The variegated bias of activity spaces derived from mobility data

Ate POORTHUIS[1]; Qingqing CHEN[2]

[1]KU Leuven, Belgium, ate.poorthuis@kuleuven.be (corresponding author)

[2]University of Buffalo, United States, qchen47@buffalo.edu

Keywords: social media data, mobility, activity spaces, big data

Analyzing functional systems, such as cities and commuter or border regions, is ultimately about the analysis of social interactions or mobility across space. These social interactions are conventionally captured in census, register or survey data through, for example, capturing both home and work locations of residents. However, relying solely on travel-to-work data limits our understanding of functional systems to a relatively narrow perspective and makes it difficult to scale up analysis to cross-regional or cross-country systems. As such, researchers have been eager to capitalize on newer mobility datasets such as GPS logs, social media data, mobile phone operator data, and – more recently – mobile phone application data. These datasets potentially cover a much wider aspect of our daily mobility (and thus functional systems) and high temporal and spatial granularity.

Despite this promise, widespread use of such mobility data for the large-scale analysis of functional systems across countries can be hampered by two specific issues. The first issue revolves around **access** to data. Specifically, accessing raw data is often restricted due to, for example economic, privacy and ethical concerns, and a range of other reasons. Even if this is overcome through the social and economic capital of researchers, specific data providers often operate only in a specific local or national geography. Furthermore, raw mobility data is often not readily usable for functional systems analysis and requires non-trivial computational skills and resources to convert to a derived mobility dataset that is more suitable for such analysis. The second issue relates to potential **bias** in such data. As mobility data is not collected in register or random-sampling frameworks, it is not always clear if and how insights derived from such data can be trusted or even generalized.

This study addresses these two challenges by constructing a large-scale mobility dataset derived from all geotagged social media posts on Twitter between 2012 and 2019 across the contiguous United States. The raw data (~3.8 billion data points) is first filtered by removing inactive and nonhuman users, and subsequently enriched by detecting the most likely home location of each user. The resulting data is aggregated temporally and spatially to a ~500m hexagonal H3 grid. To further safeguard privacy, random perturbations are added to both the location and timestamp, and sensitive locations with few observations that may compromise anonymity are excluded. As a result, we generate an aggregated and de-identified dataset of activity spaces of ~10.3 million users nationwide with a collective ~1.2 billion observations. This publicly shareable dataset enables the fast querying of activity spaces in a specific time period or location, allowing researchers to immediately access

and start an analysis on topics such as spatial inequality in activity spaces in Saint Louis, Missouri; or cross-state mobility on the Eastern Seaboard.

Moreover, the breadth of this dataset also allows an in-depth assessment of potential biases introduced by the selective use of a specific social media platform. By comparing representation of user home locations in this dataset with census statistics, we find that correlation between population statistics at the state and county is fairly good (Pearson's r ~0.98) but this correlation gets notably lower at the granular level of the census tract (Pearson's r ~0.35). More importantly, the uneven representation of residents within such a dataset is not homogeneously distributed. We analyze this through a geographically-weighted regression (GWR) approach where the number of social media users with identified home locations within the mobility dataset is explained through the census population, age, income levels, and racial characteristics of each census tract. It is often assumed that social media users are younger, higher-educated, and whiter than the general population. However, we find that this is not necessarily true for this dataset and is highly context-specific. The strength and direction of the relationship between these variables changes from state to state, from urban to rural and even within cities. This analysis shows that bias is indeed an important issue to take into account. Making these biases transparent opens the road to more widespread use of mobility datasets. For example, users within the dataset can be weighted to account for over- and under-representation; studies can oversample from specific areas of interest; and conclusions on functional systems can be contextualized with knowledge of specific biases present in the input data.