GEOPARSING COMMENTS FROM REDDIT TO EXTRACT MENTAL PLACE CONNECTIVITY WITHIN THE UNITED KINGDOM

A PREPRINT

Cillian Berragan
University of Liverpool
c.berragan@liverpool.ac.uk

Alex Singleton
University of Liverpool
alex.singleton@liverpool.ac.uk

Alessia Calafiore
University of Edinburgh
acalafio@ed.ac.uk

Jeremy Morley
Ordnance Survey
Jeremy.Morley@os.uk

2022-09-09

ABSTRACT

Place connectivity is explored between geographic locations extracted from comments on Reddit. Unlike formally structured geographic data, this corpus of unstructured text provides connections derived from co-occurring locations, capturing subconscious links between them, alongside inherent biases. Our work demonstrates the ability to link locations mentioned by unique users, building 'mental' place connections for over 50,000 unique locations in the United Kingdom. Sentiment regarding locations is compared against their levels of connectivity, demonstrating that user opinions regarding locations are likely drivers in mental place connectivity.

Keywords social media • natural language processing • social interaction

1 Introduction

Connectivity between places may be explored through the physical movement of individuals, using population movement data like transport records (Yang, Li, and Li 2019; Allard and Moura 2016; Gong et al. 2021; Farber and Li 2013), or GPS information through mobile phone data (Lin, Wu, and Li 2019; SafeGraph 2022). Due to the advent of 'Volunteered Geographic Information' (VGI) (Goodchild 2007), these connections may also be explored through geotagged social media posts (Arthur and Williams 2019; Ostermann et al. 2015; Li et al. 2021), with results that mirror true population movements [li2021;Kuchler, Russel, and Stroebel (2020)].

Literature discussing the role of human cognition in constructing mental images of cities (**lynch1964?**), and how they can be represented through mental maps (Gould and White 1986), shows that the way humans conceive spatial structures and place relationships are substantially entrenched in individuals' experiences and geographic knowledge, which only partially derive from movements. In particular, while movements are constrained by time and euclidean distance in geographic space (Miller 2018; Patterson and Farber 2015), representational spaces expressed in mental maps do not necessarily correlate with these spatio-temporal boundaries.

New forms of data, especially social media and text data, offer novel opportunities to explore place connections, emerging from peoples naïve and experiential geographic knowledge. Recent studies have investigated how digital social networks, i.e. Facebook friendships (Bailey et al. 2018), may be explored in this regard. While, geo-semantic relatedness outlines the ability to quantify relationships between geographic terms in text (Ballatore, Bertolotto, and Wilson 2014), with work applying this to co-occurring city names found in news articles, social media, and general web pages (Hu, Ye, and Shaw 2017; Ye, Gong, and Li 2021; Liu et al. 2014; Meijers and Peris 2019). Our paper considers

the use of a novel corpus of text data from comments on the online social media website Reddit as a source of VGI. A task-specific geoparsing pipeline is first used to identify place names4 related to the United Kingdom and resolve them to geographic coordinates (Purves et al. 2018), at a geographic resolution higher than is typically explored through geoparsing. We derive connectivity between each location in our corpus based on the number of times two distinct locations co-occur, normalised by the total number of users that mention each location. The context for co-occurrences is derived from the total collection of comments submitted by each unique user, meaning every location mentioned by a single user is treated as co-occurring, and exhibiting some implicit connectivity. A full interactive map showing place connections is available through Unfolded.ai.

With the connectivity between places established, we consider the ability to derive explainable characteristics from the text, to determine why different levels of place connectivity occur. While traditional connections between places may be influenced by factors like transport availability (Allard and Moura 2016), the alternative mental place connections derived through our paper may be more heavily influenced by a subconscious bias. Connectivity is therefore examined against sentiment, expected to highlight these biases, alongside a standard measure of relative material deprivation (limited to England), through the Indices of Multiple Deprivation (IMD), an alternative measure that would be expected to affect traditional place connectivity.

2 Methodology

2.1 Data sources

Reddit is a public discussion, news aggregation social network, among the top 20 most visited websites in the United Kingdom. As of 2020, Reddit had around 430 million active monthly users, comparable to the number of Twitter users (Murphy 2019; Statista 2022). Reddit is divided into separate independent subreddits each with specific topics of discussion, where users may submit posts which each have dedicated nested conversation threads that users can add comments to. In total there are 213 subreddits that relate to 'places' within the United Kingdom[1]. For each subreddit, every single historic comment was retrieved using the Pushshift Reddit archive (Baumgartner et al. 2020). In total 8,295,591 comments were extracted, submitted by 492,123 unique users, between 2011-01-01 and 2022-04-17.

To train a model to identify place names from comments, the WNUT-17 corpus was used (Derczynski et al. 2017), keeping only 'location' labels. In total this corpus covers 5,690 individual documents from Reddit, Twitter, YouTube, and StackExchange. Two gazetteers were selected to geocode place names, chosen to be UK centric, and at a high resolution. We were specifically interested in a gazetteer that did not include country names external to the UK, but included fine-grained named locations like street names. For our gazetteer we combined OS Open Names, and non-settlements from the Gazetteer of British Place Names.

2.2 Geoparsing

A custom named entity recognition (NER) model was built using a RoBERTa based transformer language model, pre-trained using Twitter data[2], as this architecture has given good results on the WNUT-17 corpus (Barbieri et al. 2020).

With place names identified, we developed a method for attributing each name to a single set of geographic coordinates. Place names typically appear multiple times in gazetteers, especially when grounding fine-grained locations like street names, meaning a disambiguation method is required. We disambiguate place names by finding their minimum distance to a collection of contextual locations. Contextual locations in this case refer to all gazetteer entries matching place names that appear in sentences with this target place name, in the same subreddit. This works under the assumption that each unique place name in a single subreddit is likely to refer to the same location, and that locations mentioned in surrounding text are likely close together (Kamalloo and Rafiei 2018).

Sentiment was attributed to each sentence in our corpus containing a place name, using an existing fine-tuned sentiment classification transformer model[3]. Each place name identified by our NER model was assigned sentiment based on its context sentence.

2.3 Measuring place connectivity

Our place connectivity methodology considers the co-occurrence between each place mentioned by every unique user in our corpus. The following equation described by (Li et al. 2021) outlines this concept:

[1] https://reddit.com/r/unitedkingdom/wiki/british_subreddits
[2] https://huggingface.co/cardiffnlp/twitter-roberta-base
[3] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

$$PCI_{ij} = \frac{\mathbf{S}_{ij}}{\sqrt{\mathbf{S}_i \mathbf{S}_j}}, i, j \in [1, N]$$

PCI_{ij} is the place connectivity index between locations i and j, $\mathbf{S}_{ij}$ is the total number of users that mention both locations (i.e. the intersect in set theory). This is normalised, given locations with higher populations are expected to be mentioned by a larger number of users, using $\sqrt{\mathbf{S}_i \mathbf{S}_j}$, the total number of users mentioning $\mathbf{S}_i$ multiplied by the total number of users mentioning $\mathbf{S}_j$, taking the square root. n is the total number of unique locations found in all comments.

3 Results

To assess the performance of our trained NER model we manually annotated 498 randomly selected Reddit comments with place names. On this test dataset, our model achieved an F1 performance of 0.845, a recall of 0.91 and precision of 0.85, above the expected performance on the WNUT test dataset. Similarly, we annotated 200 comments as neutral, positive or negative sentiment, and found the pre-built sentiment model performed as expected (Loureiro et al. 2022), with an F1, recall and precision of 0.70.

In total 26% of all comments contained at least one place name, 4,475,800 place names were identified, with 2,816,072 (63%) attributed with a set of coordinates. From these locations, 57,682 were found to be unique.

3.1 Place connectivity

Figure 1 demonstrates mental place connections when aggregated to Local Authority Districts (LAD). This aggregation is performed by treating a place equivalent to the LAD it is contained in. Aggregation allows for a higher level view of place connections to be observed, for example combining several points of interest within a major city produces a single strong connection, rather than several weaker connections. Figure 1 (a) shows connections within England and Wales. Clusters emerge where isolated urban areas and their surrounding LADs exhibit strong connectivity, for example around London, Liverpool and Manchester, and Bristol. Notably Wales appears reasonably isolated from England, but inter-connectivity between LADs is strong, including in more rural areas. This isolation between Wales and the rest of the UK has also been observed through semantic analysis of Twitter (Ostermann et al. 2015).

Figure 1 (b) shows connections within Scotland, these are much stronger generally than the rest of the UK, and highly inter-connected. These connections are also less isolated compared with Wales, with links between Glasgow, Edinburgh and London, as well as strong links with Durham and Newcastle. Links in Scotland are not restricted to urban areas, with strong connections between the rural Highland LAD and all major urban centres. These connections may however be influenced by the varying levels of ambiguity between English place names compared with names in remote Scotland. For example in the 'Highland' LAD, 27% of place names are ambiguous, compared with 40% in Manchester, meaning toponym disambiguation is likely more accurate in the Highlands.

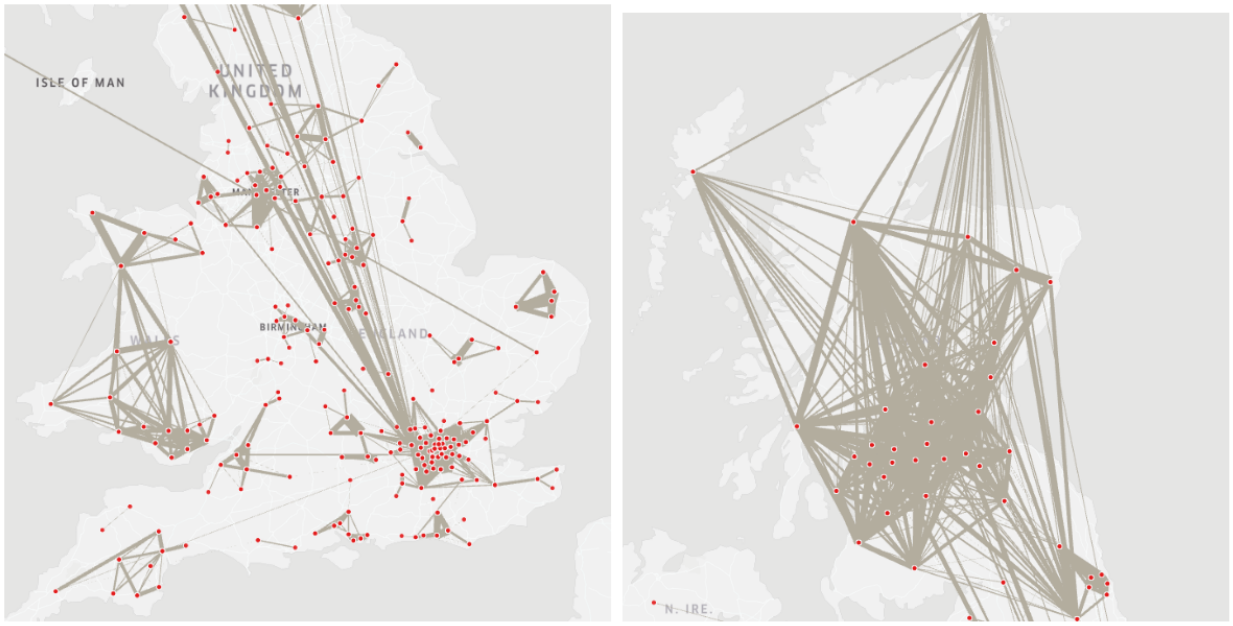
Figure 1 (c) shows a highly urbanised area of England, with two major cities; Liverpool and Manchester. While Manchester links directly with Liverpool, this figure does not appear to reflect the contiguous urban area that links these two cities (Dembski 2015), given intermediate LADs are not connected. The link between these two cities appears direct from a mental perspective, meaning intermediate urban zones are perceived as less connected through our index.

While there are a multitude of factors that affect this mental place connectivity, typical connectivity models explore the effects of distance decay (Yang, Li, and Li 2019; Gong et al. 2021; Li et al. 2021; Bailey et al. 2018; Hu, Ye, and Shaw 2017). As past work has found, with both social and physical connections between places, our mental PCI does experience distance decay, with a medium strength negative correlation between PCI and the log of distance (R = 0.55; p = 0.0; Figure 2 (c)). Additionally we explore how sentiment and deprivation may influence PCI values. Figure 2 (a) shows a significant weak correlation between PCI and sentiment (R = 0.18; p = 0.0), while Figure 2 (b) shows a very weak negative correlation between PCI and the IMD Score (R = 0.04; p = 0.0).

Lower connectivity for places with more negative sentiment may be influenced by factors like 'fear of crime' (Solymosi et al. 2021), and given deprivation appears less influential, these perceptions likely capture information that is not directly quantifiable through traditional data sources. Certain areas in the United Kingdom have been shown to have 'reputations' (Kearns, Kearns, and Lawson 2013), meaning they are known to be viewed negatively both inside and outside their occupants, particularly between the North and South of England (Gould and White 1986).

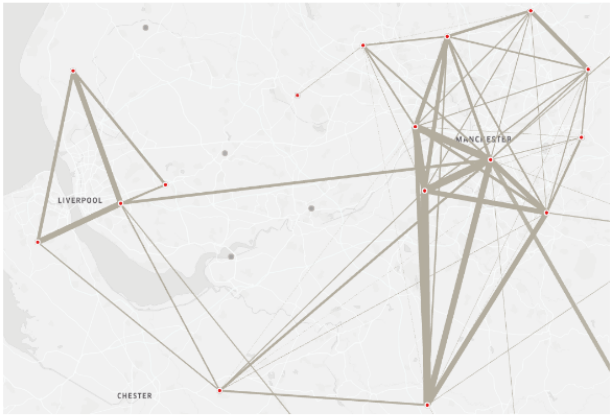
4 Conclusion

Our paper demonstrates the use of Reddit to explore mental place connectivity from a novel source of data, without the reliance on explicit geographic information. Despite Reddit being both more pseudo-anonymous than Twitter (users



(a) England and Wales

(b) Scotland



(c) Liverpool and Manchester

Figure 1: Place connections aggregated to Local Authority District, edge size weighted by PCI values, full interactive map available through Unfolded

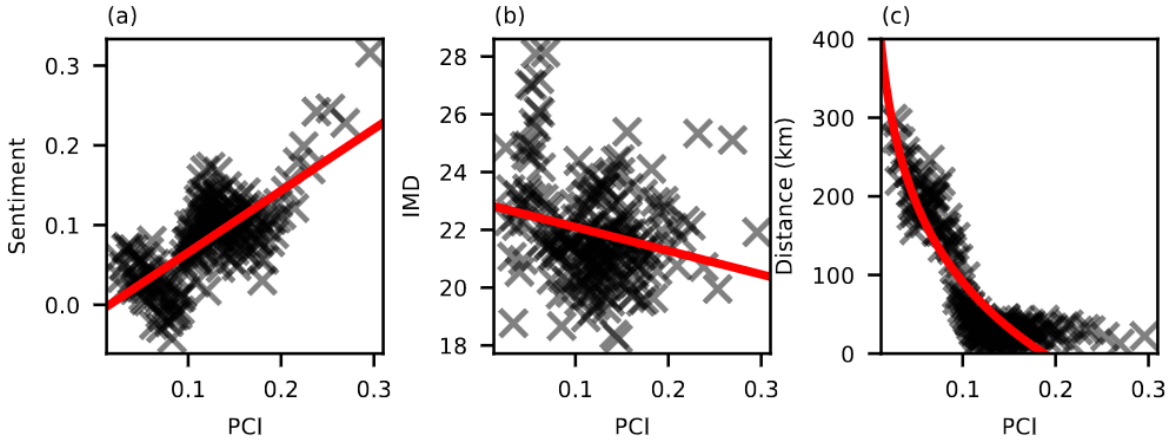


Figure 2: Correlation between PCI and target (a) IMD, (b) Sentiment, and (c) Dis- tance for unaggregated connections between locations. Connections with fewer than 100 shared users have been excluded. Values aggregated into 200 bins for readability.

more frequently use pseudonyms), and without explicit social connections between users (unlike Twitter, followers are not emphasised), mental place connectivity derived through Reddit still exhibits the distance decay that is expected when observing place connectivity (Yang, Li, and Li 2019; Li et al. 2021; Bailey et al. 2018). While Reddit users are not completely representative of the UK population, the volume of unique users that contribute to this corpus are expected to better reflect a more accurate, general view of mental connectivity, compared to alternative text sources like news articles (Hu, Ye, and Shaw 2017).

Future work may consider a further exploration of sentiment between subreddits, London for example may be viewed differently based on subreddit communities focussed on North of England, compared with the South (Gould and White 1986; Jewell 1994). There is also the opportunity to explore relationships between locations or communities through their semantic typology; clustering through Latent Dirichlet Allocation (LDA) (Gao, Janowicz, and Couclelis 2017), or finding the cosine similarity between derived lexicons (Arthur and Williams 2019). Locations identified from within these communities also likely represent urban areas of interest which may be derived based on their frequency of mentions (Chen, Arribas-Bel, and Singleton 2019), or semantic regions that reflect mental perceptions of places (Gao et al. 2017).

Allard, Ryan F., and Filipe Moura. 2016. "The Incorporation of Passenger Connectivity and Intermodal Considerations in Intercity Transport Planning." *Transport Reviews* 36 (2): 251–77. https://doi.org/10.1080/01441647.2015.1059379.

Arthur, Rudy, and Hywel T. P. Williams. 2019. "The Human Geography of Twitter: Quantifying Regional Identity and Inter-Region Communication in England and Wales." Edited by Emilio Ferrara. *PLOS ONE* 14 (4): e0214466. https://doi.org/10.1371/journal.pone.0214466.

Bailey, Michael, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. 2018. "Social Connectedness: Measurement, Determinants, and Effects." *Journal of Economic Perspectives* 32 (3): 259–80. https://doi.org/10.1257/jep.32.3.259.

Ballatore, Andrea, Michela Bertolotto, and David C. Wilson. 2014. "An Evaluative Baseline for Geo-Semantic Relatedness and Similarity." *GeoInformatica* 18 (4): 747–67. https://doi.org/10.1007/s10707-013-0197-8.

Barbieri, Francesco, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification." *arXiv:2010.12421 [Cs]*, October. https://arxiv.org/abs/2010.12421.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. "The Pushshift Reddit Dataset." arXiv. https://arxiv.org/abs/2001.08435.

Chen, Meixu, Dani Arribas-Bel, and Alex Singleton. 2019. "Understanding the Dynamics of Urban Areas of Interest Through Volunteered Geographic Information." *Journal of Geographical Systems* 21: 89–109.

Dembski, Sebastian. 2015. "Structure and Imagination of Changing Cities: Manchester, Liverpool and the Spatial in-Between." *Urban Studies* 52 (9): 1647–64. https://doi.org/10.1177/0042098014539021.

Derczynski, Leon, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. "Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition." In *Proceedings of the 3rd Workshop on Noisy User-generated*

Text, 140–47. Copenhagen, Denmark: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4418.

Farber, Steven, and Xiao Li. 2013. "Urban Sprawl and Social Interaction Potential: An Empirical Analysis of Large Metropolitan Regions in the United States." *Journal of Transport Geography* 31 (July): 267–77. https://doi.org/10.1016/j.jtrangeo.2013.03.002.

Gao, Song, Krzysztof Janowicz, and Helen Couclelis. 2017. "Extracting Urban Functional Regions from Points of Interest and Human Activities on Location-Based Social Networks." *Transactions in GIS* 21 (3): 446–67. https://doi.org/10.1111/tgis.12289.

Gao, Song, Krzysztof Janowicz, Daniel R. Montello, Yingjie Hu, Jiue-An Yang, Grant McKenzie, Yiting Ju, Li Gong, Benjamin Adams, and Bo Yan. 2017. "A Data-Synthesis-Driven Method for Detecting and Extracting Vague Cognitive Regions." *International Journal of Geographical Information Science* 31 (6): 1245–71. https://doi.org/10.1080/13658816.2016.1273357.

Gong, Junfang, Shengwen Li, Xinyue Ye, Qiong Peng, and Sonali Kudva. 2021. "Modelling Impacts of High-Speed Rail on Urban Interaction with Social Media in China's Mainland." *Geo-Spatial Information Science* 24 (4): 638–53. https://doi.org/10.1080/10095020.2021.1972771.

Goodchild, Michael F. 2007. "Citizens as Sensors: The World of Volunteered Geography." *GeoJournal* 69 (4): 211–21. https://doi.org/10.1007/s10708-007-9111-y.

Gould, Peter R., and Rodney White. 1986. *Mental Maps.* Hoboken: Taylor and Francis.

Hu, Yingjie, Xinyue Ye, and Shih-Lung Shaw. 2017. "Extracting and Analyzing Semantic Relatedness Between Cities Using News Articles." *International Journal of Geographical Information Science* 31 (12): 2427–51. https://doi.org/10.1080/13658816.2017.1367797.

Jewell, Helen M. 1994. *The North-south Divide: The Origins of Northern Consciousness in England.* Manchester University Press.

Kamalloo, Ehsan, and Davood Rafiei. 2018. "A Coherent Unsupervised Model for Toponym Resolution." In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 1287–96. Lyon, France: ACM Press. https://doi.org/10.1145/3178876.3186027.

Kearns, Ade, Oliver Kearns, and Louise Lawson. 2013. "Notorious Places: Image, Reputation, Stigma. The Role of Newspapers in Area Reputations for Social Housing Estates." *Housing Studies* 28 (4): 579–98. https://doi.org/10.1080/02673037.2013.759546.

Kuchler, Theresa, Dominic Russel, and Johannes Stroebel. 2020. "The Geographic Spread of COVID-19 Correlates with the Structure of Social Networks as Measured by Facebook." w26990. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w26990.

Li, Zhenlong, Xiao Huang, Xinyue Ye, Yuqin Jiang, Yago Martin, Huan Ning, Michael E. Hodgson, and Xiaoming Li. 2021. "Measuring Global Multi-Scale Place Connectivity Using Geotagged Social Media Data." *Scientific Reports* 11 (1): 14694. https://doi.org/10.1038/s41598-021-94300-7.

Lin, Jinyao, Zhifeng Wu, and Xia Li. 2019. "Measuring Inter-City Connectivity in an Urban Agglomeration Based on Multi-Source Data." *International Journal of Geographical Information Science* 33 (5): 1062–81. https://doi.org/10.1080/13658816.2018.1563302.

Liu, Yu, Fahui Wang, Chaogui Kang, Yong Gao, and Yongmei Lu. 2014. "Analyzing Relatedness by Toponym Co-Occurrences on Web Pages: Analyzing Relatedness by Toponym Co-Occurrences on Web Pages." *Transactions in GIS* 18 (1): 89–107. https://doi.org/10.1111/tgis.12023.

Loureiro, Daniel, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. "TimeLMs: Diachronic Language Models from Twitter." arXiv. https://arxiv.org/abs/2202.03829.

Meijers, Evert, and Antoine Peris. 2019. "Using Toponym Co-Occurrences to Measure Relationships Between Places: Review, Application and Evaluation." *International Journal of Urban Sciences* 23 (2): 246–68. https://doi.org/10.1080/12265934.2018.1497526.

Miller, Harvey J. 2018. "Time Geography." *Handbook of Behavioral and Cognitive Geography*, 74–94. https://doi.org/10.4337/9781784717544.00011.

Murphy, Nicole. 2019. "Reddit's 2019 Year in Review - Upvoted." https://www.redditinc.com/blog/reddits-2019-year-in-review/#content.

Ostermann, F. O., H. Huang, G. Andrienko, N. Andrienko, C. Capineri, K. Farkas, and R. S. Purves. 2015. "Extracting and Comparing Places Using Geo-Social Media." *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3/W5 (August): 311–16. https://doi.org/10.5194/isprsannals-II-3-W5-311-2015.

Patterson, Zachary, and Steven Farber. 2015. "Potential Path Areas and Activity Spaces in Application: A Review." *Transport Reviews* 35 (6): 679–700. https://doi.org/10.1080/01441647.2015.1042944.

Purves, Ross S., Paul Clough, Christopher B. Jones, Mark H. Hall, and Vanessa Murdock. 2018. *Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text*. now. https://doi.org/10.1561/1500000034.

SafeGraph. 2022. "Places Data Curated for Accurate Geospatial Analytics | SafeGraph." https://www.safegraph.com.

Solymosi, Reka, David Buil-Gil, Laura Vozmediano, and Inês Sousa Guedes. 2021. "Towards a Place-based Measure of Fear of Crime: A Systematic Review of App-based and Crowdsourcing Approaches." *Environment and Behavior* 53 (9): 1013–44. https://doi.org/10.1177/0013916520947114.

Statista. 2022. "Most Popular Social Networks Worldwide as of January 2022, Ranked by Number of Monthly Active Users." *Statista*. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

Yang, Yang, Dong Li, and Xiang (Robert) Li. 2019. "Public Transport Connectivity and Intercity Tourist Flows." *Journal of Travel Research* 58 (1): 25–41. https://doi.org/10.1177/0047287517741997.

Ye, Xinyue, Junfang Gong, and Shengwen Li. 2021. "Analyzing Asymmetric City Connectivity by Toponym on Social Media in China." *Chinese Geographical Science* 31 (1): 14–26. https://doi.org/10.1007/s11769-020-1172-6.