

The demographic indices from which we extract our socioeconomic status index cover the 2010 to 2014 period. ACS data is presented as five year estimates, taking the average of the period observations to reduce measuring error. Twelve indicators were identified as potentially defining socioeconomic status:

- Median monthly contract rent
- Median home value
- Percentage of persons of white race, not Hispanic origin
- Percentage of persons with at least a four-year college degree
- Percent unemployed
- Percentage of professional employees (by occupations)
- Percent poor
- Percentage of household heads moved into unit less than 10 years ago
- Percentage of persons age 17 years and under
- Percentage of vacant housing units
- Percentage of owner-occupied housing units
- Percentage of persons age 60 years and over

Only Census tracts in the City of Boston where more than 500 people have their residence where considered for the present study (n = 167).

We employed a factor analysis technique to identify a reduced set of dimensions, or variables, that can explain most of the variance between observations. Factor analysis cannot be performed when missing values are present. This was the case for median home value in 7 Census tracts. To address this issue, predictive mean matching [Horton and Lipsitz 2001] was used to estimate the most likely value for the missing attributes.

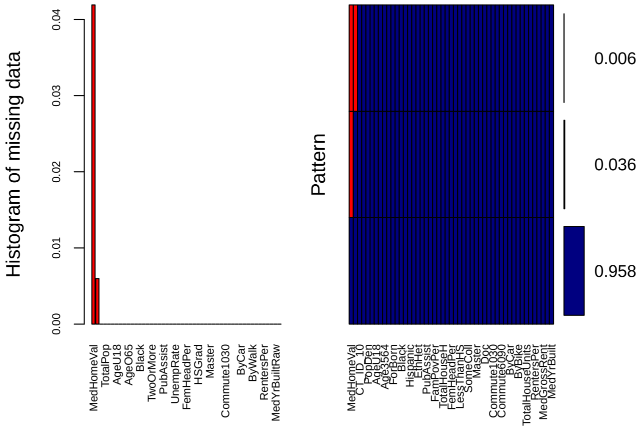

Figure 1: Missing ACS data

Factoral complexity is the number of variables loading on a given factor. Ideally a variable should only load on one factor, complexity 1. This means, theoretically, that it represents an accurate conceptual determination of how the variables will group. The logical extension of this is that it provides a relatively accurate measure of the underlying dimension or concept that is the focus of research; in our case, the socioeconomic status of the inhabitants of each Census tract.

The Very Simple Structure criterion (VSS) is an exploratory data analysis tool used to establish the optimal number of factors [Ruscio and Roche 2012]. The VSS plot for our data suggested a complexity 1 solution with a single factor as optimal. That is, a single factor was the closet to 100% of the variance explained for a complexity 1 solution.

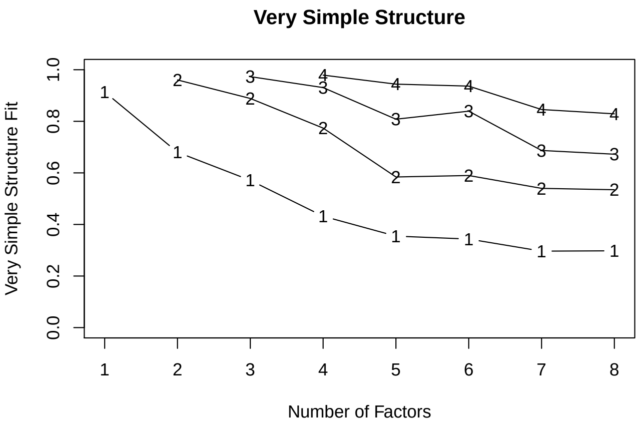

Figure 2: VSS for ACS 2010 - 2014 socioeconomic variables, Boston

The variables that form the reduced set are (1) Median household income, (2) Median monthly contract rent, (3) Median home value, (4) Percent of home ownership, (5) Percent of persons with at least a four-year (bachelor's) college degree, (6) Percent unemployed, (7) Percent of families with income falling below the poverty line, and (8) Percent of the population receiving public assistance.

As expected, the factor loadings for percentage unemployed, percentage under the poverty line and percent receiving public assistance pull in the opposite direction than the other variables. This is consistent with the notion that these three variables are associated with a lower socioeconomic status level, and the rest with a higher status.

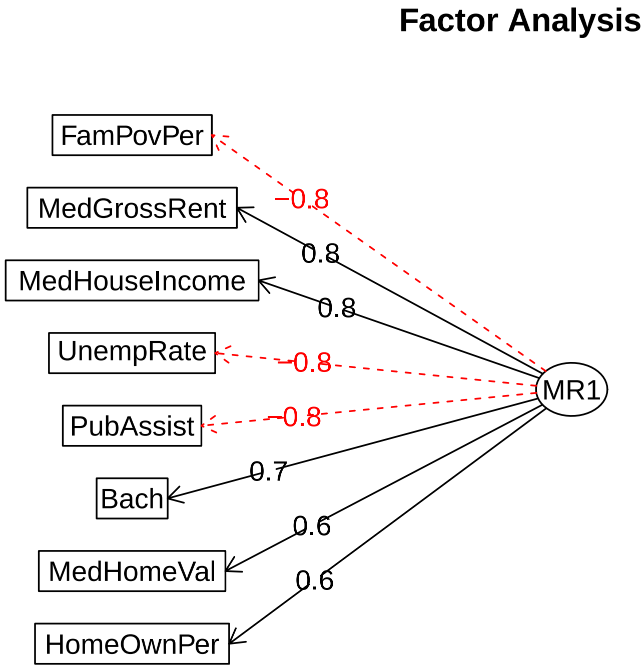

Figure 3: Factor loadings

The main factors have the following summary statistics (note the wide differences in scale and variance):

Table 1: Summary statistics for main SES factors

Statistic	Mean	Median	St. Dev.
MedHouseIncome	58,358.620	53,065	29,491.350
MedGrossRent	1,296.659	1,291	397.283
MedHomeVal	401,180.200	344,600	154,952.600
HomeOwnPer	0.333	0.321	0.196
Bach	0.248	0.246	0.133
UnempRate	0.102	0.097	0.059
FamPovPer	0.167	0.145	0.127
PubAssist	0.045	0.035	0.039

To combine the variables that compose the factor as a single attribute, we calculated an index variable via an optimally-weighted linear combination of the dimensions (factor scoring). We divide the weighted sum scores for the factor by the sum of the variable loadings in that factor.

$$SES_i = \frac{\sum loading_j x_{ij}}{\sum loading_j}$$

To account for the wide range of sale and the diversity of units of measure, the variables are normalized before being weighted (z-scores). The resulting index is distributed as follows:

Table 2: Summary statistics for SES index

Statistic	Mean	Median	St. Dev.
SESindex	0.000	0.129	4.185

Plotted on a map of Boston, the SES index variable shows the fragmentation of the city into low and high SES areas.

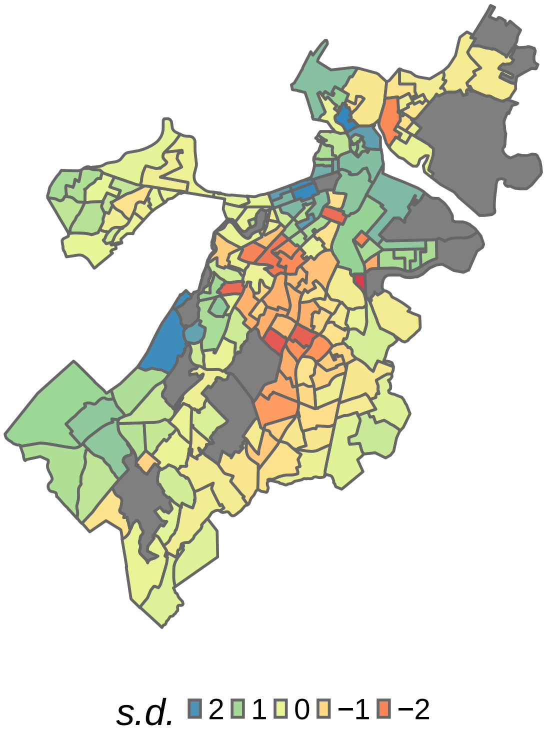

Figure 4: Boston Socioeconomic Status index, deviations above or below the mean

To enable later classification using a machine learning algorithm, we also create a dummy variable that labels certain Census tracts as "High SES". The condition to receive the High SES label is having a SES index value at least one standard deviation above the city mean.

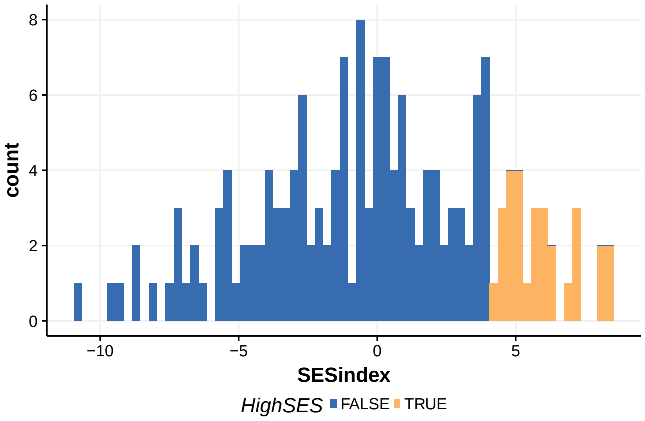

Figure 5: SES histogram, per tract

4.2 Preparing Twitter data for analysis

The source for Twitter data is a corpus of geo-referenced tweets captured between January and December 2015 (n = 3604352). Via geo-spatial processing, tweets posted from within the city limits were isolated. This smaller dataset contains 1180939 observations.

To ensure that Tweets were produced by human users (in opposition to automated software systems) the dataset was filtered to include only tweets originated from one of four sources: iOS devices, Android Devices, Instagram or Foursquare. The results (n = 921933) comprise the observation used for machine learning labeling.

Six additional dummy variables (taking a value of either 1 or 0 for "yes" or "no") were created: "iOS", "Android", "Foursquare", "Instagram", "atnight", "onweekend". The last two variables refer to the moment when the tweets were created.

We decided to limit our area of analysis to residential tracts. This was necessary to filter out non-residential areas such as airports, university campuses and downtown districts, which by nature are frequented by a considerable number of visitors whose socioeconomic status not necessarily matches that of the tract permanent residents.

Finally, the "High SES" label variable was appended to the data to act as the classification criteria.

Table 3: Percentage of tweets with derived attributes

Statistic	Mean
at_night	0.540
onweekend	0.290
iOS	0.619
Android	0.231
Instagram	0.124
Foursquare	0.026
HighSES	0.172

4.3 Text mining

After subselection, the remaining set of tweets (n = 398413) was fed into a statistical text analysis pipeline.

Table 4: Text content from a sample of 5 tweets

	Text
1	@recentpoker @thereaIbanksy how many of
2	@elgranblack @kaoscdenblanco @elgranblac
3	Look outside, Boston sunset is killing i
4	How u marry me and still make me a back
5	@ActuallyNPH You are unbelievable great!

Tweet content was mined to compile a matrix with all found terms and their frequencies. All words were converted to lower case. Punctuation and URLS were removed. "Stop words", or common English terms such as "the", "at", "in" were removed as well. The remaining words were "stemmed", combining families of word into their root terms; i.e. "mining" and "miner" being counted as two instances of the same root term, "mine". Finally, only terms with a frequency above 99.8% of all found words were kept, to reduce the computational effort required for posterior analysis.

The resulting term frequency matrix contains 398 terms, with a number of appearances in each tweet in our dataset ranging from 0 to 27.

4.4 Predicting through logistic regression

Each row representing a tweet in our text frequency matrix was paired with "HighSES", a binary label (TRUE | FALSE) expressing whether the tweet was published from a Census tract with a high socioeconomic status. The label represented our outcome variable, while the term frequencies represented the predictors.

After preliminary testing of classification algorithms, binomial logistic regression produced results of accuracy comparable to more sophisticated methods such as Support Vector Machine classification. Given that the computational complexity of logistic regression grows linearly with the number is items in the dataset ($O(n)$) [Iyer 2015], while SVM grows exponentially ($O(n^2)$) [Bottou and Lin 2007], the former was chosen for our classification. In practice, other machine learning algorithms were deemed nonviable the task of processing our dataset without the use of distributed computation or a large scale computing system.

The odds of a tweet having been sent from a "High SES" tracts were thus calculated as:

$$log(\frac{P(y=1)}{1-P(y=1)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \epsilon$$

Where $x_1, x_2, ...x_{398}$ are the counts per tweet of each of our 398 words, and $\beta_1, \beta_2, ...\beta_{398}$ their effects on the odds.

Given that binomial logistic regression estimates the chance of an outcome being true, the probability estimates allowed us to fine tune the results and reduce false positives by setting a probability threshold for a tweet to be considered "HighSES".

We trained our logistic model on 2/3 of the items on our dataset (n = 398413).

5. RESULTS

Almost a half the terms (191) were found to have a statistical significant correlation with the odds of the text being emitted from a high SES Census tract.

Many of the terms with large effects refer to place names, signaling the association between specific areas in the city with high or low levels of socioeconomic status. Unexpectedly, neutral sounding terms like "bar" or "street" have a significant positive correlation, while "midnight" appears among the terms with highly negative effects.

Table 5: Top ten terms with positive High SES correlation

	Estimate	Std..Error	z.value	Pr...z..	Term
33	4.60	0.21	22.09	0.00	copley
112	4.40	0.21	21.40	0.00	newbury
13	3.75	0.23	16.17	0.00	beacon
102	1.55	0.08	18.73	0.00	massachusetts
158	1.47	0.06	24.36	0.00	street
12	1.19	0.09	13.68	0.00	bay
156	1.16	0.09	13.29	0.00	state
77	0.99	0.08	12.75	0.00	hill
11	0.98	0.07	14.96	0.00	bar
122	0.93	0.07	14.23	0.00	patriots

Table 6: Top ten terms with negative High SES correlation

	Estimate	Std..Error	z.value	Pr...z..	Term
42	-3.77	0.34	-11.20	0.00	dorchester
23	-3.68	0.37	-9.85	0.00	brighton
3	-2.74	0.32	-8.53	0.00	allston
116	-1.98	0.18	-11.16	0.00	niggas
44	-1.72	0.15	-11.67	0.00	east
115	-1.38	0.10	-13.19	0.00	nigga
157	-1.38	0.14	-10.03	0.00	station
95	-1.34	0.16	-8.41	0.00	lmfao
105	-1.06	0.15	-7.08	0.00	midnight
18	-0.88	0.14	-6.12	0.00	bitches

As effect interpretation is not the goal of our study, but classification, we turn to confusion matrices to assess performance. Our focus was on minimizing false positives, sacrificing recall but ensuring specificity (minimal false positives). We are not predicting the socioeconomic status of the individuals that produce the tweets; instead, we are trying to infer the characteristics of the place from where the tweet is being published. The rationale was to place a high confidence on the detection of HighSES tweets, and then use the ratio of total HighSES tweets detected in a census tract as a predictor of its socioeconomic status.

For logistic regression, it is common to classify an outcome as TRUE if its odds are above 0.5, and FALSE otherwise (Table 9). We found that rising the threshold up to 0.9, the precision incremented at a higher than the loss of recall (Table 10).

Table 7: Tweet origin: prediction accuracy at threshold 0.5

	FALSE	TRUE
FALSE	109075	808
TRUE	20323	2599

With our classification threshold set at 0.9, we sum the total of tweets labeled as HighSES per tract, and divide them by the total amount of tweets published there to obtain a ratio. This aggregated measurement is used for a new logistic regression model, with an outcome variable now at the Census tract level instead of individual tweets. To improve the accuracy of the model, we also include aggregated ratios of our Twitter-derived variables.

Table 8: Tweet origin: prediction accuracy at threshold 0.9

	FALSE	TRUE
FALSE	109835	48
TRUE	21454	1468

The regression table for our tract-level SES prediction model shows the occurrence of manifold statistically significant correlations (Table 9; confidence intervals between parentheses):

While all of the variables that we derived from tweet metadata help increase the accuracy of the model, three of them are particularly interesting. Ratio of tweets from an iOS device, ratio of Foursquare and ratio of Instagram posts have a statistically significant effect on socioeconomic status, suggesting the existence of strong links between social media output and local status.

As a classifier, the tract-level logistic model yields an overall accuracy slightly above 90 %, with few false positives. Precision was good at 92%, at the cost of missing some true positives (Specificity only slightly above 60%). Still, the trade-off is acceptable in terms of our objective: explore method for early detection rises in local urban socioeconomic level, such as those associated with rapid gentrification.

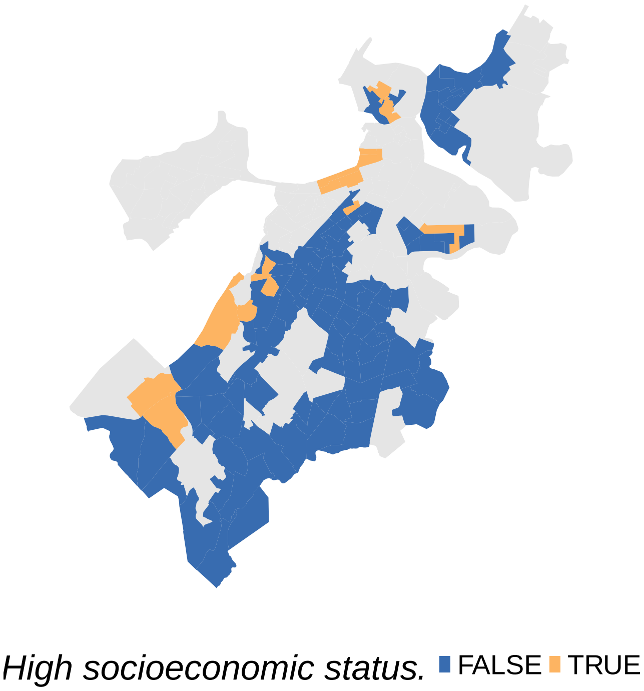

Figure 6: SES status, Boston residential tracts (Measured)

Table 9: Tract-level logistic regression model

	Dependent variable:
	HighSES
n	0.0004 (0.002)
unique_users	-0.013^{*} (0.008)
tweets_per_user	-6.876 (7.680)
lang_diversity	0.063 (0.229)
en_ratio	5.949 (6.483)
median_followers	-0.0002 (0.001)
median_following	
median_friends	0.0001 (0.002)
median_tweet_count	-0.0001 (0.0001)
at_night_ratio	6.167 (10.159)
onweekend_ratio	12.933 (9.909)
iOS_ratio	11.244^{**} (5.147)
instagram_ratio	22.004^{**} (8.990)
foursquare_ratio	60.460^{*} (30.865)
predicted_ratio	563.393^{**} (234.165)
Constant	-20.297^{*} (11.718)
Observations	110
Log Likelihood	-21.832
Akaike Inf. Crit.	73.663
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 10: Tract SES status: prediction performance by class

	Measure	Performance
1	Sensitivity	0.97
2	Specificity	0.61
3	Pos Pred Value	0.93
4	Neg Pred Value	0.79
5	Precision	0.93
6	Recall	0.97
7	F1	0.95
8	Prevalence	0.84
9	Detection Rate	0.81
10	Detection Prevalence	0.87
11	Balanced Accuracy	0.79

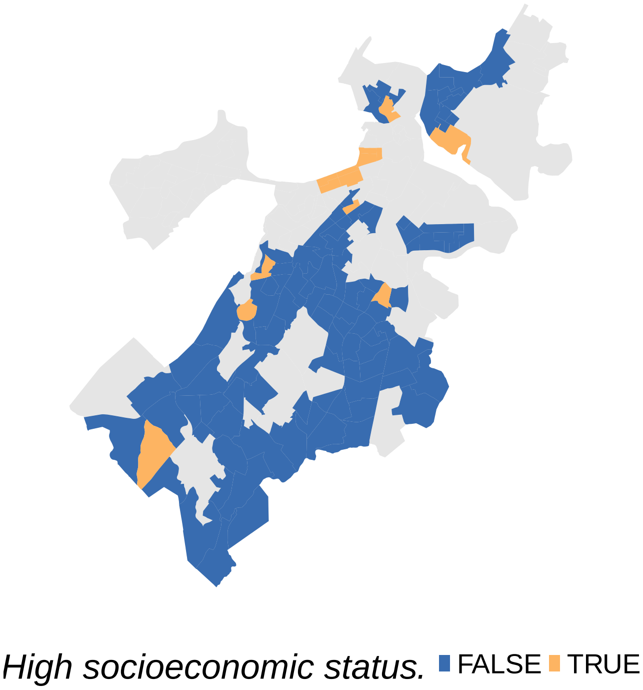

Figure 7: SES status, Boston residential tracts (Predicted)

6. CONCLUSION

We have described a methodology to reduce the multiple dimensions of Census demographic data as a single index of socioeconomic status. We achieved this by finding the linear combination of dimensions, or factor, that explains per se most of the variance between observations. By setting an arbitrary threshold, we also labeled the set of Census tracts with the highest SES level. This "High SES" status became our outcome variable, measured within the City of Boston.

As input for our SES prediction, we selected tweets originated from inside Boston's residential areas. We extracted new measurements from the tweet collection based on their metadata. We also applied unstructured text mining techniques to generate a word frequency matrix representing the most common terms found in the tweets content.

We created a predictive model to classify tweets according to the SES status of their place of origin. We established the advantage of logistic regression as the machine learning algorithm of choice, given it's performance at low computational cost.

Analysis of the term frequency matrix and its correlation with tweet origin shows how the classification algorithm easily picks how certain neighborhood names have either highly positive or highly negative associations with SES.

We fed the results of this model into a second one, that predicted whether a Census tract has a high socioeconomic status as measured by our Census based index. By adding our variables derived from twitter metadata, we were able to infer correlation between social media use patterns and local socioeconomic status. In particular, we found that the use of iOS devices, as well as the use of both Instagram and Yelp social networks is statistically correlated with high SES.

Our second model fulfills our goal, predicting which Census tract have the population with the highest levels of socioeconomic status in the city. The model performs well, with very few false positives and an overall accuracy of 90%. These results invite further research, extending our current findings both by exploiting novel datasets of online activity, and by predicting other social metrics.

7. REFERENCES

BARI. Boston Data Portal. Retrieved December 11, 2016 from `https://www.northeastern.edu/csshresearch/boston`arearesearchinitiative/boston-data-portal/index.htm

Léon Bottou and Chih-Jen Lin. 2007. Support vector machine solvers. Large scale kernel machines (2007), 301–320.

US Census Bureau. About the Survey. Retrieved December 11, 2016 from `https://www.census.gov/programs-surveys/acs/about.html`

Justin Cranshaw, Raz Schwartz, Jason I. Hong, and Norman Sadeh. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. In International AAAI Conference on Weblogs and Social Media. 58.

Mark Graham. 2014. The Knowledge Based Economy and Digital Divisions of Labour, Rochester, NY: Social Science Research Network.

Matthew Ingram. 2015. Twitter has a serious growth problem. Fortune (October 2015).

Nicholas J. Horton and Stuart R. Lipsitz. 2001. Multiple Imputation in Practice. The American Statistician 55, 3 (August 2001), 244–254. `DOI:https://doi.org/10.1198/000313001317098266`

Desislava Hristova, Afra Mashhadi, Giovanni Quattrone, and Licia Capra. 2012. Mapping community engagement with urban crowd-sourcing. In Proceedings of When the City Meets the Citizen Workshop, Dublin, Ireland.

Karthik Thyagarajan Iyer. 2015. Computational complexity of data mining algorithms used in fraud detection. (August 2015).

John Ruscio and Brendan Roche. 2012. Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. Psychol Assess 24, 2 (June 2012), 282–292. DOI:https://doi.org/10.1037/a0025697

Matthew A. Russell. 2013. Mining the Social Web, O'Reilly Media, Inc.

Taylor Shelton, Ate Poorthuis, and Matthew Zook. 2015. Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. Landscape and Urban Planning 142 (October 2015), 198–211. DOI:https://doi.org/10.1016/j.landurbplan.2015.02.020

Monica Stephens. 2013. Gender and the GeoWeb: divisions in the production of user-generated cartographic information. GeoJournal 78, 6 (December 2013), 981–996. DOI:https://doi.org/10.1007/s10708-013-9492-z