Profiling the Dynamic Pattern of Bike-sharing Stations: a case study of Citi Bike in New York City

Yunzhe Liu*[12], Meixu Chen†[13], Daniel Arribas-bel[1], Alex Singleton[1]

[1] Geographic Data Science Lab, Geography and Planning, University of Liverpool, Liverpool, L69 7ZT

[2] SpaceTimeLab, Civil, Environmental and Geomatic Engineering, University College London, London, WC1E 6BT

3 Department of Geography, University College London, London, WC1E 6BT

February 12, 2021

Summary

This research applies a hierarchical k-means clustering method on the TF-IDF weighted 2019 cycling transactions from the Citi Bike bike-sharing system operating in New York City, with the primary goal of investigating the spatiotemporal usage pattern of its docking points. With a particular focus on bike-sharing stations in Manhattan, we classify 504 stations into four main clusters featuring heterogeneous dynamic usages, including leisure-oriented, residential-oriented, workplace-oriented, and off-peak oriented. We interpret each cluster based on their salient characteristics and anticipate possible future directions of this work.

KEYWORDS: Bike-sharing, Mobility, Public Transit, Urban Dynamics, Spatiotemporal Data Mining

1. Introduction

Bike-sharing system (BSS) is “a short-term bicycle rental service for inner-city transportation providing bikes at unattended stations” (Vogel et al., 2011, 514). Unlike traditional bicycle rental services, BSS is usually designed as part of the urban transit system, with lesser cost, increased flexibility, and easier access (Midgley, 2009). With more implementations of initiatives promoting active travel (i.e., walking and cycling), BSS has gained increasing popularity that over 2000 systems are currently in operation worldwide by 2020 (DeMaio et al., 2020), positively contributing to public transit efficiency, public health and well-being, and environmental and socioeconomic affairs (Public Health England, 2016).

The freely available data and diversity in business models have drawn many researchers’ interest in gaining insights into the BSS. For example, existing studies have focused on statistical patterns of bike-sharing trips, analysing cyclists’ travel behaviour, optimising system operation, and the relationship between multiple variables and the BSS ridership (Kou & Cai, 2019; Noland et al., 2016; O’Brien et al., 2014). However, limited studies have unpacked the dynamic usage patterns of bike-sharing docking stations, which is another significant research subject in the study of urban dynamics and human mobility since the outcomes of such research could be utilised to monitor travel demand, inferring functional characteristics and eventually maintain a sustainable BSS (Zhou, 2015).

This study’s primary objective is to profile the dynamic pattern of bike stations based on their usage within the context of a Citi Bike system dataset collected for the case study area, i.e., New York City. The cycling trip transactions are processed into hourly ingress and egress

*psyliu7@liverpool.ac.uk
†meixu@liverpool.ac.uk

frequency by stations, and a TF-IDF weighted hierarchical clustering is utilised to unveil their spatiotemporal patterns. This research has the potentiality of informing urban planners or decision-makers to identify the primary usage of each docking station, which helps them to examine the current performance of BSS operation and hence improve their services.

2. Study Area and Data Description

New York City (NYC), serving as the whole world's finance capital (Yeandle, 2015), is selected as the case study area, which is characterised by the densest population, the most compact urban land use, and the busiest public transit system in the US (US Census Bureau, 2019). The Citi Bike system operating in NYC is the largest privately-owned 24/7 BSS scheme in North America since 2013, which has possessed over 700 docking stations and more than 12000 bikes, with further expansion underway NYCDoT, 2017).

We extracted 2019 trip histories from the Citi Bike's open data repository[4]. The general data structure is presented in Table 1, where each row represents a finished cycling journey with origin and destination docking stations of one user. As the system continues to expand, several docking stations were only built and commissioned in the second half of 2019 (DiBarba, 2020). For data integrity concern, only stations that were utterly operational before 2019 are considered in this research. Additionally, since those early existed docking stations are mainly located in Manhattan, we only selected stations located in Manhattan to conduct the following-up analysis. Figure 1 displays the spatial distribution of Citi Bike's docking stations in NYC and the aggregated inter-station origin-destination (OD) flows in the Manhattan area. After data cleaning, about 86% (17,650,069 out of 20,551,697) bike trip histories were retained.

Table 1 Examples of Citi Bike data in Manhattan, NYC (after data pre-processing)

Trip ID	**Start Station ID (with location)**	**Start Time (Date & Time)**	**End Station ID (with location)**	**End Time (Date & Time)**[5]	**User Type**[6]
1	3160	2019-01-01 00:01:47	3283	2019-01-01 00:07:07	Subscriber
2	519	2019-01-01 00:04:73	518	2019-01-01 00:10:01	Subscriber
…	…	…	…	…	…
17650069	437	2019-12-31 23:54:55	344	2019-12-31 23:58:09	Subscriber

[4] The dataset is available here: https://www.citibikenyc.com/system-data
[5] End Time >= 2019-12-31 23:59:59
[6] Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member; Only Subscribers were retained in this study due to the regularity

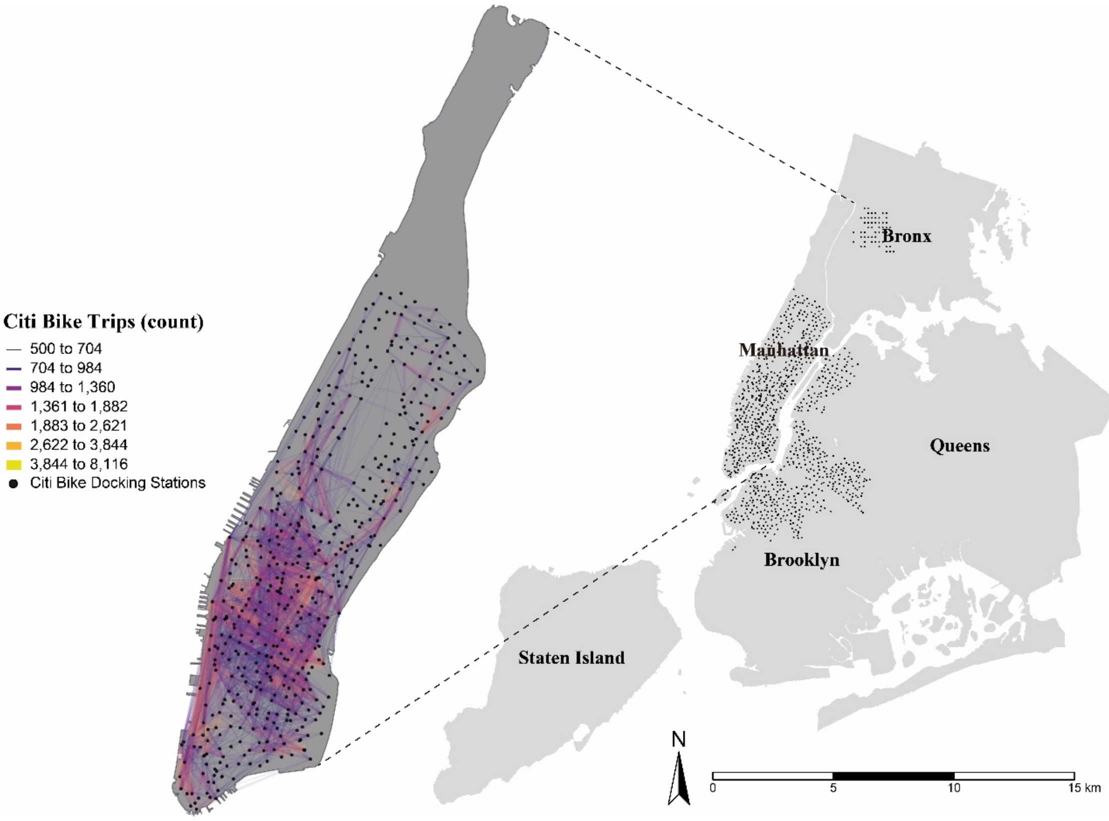

Figure 1 Geographic distribution of Citi Bike docking stations in NYC and OD flows between stations in Manhattan (flow > 500)

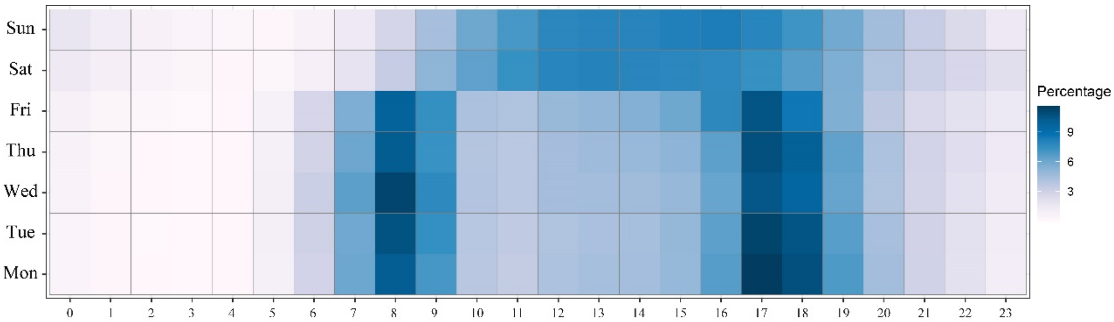

Figure 2 'Weekly travel profile': temporal distribution of the Citi Bike trips in Manhattan, 2019.

3. Methodology

To profile the docking stations' temporal usage pattern, we utilise the ingress and egress information to formulate the 'weekly travel profile' (Figure 2) for all 504 stations. The trip data was aggregated into twenty-four-hour time bands by days of the week, formulating 336 temporal variables, meaning that each station contains 168 variables (24 hours multiply seven days) representing start/egress count, and another 168 variables for end/ingress count. The figure observes two major peaks during weekdays and random diffusion trips during weekends.

Term Frequency-inverse document frequency (TF-IDF), one of the commonly used weighting schemes in text mining, was implemented to weight the egress and ingress frequency assembled in each station, assisting follow-up clustering analysis in providing more distinctive and robust results. Initially, in the text mining field, TF-IDF is to weight 'words' over 'sentences' formulating a 'document'. TF-IDF decreases the importance of 'words' if they appear everywhere in the whole ‘document’, while increases the magnitude of those that only have a high frequency at particular ‘sentences’ (Hu et al., 2015; Leskovec et al., 2020). Inspired by this mechanism, we implemented TF-IDF on our dataset to weight the ‘word’ (i.e., a specific temporal interval) over ‘sentence’ (i.e., 336 temporal intervals), assembling a ‘document’ (i.e., a single station). The analogy is presented in **Eq.1**.

$$W_{ij} = tf_{ij} \times \log \frac{N}{df_i} \qquad \textbf{(Eq.1)}$$

Where W_{ij} is the weight of a temporal interval T_j in Citi Bike docking station S_i; tf_{ij} is the frequency count of T_j among all temporal intervals in S_i; N is the total number of Citi Bike docking stations in the study area, and df_i is the number of stations that contain the temporal interval T_j.

Consequently, higher weights will be assigned to a specific period in stations experiencing a high volume of cycling flows, which can be rarely found elsewhere.

The distinction between station characteristics was assessed by creating a distance matrix based on the cosine similarity of the TF-IDF scores. The hierarchical k-means (H-K-means) clustering algorithm is a hybrid of hierarchical clustering and k-means clustering (Arai & Ridho Barakbah, 2007), was subsequently implemented to classify bike stations into clusters based on the underlying similarities in their dynamic pattern. The optimal cut-off point for the number of clusters was identified as k=4 by Gap Statistics method introduced by (Tibshirani et al., 2001)

4. Results and future work

A series of heatmaps presented in Figure 3 display four generated temporal clusters from 504 stations. A block with a light colour indicates a low probability of appearance of the temporal interval, while the darker colour indicates a higher probability. Additionally, the geographic distribution of the four generated docking station clusters is mapped in Figure 4.

Stations categorised in Cluster 1 are more likely to be leisure-oriented. They witness high flows at both inbound and outbound usage during the non-working time (19:00 to 00:00 at weekday night and 10:00 to 0:00 at the weekend), but low appearance during working hours. The overall docking stations located across Manhattan, while the majority located in Lower Manhattan (Downtown), featured as the home to some of the city’s most prominent buildings and tourist attractions in Manhattan, indicating that these stations are more likely for random entertainment usage.

Stations in Cluster 2 predominately witness high outbound flows during the weekday morning and high inbound flows during the evening peak times. Such pattern indicates a typical residential-oriented functionality. The insights have been further confirmed by examining their spatial distribution: stations are primarily aggregated at Upper West Side and Upper East Side Manhattan, which are known as residential areas.

Stations classified in Cluster 3 shows a reverse pattern compared to those in Cluster 2, implying a typical workplace-oriented usage. The docking stations primarily located in the Midtown East and Downtown business centres, usually featured by many commercial centres, offices and skyscrapers.

The dynamic pattern exhibited by docking stations from Cluster 4 is similar to Cluster 2, thus, these stations might also serve as residential-oriented usage. However, they witness a relatively high flow volume at one or a few temporal intervals before the conventional peak times, implying a preference of off-peak travel.

Based on the spatiotemporal patterns, one of the future directions of this study could be extended by in-depth analysing into the neighbourhood around these bike-sharing stations, providing a detailed urban contextual analysis (Liu et al., 2020; Liu et al., 2021). For example, by looking into the socioeconomic, demographic, and land-use characteristics of these proximate neighbourhoods to further evaluate and characterise the identified docking station clusters.

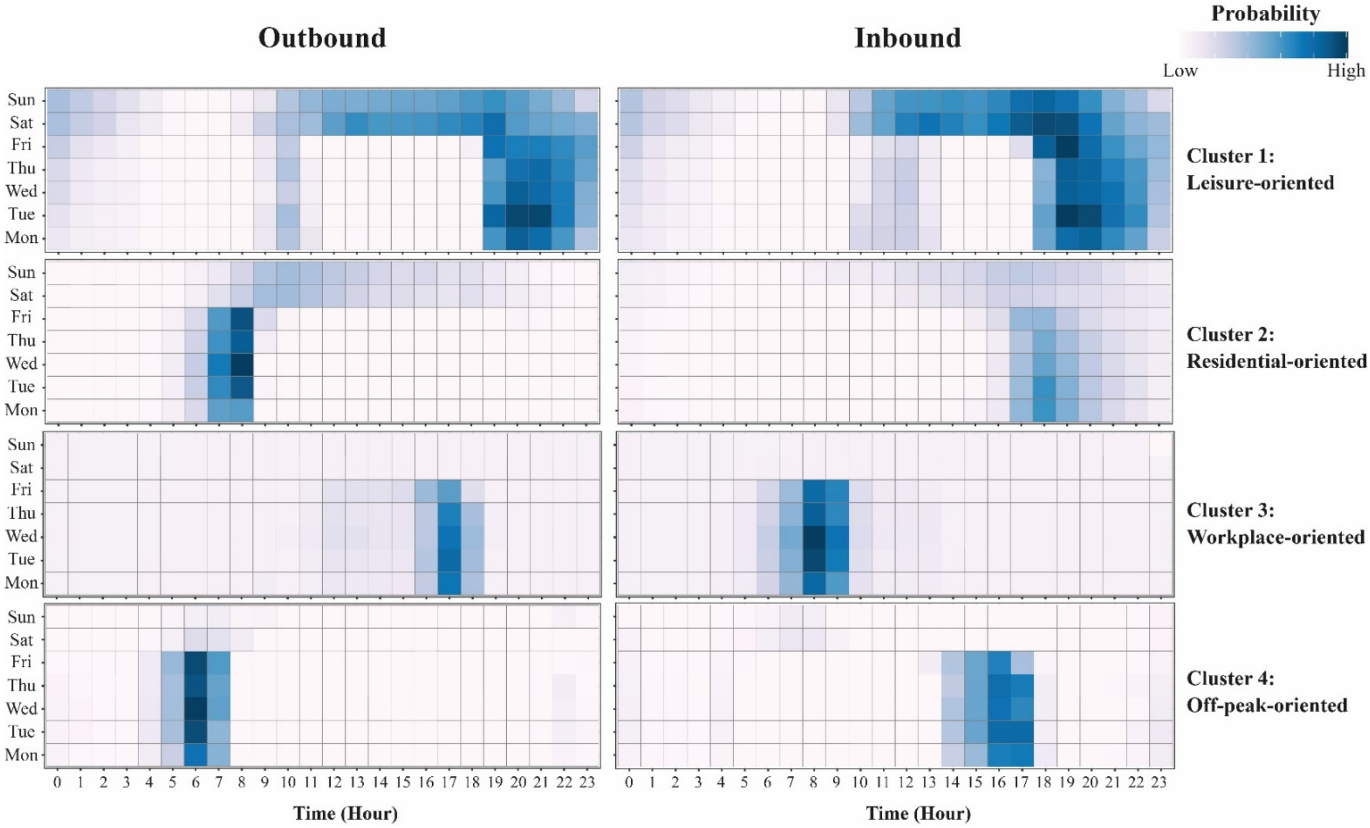

Figure 3 H-K-mean clustering results of four clusters of Citi Bike stations; named by their salient characteristics

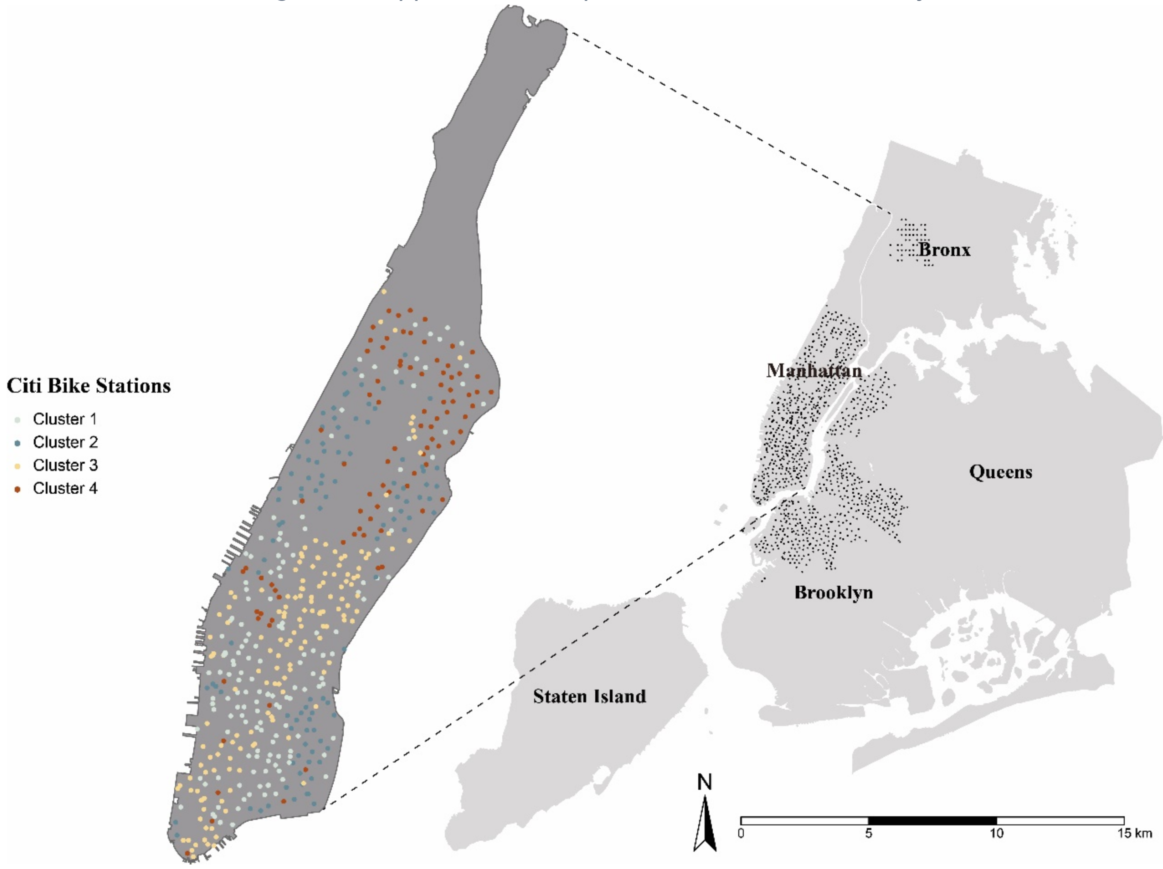

Figure 4 Spatial distribution of the four clusters in Manhattan

5. Acknowledgements

This research has no funding.

References:

Arai, K., & Ridho Barakbah, A. (2007). Hierarchical K-means: an algorithm for centroids initialization for K-means. *Rep. Fac. Sci. Engrg. Reports of the Faculty of Science and Engineering*, *36*(1), 25–31.

DiBarba, A. (2020). *Major Citi Bike Expansion Map Revealed!* https://www.citibikenyc.com/blog/major-citi-bike-expansion-map-revealed

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/j.compenvurbsys.2015.09.001

Kou, Z., & Cai, H. (2019). Understanding bike sharing travel patterns: An analysis of trip data from eight cities. *Physica A: Statistical Mechanics and Its Applications*. https://doi.org/10.1016/j.physa.2018.09.123

Leskovec, J., Rajaraman, A., & Ullman, J. (2020). Data Mining. In *Mining of Massive Datasets* (Third). Cambridge University Press.

Liu, Y., Singleton, A., & Arribas-Bel, D. (2020). Considering context and dynamics: A classification of transit-orientated development for New York City. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2020.102711

Liu, Y., Singleton, A., Arribas-bel, D., & Chen, M. (2021). Identifying and understanding road-constrained areas of interest (AOIs) through spatiotemporal taxi GPS data: A case study in New York City. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/j.compenvurbsys.2020.101592

Midgley, P. (2009). The role of smart bike-sharing systems in urban mobility. *Journeys*.

New York City Department of Transportation. (2017). *Safer cycling. Bicycle Ridership and Safety in New York*. *132*(6), 265. https://search.proquest.com/docview/1313182708?accountid=136944

Noland, R. B., Smart, M. J., & Guo, Z. (2016). Bikeshare trip generation in New York City. *Transportation Research Part A: Policy and Practice*, *94*, 164–181. https://doi.org/10.1016/j.tra.2016.08.030

O'Brien, O., Cheshire, J., & Batty, M. (2014). Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography*, *34*, 262–273. https://doi.org/10.1016/j.jtrangeo.2013.06.007

Public Health England. (2016). *Working Together to Promote Active Travel A briefing for local authorities*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/523460/Working_Together_to_Promote_Active_Travel_A_briefing_for_local_authorities.pdf

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*. https://doi.org/10.1111/1467-9868.00293

United States Census Bureau. (2019). *QuickFacts*. https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork/PST120219

Vogel, P., Greiser, T., & Mattfeld, D. C. (2011). Understanding bike-sharing systems using Data Mining: Exploring activity patterns. *Procedia - Social and Behavioral Sciences*, *20*, 514–523. https://doi.org/10.1016/j.sbspro.2011.08.058

Yeandle, M. (2015). *The Global Financial Centres Index 17*. *March*, 56. http://www.longfinance.net/images/GFCI15_15March2014.pdf

Zhou, X. (2015). Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in Chicago. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0137922

Biographies

Yunzhe Liu is a postdoctoral associate at SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London. He is also a PhD student in the Geographic Data Science Lab, University of Liverpool. His research is on geodemographics, urban analytics, geographic data science, human mobility, and spatiotemporal data mining.

Meixu Chen is a research fellow at Department of Geography, University College London. She is also a PhD student in the Geographic Data Science Lab, University of Liverpool. Her research is on urban analytics, geographic data science, social media, and social mobility.

Alex Singleton is a Professor of Geographic Information Science at the University of Liverpool, Deputy Director of the ESRC Consumer Data Research Centre (CDRC) and Director of the ESRC Data Analytics & Society CDT. His research is on geodemographics, geographic data science, and urban analytics.

Dani Arribas-Bel is a Senior Lecturer in Geographic Data Science at the Department of Geography and Planning, University of Liverpool. He is also an ESRC Fellow at the Alan Turing Institute. His research is on open science, geographic data science, urban economics.