Spatio-temporal variability in Wikipedia: The case of Greater London

Shahreen Muntaha Nawfee
University of Leicester, United Kingdom
smn18@leicester.ac.uk

Andrea Ballatore
Department of Digital Humanities, King's College London
andrea.ballatore@kcl.ac.uk

Stefano De Sabbata
University of Leicester, United Kingdom
s.desabbata@leicester.ac.uk

Dr Nicholas Tate
University of Leicester, United Kingdom
njt9@leicester.ac.uk

Abstract

Spatial user-generated content (UGC) is increasingly being used to study a variety of geographical phenomena, including urban change in social and economic dimensions. Wikipedia content evolves over time and includes articles about geographical areas, points of interest, and geo-located events. In this article, we explore the spatio-temporal variability of geo-located Wikipedia pages, considering their complete editing history. Selecting Greater London as a case study, we study the association between Wikipedia activity and the socio-demographic characteristics of the spatial context. Editing activity grows rapidly at first, and is then followed by a slowdown, reaching a stable rate, with occasional spikes. The initial growth is distributed throughout the study area, but activity becomes gradually more concentrated in central areas. The socio-demographic variability is strongly related to the presence of Wikipedia pages, but only partially to the editing. This approach may support the detection and characterisation of socio-economic change at the urban scale.

Acknowledgements The authors would like to thank the Wikipedia contributors, the R community for the spatial analysis resources and the College of Science and Engineering, University of Leicester for the PhD studentship of the first author.

1 Introduction

Over the last two decades, there has been a substantial rise in the use of spatial user-generated content (UGC) in geographical research. UGC's ability to provide more granular and timely data than traditional sources in some aspects brings value to many contexts and applications. For instance, different UGC sources are increasingly used to understand urban dynamics, neighbourhood changes, urban land use classification and gentrification [6]. Data sources range from Flickr images [3] to Airbnb reviews [6] and social media, such as tweets [9].

Wikipedia is one of the largest crowdsourcing platforms, including a large variety of geospatial information, and one of the most frequently accessed websites worldwide. It acts as a collective knowledge-building platform that can influence our understanding of place

[4]. The spatial distribution of Wikipedia content is uneven. At a global scale, there are more articles about the Global North than the Global South [4] and at the urban scale, it over-represents urban centers [2, 8]. Urban areas have higher quality content produced more by local contributor compared to its rural counterparts [7]. Non-geographic biases are also present, including self-focus bias, topical coverage bias [5], and biases in linguistic representation [8]. In this paper, we aim to explore the spatio-temporal variability of Wikipedia pages and their editing, considering Greater London as a case study. Focusing on geo-located pages, we address two questions: (RQ1) what are the temporal trends in the variation of Wikipedia editing activity? (RQ2) How is the activity spatially distributed?

The variability in Wikipedia is influenced by the socio-demographic context. Socio-demographic characteristics such as workday population, age, education, and presence of dependent children, as well as house prices, have been found to correlate to 44% of the spatial variability of Wikipedia content in London [1]. However, levels of education, income, population, and ethnicity explain only about 18% of variability in Los Angeles [2]. The number of Wikipedia pages correlates positively to the presence of certain Points of Interest [2]. The global variation in the number of Wikipedia pages has been linked to access to the internet connection, the number of edits, and population size [4]. However, the relationship between editing activity and socio-demographic context has not been studied in sufficient detail, we aim to answer the question, (RQ3) how the occurrence of Wikipedia pages and edits are associated to the underlying socio-demographic context?

2 Methodology

2.1 Data collection

To study the spatio-temporal variability, we collected geotagged Wikipedia data from the Wikimedia data dump.[1] This includes pages of every article in Wikipedia that has a primary geotag, thus identifying it as an article about an entity or event with geographic scope. From the original dataset, we extracted English Wikipedia articles with a primary geotag within the boundaries of the Greater London region (United Kingdom) as a dataset for our case study. This gave us 3,607 articles. Using the page ID of each article, we queried the extensive editing history of each page available in the ArticleInfo Xtools/Page History tool,[2] extracting the total number of edits – classified by type of contributor, size, and year.

2.2 Spatio-temporal analysis of Wikipedia editing

To assess the spatio-temporal variation in Wikipedia page distribution, we first analyzed the edit history (RQ1). Our collection of editing activity data shows the number of edits in each year (divided into major, minor, and IP edits) as well as the total size of the edits for each page. We aggregated the page editing history to show the total number and size of edits for each year. We then visualized both mean and sum of edits per year by types of edit to identify patterns, and we compared editing to the overall trend for English Wikipedia.

To find out the Spatio-temporal distribution of changes in edits (RQ2), we first identified patterns in the temporal dimension of editing. We segmented the time period into sections based on the variation in editing activity, and then looked for the spatial distribution of

[1] Data dumps - Meta `https://wikimedia.org`, Accessed on January 2021

[2] E.g.,`https://xtools.wmflabs.org/articleinfo/en.wikipedia.org/London`, Accessed on January 2021

the Wikipedia pages during each time period. This involved aggregating number of pages that were edited at the start and end of a time period using a $1km \times 1km$ regular grid, then calculating the differences between them.

Based on the existing literature discussed above, the underlying assumption was that areas with more population of working age, economically well-off having English as the first language are represented more through the number of Wikipedia pages (RQ3). We also wanted to explore the hypothesis that the house prices, transport accessibility and indicators of architectural heritage can provide an insight into the number of Wikipedia pages in an area (RQ3). As such, we explored those relationships through a series of linear regression models (using the ordinary least squares approach) using the number of Wikipedia pages and edits, normalised based on size of the areas in square kilometer as well as number of inhabitants as the dependent variable. We identified demographic (population count and density, as well as age), economic (employment rate, median household income, household and car ownership, house prices) and social (number of immigrant workers, non-English speaking households, public transport accessibility) factors, as well as an indicator of the presence of heritage architecture (number of listed buildings) as our independent variables. In the next section, we present our two best performing models.

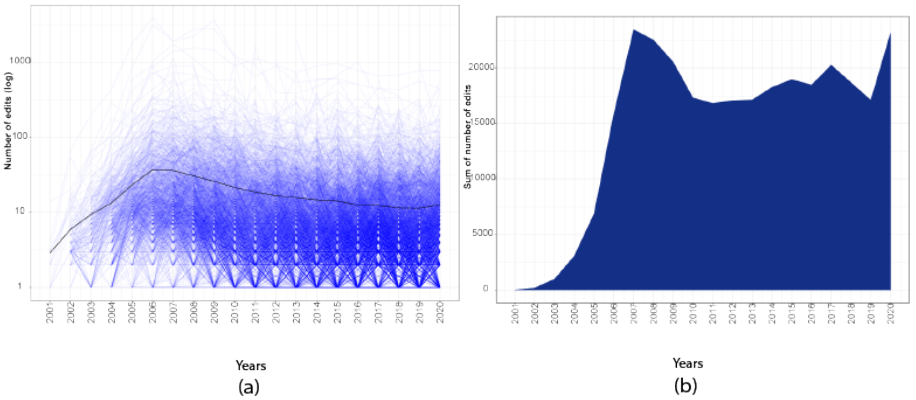
3 Results

3.1 The yearly changes in Wikipedia editing

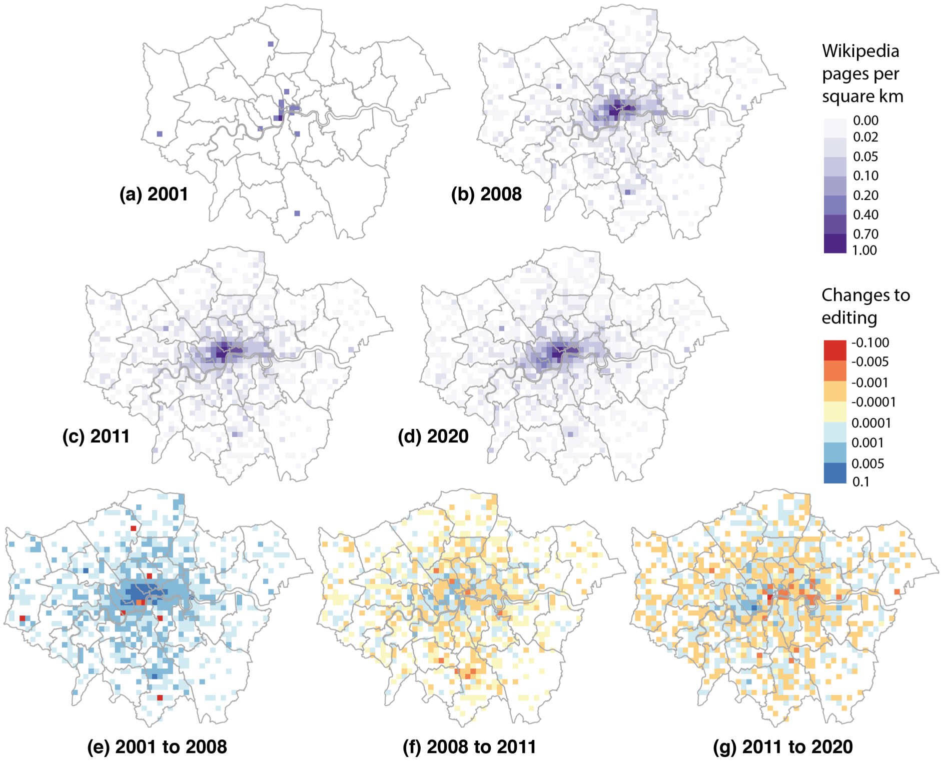
Figure 1a shows changes in the number of edits over time (RQ1). The average number of edits increases steeply from 2001 to 2006, then falling gradually, except for the occasional high peaks in 2015, 2017, and 2020. Also noticeable in the figure are the higher number of outliers from 2005 onwards, although editing decreased after 2006 it became concentrated on few pages. This led us to the question of which pages they are and where are they located. In figure 1(b) the variation in the sum of the number of edits from year to year is noticeable. The sum of edits rises steeply until 2008, then falling to a more constant level, with occasional spikes corresponding to those seen in figure 1. The growth rate of editing for Wikipedia pages is extremely fast in the first two years, with the rate falling until 2010, then stabilizing with occasional high peaks in 2015, 2017, and 2020, as already seen for the other series.

We further further explores the trend by differentiating by type of edit. The contribution is mostly in the form of major edits, which follow the same trend as the overall count. The amount of minor edits also show a similar types pattern. The occasional peaks are mostly constituted of major edits related to the addition of new pages (especially related to events, e.g. ‘Parsons green train bombing in 2017’) or significant changes to the existing ones (e.g., ‘Big Ben’ was one of the most active pages in 2017, when renovation works started). We also established that the trend of edits we see for English Wikipedia in Greater London is in line with the trend for overall English Wikipedia. As such, we can identify three periods:

1. From 2001 to 2008, when the mean number of edits, as well as the sum of all edits, increased steeply, reaching its highest value, despite a noticeable fall in growth rate after the year 2003/2004. The increase was high among all types of edits.
2. From 2008 to 2011, when a gradual fall in the number of edits and the sum of edits size is visible. All types of edits fell, but a large number of outliers was also noticeable.
3. From 2011 to 2020, when a stable editing activity is visible and the number of pages edited and the overall edits remain stable. However, occasional peaks in both the sum of edits and the average number of edits were noticeable in 2015, 2017 and 2020.



■ **Figure 1** Number of edits per Wikipedia page in Greater London per year (a), their average – black line in (a) and their overall total (b).



■ **Figure 2** Spatial distribution of Wikipedia pages by editing(a-d) and changes between them(e-g), rise in blue and fall in red. Contains National Statistics data © Crown copyright and database right [2015] and Contains Ordnance Survey data © Crown copyright and database right [2015].

3.2 Spatial distribution of Wikipedia pages

The spatial distribution of Wikipedia pages (RQ2) per square kilometer in 2020, aggregated to ward level, shows a higher concentration in the central part of Greater London, covering mostly the City of London, Westminster and Royal Borough of Kensington and Chelsea. We have also explored where the changes occurred in the time period we have identified in the section before. Figure 2 illustrates where the growth or decline in editing activity has taken place when Wikipedia pages experienced maximum or minimum editing.

1. From 2001 to 2008 the growth of Wikipedia is spread throughout the study area, with the highest growth around central London (Figure 2e). As figure 2a shows, the editing activity was concentrated at very few locations especially in Central London in 2001. In 2008 (the year with the highest average and the total number of edits) the editing took place in a wider range of areas (Figure 2b).
2. From 2008 to 2011, the editing activity shrank largely in all areas, in this period, with few exceptions to areas west of the City of London. (Figure 2f).
3. From 2011 to 2020, when the editing activity remains quite stable, the fall in editing number is visible throughout Greater London, except in some scattered areas and the western side of Central London (Figure 2g), covering the boroughs of Westminster, Kensington and Chelsea, and Hammersmith and Fulham.

In addition, we looked into which pages have been most active, to understand what was changing at each phase of Wikipedia growth. It shows that some pages are repeatedly active (Heathrow Airport, British Museum, Chelsea F.C.). Our analysis also shows the yearly variation result from the occurrence of specific events in a year (e.g. Olympic in 2012).

3.3 Association between socioeconomic variability and Wikipedia pages and edits

To explore how the presence of Wikipedia pages and edits are influenced by the underlying socio-demographic context (RQ3) we created a regression model using the number of pages per square kilometer as the dependent variable. However, not all the assumptions of the regression model were met, as the residuals were positively auto-correlated and had a skewed distribution. As such, we decided to conduct further analysis using a Moran'I test, which identified spatial autocorrelation (0.155 at a p-value 6.473e-12) among the residuals. Spatial dependence is also proven by the significant Lagrange Multiplier LM (lag) and LM(error) tests. The Robust LM (lag and error) tests are then conducted to understand the type of spatial dependence, both of which are significant for our model. Since Robust LM (error) has a lower significance value we ran the Spatial Error model (SEM). The spatial error variable is statistically significant and has a positive effect (lamda=0.40299), the model fit has thus improved, as indicated by log-likelihood and a smaller AIC value (1147.4 for SEM compared to 1187.2 for OLS model), when the spatial dependence of error is considered. The Nagelkerke pseudo-R^2 value indicates almost 69% of the variability can be accounted for by the independent variables. Our second regression model was run on Wikipedia edits per 1000 inhabitants as the dependent variable. Similarly, the assumptions of regression were not met. Like our first model, we check for spatial autocorrelation and dependence. The Moran's I test return a small, positive (0.041) but significant value (p-value of 0.03). The spatial dependence test indicates the presence of spatial lag dependence (significant value for the LM lag test) but not a spatial error. We thus ran a Spatial Lag model (SLM), which increased the model fit, as indicated by a positive, significant rho value (0.141 at a p-value of 0.004) and lower AIC (3136.4 for SLM against 3142.6 for OLS). However, only around 21% of the variability in editing can be accounted for by the independent variables.

4 Discussion and Conclusion

In this study, we showed how, in Greater London, Wikipedia content was created quickly in the early stages, reaching a peak in the first couple of years, before falling and stabilizing (RQ1). The nature of change through time seems to be specific to this platform, as English Wikipedia follows the same trend. We also found that the initial growth is spread throughout the study area, but activity gradually became more concentrated in some inner boroughs (RQ2). Such variability is likely to be dependent on what features of an urban area are represented, and where they are more likely to be located. As most activity is concentrated on renowned landmarks, heritage sites, tourist attractions and remarkable events, areas where more of those occur are more likely to be over-represented. In our future work, we aim to explore how different points of interest are represented in Wikipedia, to better understand what is represented and how it changes with time. Finally, we were able to show that socio-demographic context of an area can account for a large amount (almost 69%) of the variability in the presence of Wikipedia, when accounting for the spatial errors (RQ3).

The spatio-temporal distribution of Wikipedia editing is clearly related to the dynamics of Greater London, where areas such as Central London seem more likely to experience change (in Wikipedia and on the ground) than the rest. In our future research, we aim to explore which kind of changes in an urban area are reflected through a UGC source like Wikipedia and to what extent, as well as whether the association between editing and socio-demographic variability might lead to a potential of use of the Wikipedia platform in the study of urban change.

References

1 Andrea Ballatore and Stefano De Sabbata. Charting the geographies of crowdsourced information in greater london. In *The Annual International Conference on Geographic Information Science*, pages 149–168. Springer, 2018.

2 Andrea Ballatore and Stefano De Sabbata. Los angeles as a digital place: The geographies of user-generated content. *Transactions in GIS*, 24(4):880–902, 2020.

3 Meixu Chen, Dani Arribas-Bel, and Alex Singleton. Understanding the dynamics of urban areas of interest through volunteered geographic information. *Journal of Geographical Systems*, 21(1):89–109, 2019.

4 Mark Graham, Bernie Hogan, Ralph K. Straumann, and Ahmed Medhat. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*, 104(4):746–764, 2014.

5 Brent Hecht and Darren Gergle. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on communities and technologies*, pages 11–20, 2009.

6 Shomik Jain, Davide Proserpio, Giovanni Quattrone, and Daniele Quercia. Nowcasting gentrification using airbnb data. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

7 Isaac L. Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. *Not at Home on the Range: Peer Production and the Urban/Rural Divide*, page 13–25. Association for Computing Machinery, New York, NY, USA, 2016.

8 Cailean Osborne, Mark Graham, and Martin Dittus. Edit wars in a contested digital city: Mapping wikipedia's uneven augmentations of berlin. *The Professional Geographer*, 73(1):85–95, 2021.

9 Ate Poorthuis, Taylor Shelton, and Matthew Zook. Changing neighborhoods, shifting connections: mapping relational geographies of gentrification using social media data. *Urban Geography*, 0(0):1–24, 2021.