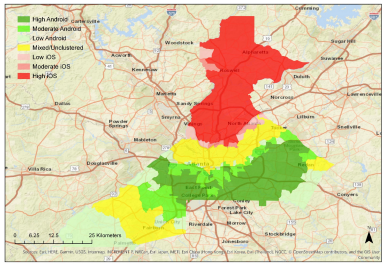
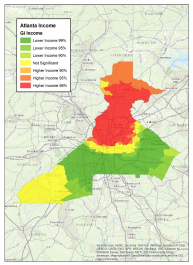
How Good is Twitter Data: Comparing Phone Usage Patterns Using Social Media vs. Calculated Home Location

Allen Finchum and Matthew Haffner
Department of Geography
Oklahoma State University

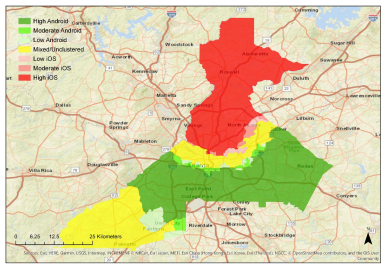
Atlanta

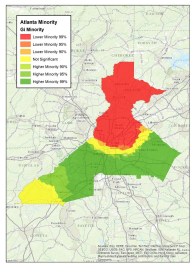
Atlanta Phone OS Usage - Twitter



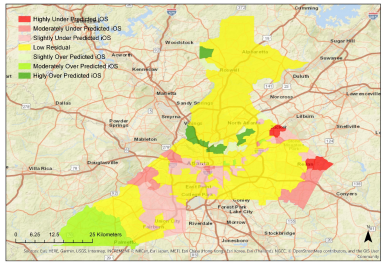


Atlanta Phone OS Usage - Calculated Home





Atlanta Phone OS Usage - Hot Spot Residuals



Statistical Analysis

Pearson's R	.9553
Multiple R^2	.9126
Adjusted R^2	.9124

These maps show two core counties of the Atlanta Metropolitan Area (Fulton and DeKalb counties). These counties are comprised of 346 census tracts which were used for this analysis. In reviewing the two hot spot maps we see a strong similarity between the hot spot maps. The Pearson's R for between the hot spot classifications is .9553 and the Multiple R^2 is .9124.

The regression was built on using the earlier Twitter iOS percentage to "predict" the Calculated Home Location iOS percentage used to develop the hot spot calculations. The residual map shows those areas where the Home Location map differs from the Twitter map, and in which "direction" these differences exist.

The two hot spot maps for Atlanta show a very strong similarity and the statistical analysis simply corroborates the visual observations.

Discussion

Introduction: In this project we were attempting to determine what differences and similarities exist between readily available Twitter data and a proprietary dataset produced by a commercial provider of home locations for user devices. The publically available data itself is culled from Twitter (Geo-Coded Tweets), which we accept are not representative of the entire population, but they can represent a portion of the population – one that marketing and advertising are keen to understand – 18-35 year old technically savvy individuals. The proprietary data was acquired from Safegraph Inc of San Francisco and is based on repeated usage of cooperating third party applications.

We scraped the Twitter feed using Tweepy to cull out geo-coded tweets for one full week (24 hour/7 days) to compile over 240,000 tweets in the two cities shown here. These were then joined to census tracts to match with generalized demographic information.

The proprietary data from Safegraph is based on following location data from over 50 cooperating applications (i.e. third party weather applications) that use the location information from the device. After following the device for at least one month custom algorithms are used to determine a "home" location for the device based on repeated visits, time of day, and land-uses of the visited locations.

Data Discussion: Due to the proprietary nature of the data we are not allowed to give specifics of the algorithms other than to state that based on sampled users the "home" location is correct up to 90% of the time depending on locale specifics. There are some general variations between these datasets due to the timing of when the data was gathered based on overall phone type usage changes – between 2015 and 2017 when the two datasets were developed the overall proportion of iOS vs Android usage has seen a change of approximately 5 percent.

Hot Spot Analysis: At this point we used a Getis-Ord Gi Hot Spot Analysis to determine generic patterns for all of the variables based on each dataset. Use of this analysis tool is described below:

"For some tools, like Hot Spot Analysis, a fixed distance band is the default conceptualization of spatial relationships (we used this approach). With the Fixed Distance Band option, you impose a sphere of influence, or moving window conceptual model of spatial interactions onto the data. Each feature is analyzed within the context of those neighboring features located within the distance you specify for Distance Band or Threshold Distance. Neighbors within the specified distance are weighted equally. Features outside the specified distance do not influence calculations (their weight is zero). Use the Fixed Distance Band method when you want to evaluate the statistical properties of your data at a particular (fixed) spatial scale. If you are studying commuting patterns and know that the average journey to work is 15 miles, for example, you may want to use a 15-mile fixed distance for your analysis. We felt this approach would provide a basic understanding of the spatial patterns within our various types of data". (ESRI, 2015)

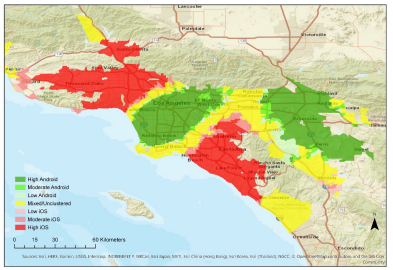
As can be seen in the results for Atlanta and Los Angeles we believe that these two data sources provide similar, albeit slightly different, results. However, since our overall analysis goal is on a macro level for metropolitan areas we feel that the results from both cities demonstrate that the publically available Twitter data provides similar results to those provided from the commercially produced data that under normal circumstances would be prohibitively expensive for most academic research projects.

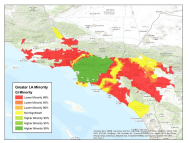
ACKNOWLEDGEMENTS:
We wish to thank the following individuals from Oklahoma State University for their advice and assistance: Jon Comer, Michael Larson, and Nick Rose of the OSU Department of Geography.

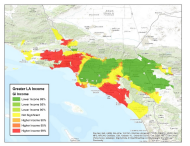
Data Sources: American Community Survey, U.S. Census, 2013 and Twitter API/Tweepy. Custom Home Location Data produced by Safegraph for July 2017.

Los Angeles

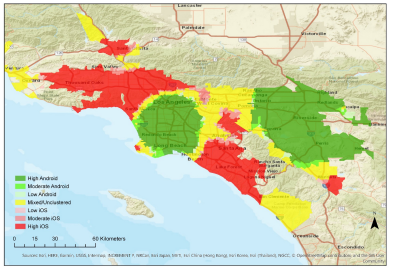
Los Angeles Phone OS Usage - Twitter



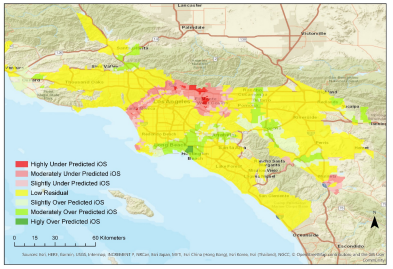




Los Angeles Phone OS Usage - Calculated Home



Los Angeles Phone OS Usage - Hot Spot Residuals



Statistical Analysis

Pearson's R	.8279
Multiple R^2	.6855
Adjusted R^2	.6854

These maps show the primary urbanized area of the Los Angeles Metropolitan Area as defined by United States Department of Transportation. This urbanized area is comprised of 3542 census tracts which were used for this analysis. In reviewing the two hot spot maps we see a strong overall similarity, but there are some variations as well. The Pearson's R for between the hot spot classifications is .8279 and the Multiple R^2 is .6855.

The regression was built on using the earlier Twitter iOS percentage to "predict" the Calculated Home Location iOS percentage used to develop the hot spot calculations. The residual map shows those areas where the Home Location map differs from the Twitter map, and in which "direction" these differences exist.

While the Los Angeles map does not demonstrate the level of correlation between the hot spots as in Atlanta these maps still demonstrate a relatively strong similarity, especially in the "core" of each hot spot as can be seen on the maps.