Discover mobility of twitter users

Are geo-tagged tweets a good proxy of all commuters in London?

Intro

Traditional travel behaviour surveys provide invaluable insights for supporting human mobility analysis and forecast more effective use of transport infrastructure for policy makers. However, few institutions could afford yearly or more frequent surveys because traditional ones rely on interviews and home visits (Zhang, He and Zhu, 2017), being time-consuming, labour intensive and costly.

The emerging location based social networks give a faster and more affordable alternative to achieve similar goals. In particular, Twitter is one of the most popular platforms for GIS researchers as it allows users to specify location with tweets (hereinafter called “geo-tagged tweets”), and grants researchers access to its data by its Application Programming Interfaces (API).

. With care to filter and stream location tagged tweets,

Therein lies the problem of the composition of users

This paper presents a simplified approach to examine whether the geo-tagged tweets are a good proxy for the overall flow of London commuters, by clustering those tweets using DBSCAN for preliminary observations and conducting spatial regression for validation.

Literature review

Much has been written to extract user movements from geo-tagged tweets to fulfil the goals of traditional transportation surveys, for example, extracting popular trajectories of users, predicting whereabouts of people

Most of them do cluster with various modifications like:

Major drawbacks

They focus on extracting general mobility information related to a particular urban area or region without distinguishing the time of the day in which the information was generated (time slots). Consequently, existing solutions do not study the relationship between the moment of the day when social-media documents are posted and its associated spatial place. This missing information could provide a global vision of the movement of a population along the day. Therefore, these works are not taking full advance of social-media datasets.

Most of the employed algorithms do not consider the fuzzy and noisy nature of the kind of data generated by humans. Thus, the results obtained are not as precise as they could be.

The above approaches

This paper takes a simpler approach

Methodology

Scrape tweets (<https://towardsdatascience.com/how-to-scrape-more-information-from-tweets-on-twitter-44fd540b8a1f#aa2a>)

density-based spatial clustering of applications with noise is used to

DBSCAN has two parameters, eps, minpts

This approach minimal user input (pre assumptions), ripley’s k infer eps

knn to infer both parameters.

10-fold CV mean dividing your training dataset randomly into 10 parts and then using each of 10 parts as testing dataset for the model trained on other 9. We take the average of the 10 error terms thus obtained.

In 5 repeats of 10-fold CV, we’ll perform the average of 5 error terms obtained by performing 10 fold CV five times. Important thing to note is that 5 repeats of 10 fold CV is not same as 50 fold CV.

Discussion

Conclusion

Do this analysis for longer period

Contains details about the location tagged by the user in this Tweet, if they specified one.

However, there are limitations in using tweets in social science studies since the data may be biased for various reasons.

socioeconomic structure of Twitter users

Twitter users may not be a complete representative of the public.

the tweet data used in this study are only the geo-tagged ones, which are a small part of all tweets that the Streaming API can collect and even a tiny part of the entire Twitter dataset.

As such, the discovered knowledge may only reflect the human activity and mobility patterns for a portion of the total population, e.g., college students. The findings found here in college cities in the Midwest of the U.S. may not be fully observed for other places with different urban and demographic structures.

We foresee that the findings of this research can be used to train advanced machine learning techniques to infer the patterns and activities of a larger population of Twitter users, including those who opt-out geo-tagging. As a result, the established method can then be expected to study human dynamics of general public. In addition, inspired from this research using Twitter data, when more social media, cellphone, card transaction, and other data from a broader population are available, we could build deeper and more complete insights on mobility patterns of general public. Another possible and necessary future direction of this type of research is to utilize the content of tweets for text mining, topic modeling, and natural language processing so as to discover deeper knowledge and patterns about human and human behavior. This can facilitate the understanding on the nature of user’s activities and the functions of places where frequent tweeting occurs. It can also help to detect space-time tweet clusters and infer the types of gatherings or events. Finally, we may investigate the possibilities of applying the found spatial and temporal patterns into broader fields such as traffic planning, market analysis, urban development, politics, and social science studies.