UNIVERSITY OF CAMBRIDGE
Centre for Digital Built Britain

2017/18 Mini-Project

Crowdsourcing data in mining spatial urban activities: the case of multi-dimensional analysis of Urban Segregation in Cambridge and Ningbo

Final Report

Elisabete A. Silva, PI, Reader in Department of Land Economy [es424@cam.ac.uk]
Haifeng Niu, CO-I, Ph.D student in Department of Land Economy [hn303@cam.ac.uk]

Abstract
“Crowdsourcing data” has been generated in great quantities with the development of the information and communications technology (ICT) from large and diverse groups of people or internet users. These new non-traditional (i.e. census data) datasets also have been introduced as a data source for urban analysis in recent studies. The new data sets provide geo-coded geographic information for spatial analysis but also contain urban human behaviour characteristics that enrich the quality of the positional data that we acquire (i.e. trajectory from continuous GPS records, emotions from social media content, the perception from geo-tagged photos, etc.). This project focuses on the crowdsourcing data harvesting and data-mining of the multi-dimensional mechanisms of urban segregation combining the geo-coding of information with the abundant attributes of this type of data. This project conducts pilots at Cambridge in the UK and then compare it with prior study of Ningbo in China trying to synchronise some of the data collection methods across the two case studies. We realized that by utilising crowdsourcing data, it can overcome some of the limitations of geographic data and provide insights into socio-economic mechanisms behind the spatial-temporal dimension of urban behaviours. Also, to extend the research focus to social-spatial and economic-spatial characteristics instead of the spatial structure, this research provides a conceptual and methodological framework for analysing crowdsourcing data that is more sensitive to the social and economic relations embodied in spatial-temporal behaviours.

Research Question
1. How does check-in data from social media is distributed around Cambridge? What kinds of spatial segmentation could be identified?
2. How to validate the social media data on urban segregation? And how to analysis it socially and economically with other data sources such as questionnaires?
3. What are different findings between case studies in Cambridge, UK and Ningbo, China?

Methodology
This project focuses on the crowdsourcing data harvesting and data-mining of the multi-dimensional mechanisms of urban segregation combining the geo-coding of information with the rich attributes of this type of data. This project will conduct pilots at Cambridge in the UK and then compare it with prior study of Ningbo in China trying to synchronize some of the data collection methods across the two case studies.

Firstly (goal 1), based on an understanding of the spatial fragmentation of urban districts, specific urban matrices are selected to present the spatial features of Cambridge. Next (goal 2), user-generated content (UGC) social media and images data are collected to characterise the social and built environment in different parts of Cambridge to assist in finding the link between social segregation and the built environment. For both goals in Cambridge previous work done in Ningbo, China will allow to compare and contrast realities.

Thereafter, in a second stage, we validated the above ‘big data’ approach with data collected by ‘eyes on the street’ type of questionnaires (soft data collection) and will also perform smartphone detection (linking mixed methods of qualitative/quantitative approaches) (goal 3). This phase in the study of urban segregation answered the common criticism that crowdsourcing doesn’t capture important groups of society because these groups don’t own or use the devices producing such data (this is particularly important in low income and jobless groups of society). While this is a mini project pilot study, the questionnaires needed to be performed for both Cambridge and Ningbo in China in order to synchronize methodologies.

Lastly, as a final step, a comparison between two historical cities, Cambridge in UK and Ningbo in China was performed, it allowed us to summarize the key features of urban segregation and extract the general principles.

Discussion

The starting point of this research was based on the data harvesting from social media. The main goal was to be able to link social media activities to the built environment. **Image 1** points to the England vs Cambridge production of data and social media activity and **Image 2** points to the Cambridge city centre social media activity.

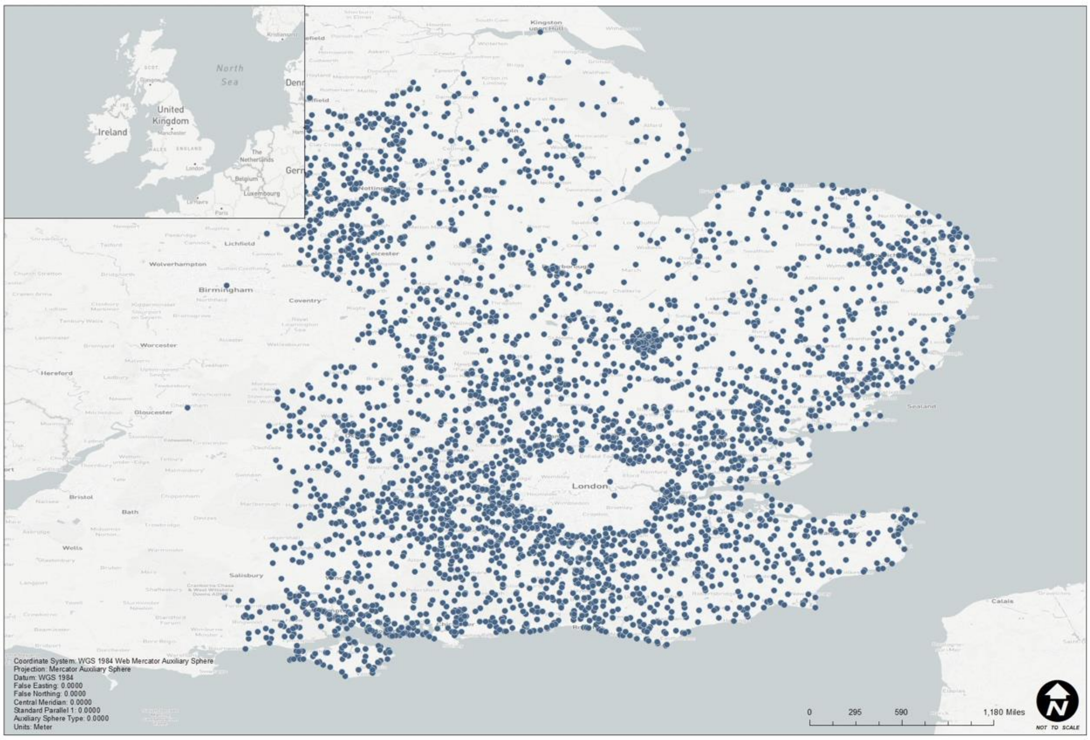

Figure 1. Social media extracted using API developed for this research

With the completion of this phase of information harvesting and analysis (performed during the first month of the project) we were able to set the foundation for the next phase: identification of areas to sample people using questionnaires and for the location of the mobile telecommunication devices (performed during the second month of the project).

By using open developer API from Twitter, we collected data from tweets during 8th February to 28th March. Among those tweets, 37497 tweets with geo-tag (geographic coordinate) are refined with data cleaning script, distributing through Eastern England except for the Great London. To get the geo-tagged tweets from Cambridge, we add a location filter as *locations=[0.068639,52.15794,0.184552,52.237228]* to narrow down the dataset, and amount of tweets in Cambridge is 2338. Based on this, we introduced kernel analysis on the ArcGIS platform and generated a tweets heat map as showed as **Image 2**.

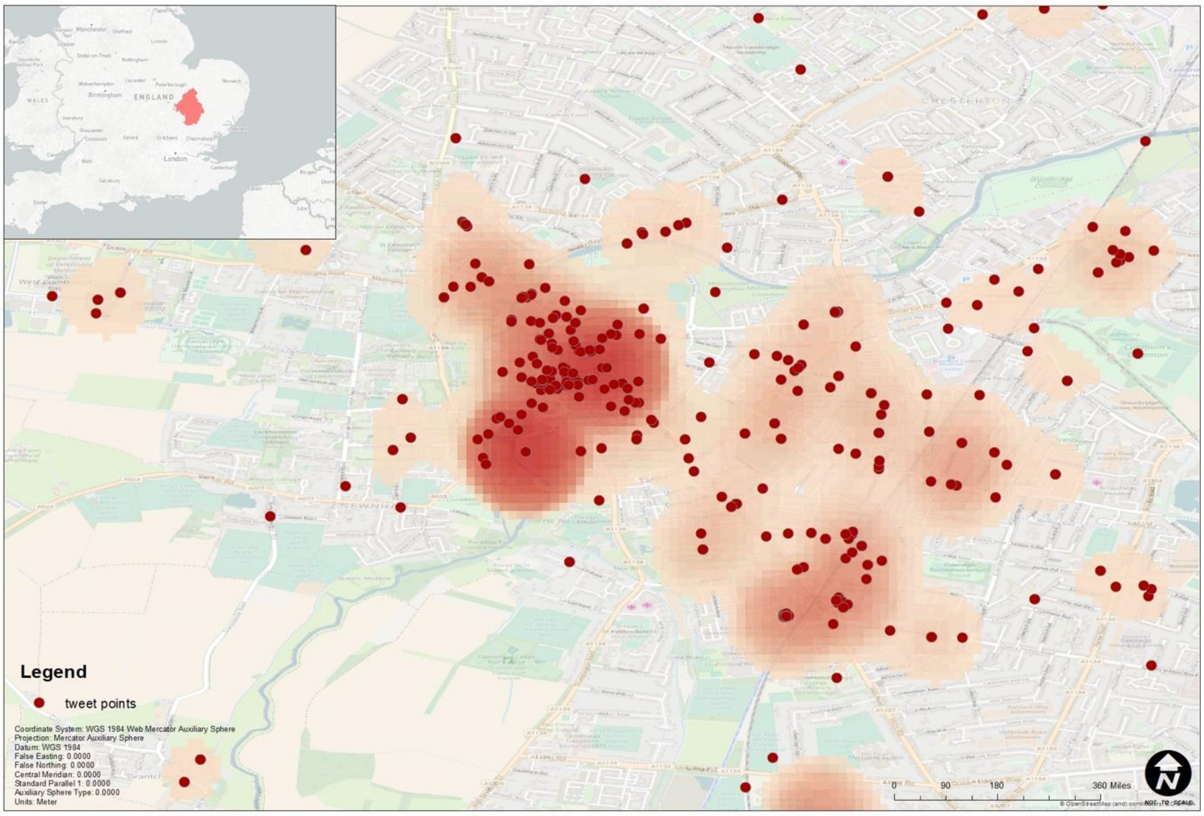

Figure 2 – Social media hotspots for Cambridge

With the second goal complete, it was possible to identify the 5 locations for the questionnaires: 1 King's Parade, 2. Guildhall and Market square, 3. Train Station, 4. Grafton and Mill road, 5.Mesuem of Cambridge. The development of the questionnaires also obeyed a set of rules: we divided the questionnaire into 3 parts, the first part was used for general information (i.e. age group, ethnicity, etc.); the second group of questions related to social media activity; the third objective dealt with socio-economic characteristics, housing affordability and homeless. (Questionnaire attached to this report as Appendix 1).

Table 1 - Cambridge questionnaires and results

Location	Number of questionnaires complete	Observations	Key findings
1. King's Parade	40/50	1.Tourists groups crowed around this area, 2. Collection point for tourists,	1.Pedestrians around King's Parade stay longer on the street, 2. Respondents' usage of social media is high, and they believe the frequent social media activities happen around.

2. Guildhall and Market square	38/50	1. more homeless than other areas, 2. people eat on the bench.	1.More locals crowd in this area, 2.Most people think it is affordable for accommodation in this area, 3.Respondents spend more time in this area.
3. Train Station	30/50	1. people do not cluster together, 2. people waiting outside the station and use their phone a lot	1.People similarly spend 5-20 mins in this area, 2.Most respondents are locals and students, 3. social media activities may not be crowded here.
4. Grafton and Mill road	33/50	1. people always carry bags, 2. the homeless live on the lanes	1.respondents are more locals but their background is diverse, 2.prefer to stay here more than 20mins, 3.no mixed-use function.
5. Museum of Cambridge	20/50	1. sidewalks are crowed 2. busy intersection for pedestrian, cyclist, vehicles.	1. Do not like to stay for long and they just passed by. 2. It is affordable for respondents if they move into this area.

Conclusion

Information and Communication Technologies (ICT), in particular associated with new internet platforms that produce user generated content are becoming a popular source of data to associate to more traditional data sets such as census and other spatial explicit data. In this study, data harvested from tweets was geocoded, allowing to identify hot-spots of activity. The identification of five key hotspots promoted the development a second set of analysis trough the use of questionnaires in order to link quantitative ad quantitative research and refine the results.
The key findings for both case studies: (1) High concentration in five key areas are identified, but the area in Grafton and Mill road doesn't show a clear cluster; **(2)** Young people prefer to use internet for housing information and easily identify the housing information on social media; **(3)** Among the respondents who use social media, the elders also make up for a higher certain percentage than we expected initially; **(4)** Facebook is the most popular social media software. It may be a good research source in the future studies; **(5)** For people who are already homeowners they are unlikely to follow housing information through the internet or social media; **(6)** Respondents basically think the function of the five observed sites is mix-used type of land use.

Acknowledgements

This project was supported by a mini-projects award from the Centre for Digital Built Britain, under InnovateUK grant number 90066 and report was written in collaboration with the Lab of Interdisciplinary Spatial Analysis (LISA lab) at the Dep. of Land Economy, University of Cambridge. We are grateful to all our founders, who made it possible for us to develop urban studies in order to advance this mission. The views expressed are from the authors only and should not be attributed to the CDBB or its founders. Funders do not determine research findings or the insights and recommendations in this report. This report is the first component of on-going Ph.D. research at the LISA lab, in particular of Mr. Haifeng Niu's research. The research will continue to focus on utilising crowd-sourcing data on urban segregation research in large-scale cities. We are very grateful also to Miss Yiwen Qiu (Ph.D. candidate) for her help with the questionnaires' survey at the five locations in Cambridge and to the team in China that kindly agreed to do the work and waited as long as possible.

References

1. AlSayyad, N., & Guvenc, M. (2015). Virtual Uprisings: On the Interaction of New Social Media, Traditional Media Coverage and Urban Space during the `Arab Spring'. *URBAN STUDIES*, *52*(11, SI), pp. 2018–2034.
2. Bailey, N. (2012). How Spatial Segregation Changes over Time: Sorting Out the Sorting Processes. *Environment and Planning A*, *44*(3), pp. 705–722.
3. Bibri, S. E., & Krogstie, J. (2017). ICT of the new wave of computing for sustainable urban forms: Their big data and context-aware augmented typologies and design concepts. *Sustainable Cities and Society*, *32*(Supplement C), pp. 449–474.
4. Bonzanini, M. (2016). *Mastering Social Media Mining with Python*. Packt Publishing.
5. Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city.
6. Fayoumi, A., Jackson, C., Lewis, C., Straw, J., Sharpe, J., & Nicol, D. (2017). *What They Are Tweeting About Me? Social Media Data Analytics with Geographical Visualisation*.
7. Frias-Martinez, V., & Frias-Martinez, E. (2014). Spectral clustering for sensing urban land use using Twitter activity. *ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE*, *35*, pp. 237–245.
8. Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media, in: *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, p. 6. ACM.
9. Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), pp. 260–271.
10. Hecht, B. J., & Stephens, M. (2014). A Tale of Cities: Urban Biases in Volunteered Geographic Information. *ICWSM*, *14*, pp. 197–205.
11. Huang, H., Gartner, G., & Turdean, T. (2013). SOCIAL MEDIA DATA AS A SOURCE FOR STUDYING PEOPLE'S PERCEPTION AND KNOWLEDGE OF ENVIRONMENTS. *MITTEILUNGEN DER OSTERREICHISCHEN GEOGRAPHISCHEN GESELLSCHAFT*, *155*, pp. 291–302.

12. Huang, Q., & Wong, D. W. S. (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, *30*(9), pp. 1873–1898.
13. Huang, Q., & Xu, C. (2014). A data-driven framework for archiving and exploring social media data. *Annals of GIS*, *20*(4), pp. 265–277.
14. Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., & Newth, D. (2015). Understanding human mobility from Twitter. *PloS one*, *10*(7), p. e0131469.
15. Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: Issues, challenges, tools and Good practices, in: *2013 Sixth International Conference on Contemporary Computing (IC3)*, pp. 404–409.
16. Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40*(2), pp. 61–77.
17. Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place (A. Sánchez, Ed.). *PLoS ONE*, *8*(5), p. e64417.
18. Rashidi, T. H., Abbasi, A., Maghrebi, M., Hasan, S., & Waller, T. S. (2017). Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies*, *75*(Supplement C), pp. 197–211.
19. Shelton, T., Poorthuis, A., & Zook, M. (2015a). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, *142*, pp. 198–211.
20. Steiger, E., Resch, B., & Zipf, A. (2016). Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE*, *30*(9, SI), pp. 1694–1716.
21. Tammaru, T., Musterd, S., van Ham, M., & Marcińczak, S. (2016). *A multi-factor approach to understanding socio-economic segregation in European capital cities*. Taylor & Francis. Retrieved January 4, 2018, from http://www.oapen.org/search?identifier=615512
22. Tsui, K. L., & Zhao, Y. (2017). Discussion of "Analyzing Behavioral Big Data: Methodological, practical, ethical, and moral issues." *Quality Engineering*, *29*(1), pp. 79–83.
23. Wu, L., Zhi, Y., Sui, Z., & Liu, Y. (2014). Intra-Urban Human Mobility and Activity Transition: Evidence from Social Media Check-In Data. *PLOS ONE*, *9*(5).
24. Xu, Z., Liu, Y., Xuan, J., Chen, H., & Mei, L. (2017). Crowdsourcing based social media data analysis of urban emergency events. *MULTIMEDIA TOOLS AND APPLICATIONS*, *76*(9), pp. 11567–11584.
25. Yang, J., Hauff, C., Houben, G.-J., & Bolivar, C. T. (2016). Diversity in Urban Social Media Analytics, in: Bozzon, A and CudreMauroux, P and Pautasso, C (Ed.), *WEB ENGINEERING (ICWE 2016)*, pp. 335–353. Lecture Notes in Computer Science.
26. Zhan, X., Ukkusuri, S. V., & Zhu, F. (2014). Inferring Urban Land Use Using Large-Scale Social Media Check-in Data. *NETWORKS & SPATIAL ECONOMICS*, *14*(3–4), pp. 647–667.
27. Zhou, X., & Zhang, L. (2016). Crowdsourcing functions of the living city from Twitter and Foursquare data. *CARTOGRAPHY AND GEOGRAPHIC INFORMATION SCIENCE*, *43*(5, SI), pp. 393–404.

28. Overview — Twitter Developers. Retrieved December 7, 2017, from https://developer.twitter.com/en/docs/tweets/batch-historical/overview

Appendix 1

Survey

Please take a few minutes to fill out this survey on the building environment around you now. Thank you for your participation.

Site:

☐ **King's parade** ☐ **Guildhall and market square** ☐ **Railway station**
☐ **Grafton and mill road** ☐ **Other**

Part I: General Information

1. Are you male or female?

☐ male ☐ female

2. To which of the following age groups do you belong?

☐ under 17 years old ☐ 18-24 years old ☐ 25-34 years old
☐ 35-44 years old ☐ 45-54 years old ☐ 55-64 years old
☐ 65-74 years old ☐ 75+ years old

3. To which of the following ethnic groups do you belong?

☐ White ☐ Hispanic or Latino ☐ Black or African American
☐ Asian / Pacific Islander ☐ Other

4. what is your resident identity?

☐ Locals ☐ Tourists ☐ Students
☐ University staff ☐ Other ☐ Region – East Anglia

Part II: Questions relate to social media result

1. How would you rate this cluster of social media activity of this area?

Very crowded ☐ ☐ ☐ ☐ ☐ ☐ ☐ Comfortable

2. Which main function will you identify this area?

☐ Commercial ☐ Transportation ☐ Cultural ☐ Education ☐ Business ☐ Residents

3. How much time do you usually spend in this area?

☐ 0 to 5 minutes ☐ 5 to 20 minutes ☐ 20 to 40 minutes Other ____________

4. How would you rate the openness of the buildings and external environment?

Only wealthy ☐ ☐ ☐ ☐ ☐ ☐ ☐ friendly to everyone

(especially for disabled and low-income)

5. Have you ever feel that this area is not designed for you or how would you improve it?

Part III: Economic and Social characteristics

1. Do you live around?

☐ Yes | ☐ No

2. What is your highest level of education?

☐ Elementary school ☐ High school ☐ College ☐ Master's degree ☐ Ph.D

3. Which options below is your current housing situation?

☐ Homeowner ☐ Tenant\College accom ☐ Temporary dwellings ☐ with no home or shelter

4. How would you rate the affordability of yourself if you move to this area?

Affordable ☐ ☐ ☐ ☐ ☐ ☐ ☐ Unaffordable

5. Do you use social media software/website? (Facebook, Twitter, Foursquare, Yelp…)

☐ Yes | ☐ No

if yes, what social media do you use_________________________

6. How do you use social media?

☐ Mobile phone ☐ Computer ☐ Tablet ☐ other

7. Where do you use wireless internet from coffe-shop?

☐Cafe ☐ University ☐ Your own paid for

8. For those with temporary dwellings and no home/shelter:

how do you use internet ____

9. Do you think that access to internet would get more housing information?

☐ Yes | ☐ No

if yes, how_________________________

10. Do you think that access to social media would improve you housing condition?

☐ Yes | ☐ No

if yes, how____________________________