A Data Fusion Framework for Exploring Mobility Around Disruptive Events

Evgeny Noi[1]
Department of Geography, University of California Santa Barbara, CA, USA

Somayeh Dodge
Department of Geography, University of California Santa Barbara, CA, USA

Abstract

This paper proposes a data fusion framework that seeks to investigate joint mobility signals around wildfires in relation to geographic scale of analysis (level of spatial aggregation), as well as spatial and temporal extents (i.e. distance to the event and duration of the observation period). We highlight the usefulness of our framework using intra-urban mobility data from Mapbox and SafeGraph for two wildfires in California: Lake Fire (August-September 2020, Los Angeles County) and Silverado Fire (October-November 2020, Orange County). We identify two distinct patterns of mobility behavior: one associated with the wildfire event and another one - with the routine daily mobility of the nearby urban core. Using the combination of data fusion and tensor decomposition, the framework allows us to capture additional insights from the data, that were otherwise unavailable in raw mobility data.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases geographic extent, geographic scale, tensor decomposition, spatio-temporal analysis

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.57

Category Short Paper

Funding *Somayeh Dodge*: NSF Award # 2043202: Modeling Movement and Behavior Responses to Environmental Disruptions.

Acknowledgements The authors gratefully acknowledge the support from the National Science Foundation through award BCS # 2043202. Mobility data provided by ©MapBox and ©SafeGraph.

1 Introduction

The issue of geographic scale (i.e. level of detail / aggregation), geographic extent (i.e. area of analysis, measured in terms of proximity to the phenomena of interest), and temporal extent (i.e. the period of observation in relation to the phenomena of interest) has been an important research topic in GIScience and in movement research within the last decade [5, 3]. Many conventional methods of geographic analysis are traditionally designed for univariate spatial or temporal series (e.g. geographically-weighted regression, local indicators of spatial association, spatial scan statistics). Yet, data on human movement is increasingly heterogeneous [1, 6], multivariate, and dependent on local land-use and transportation patterns, necessitating further development of complex multivariate spatio-temporal methods that can leverage and integrate numerous data sources.

We propose an analytical framework that allows to fuse various indicators of human mobility (in this paper, only two are considered) at different geographic and temporal scales to identify the impact zone of disruptive events, such as wildfires, on mobility. The framework consists of several processes: 1) Multi-scale spatio-temporal matching of data to

[1] corresponding author

combine various types of geographic data (e.g. POI point vector data and regular grid-based OpenStreetMap tiles) geographically and temporally. 2) Calculating mutual information score combining the multiple data sets for different geographic and temporal scale and extent. 3) Fitting the fused data to PARAFAC2 tensor decomposition model [4] to elicit shared patterns of movement at different locations. As a case study, we test this framework for exploring mobility patterns around two wildfire events.

2 Methods

2.1 Multi-Scale Spatio-Temporal Matching and Aggregation

The first step in the proposed framework relies on the matching of the various data sources (in this case two). These data sources vary in coverage, have different units of analysis and units of measurement. The process of matching is described in Algorithm 1. In short, it sequentially increases the spatial scale of the study area (unit of analysis), the spatial extent (e.g. distance from the fire event), and the temporal extent of the observations (e.g. time from fire ignition) to fuse the two mobility indices into a mutual information (MI) score as described in the next section (see Figure 1c).

Algorithm 1 Multi-Scale Spatio-Temporal Matching.

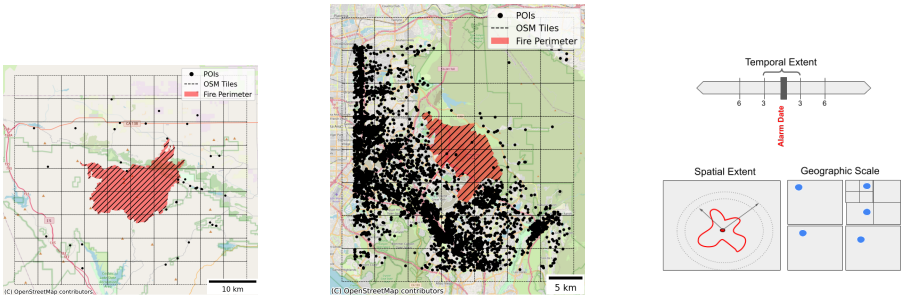
Input : fire perimeter, mobility index 1 ($M1$), mobility index 2 ($M2$), spatial extent (distance from fire perimeter - S_i), temporal extent (days from the fire ignition - T_j), spatial scale / zoom levels (OpenStreetMap zoom level tiles - Z_k)
Output : Mutual information MI_{zst} for various spatial extent ($S \in 1...j$), temporal extent ($T \in 1...j$) and spatial scale ($Z \in 1...k$)

Discretize study area using OSM tiles at level Z_k and aggregate mobility indices (M_1, M_2) at selected spatial S_i and temporal T_j extent levels;
for $z \leftarrow 1$ **to** k **do**
  for $t \leftarrow 1$ **to** j **do**
    for $s \leftarrow 1$ **to** i **do**
      1. Keep only spatial units that have non null values across M_1 and M_2;
      2. Calculate mutual information MI_{zst} from $M1_{zst}$ and $M2_{zst}$;
    end
  end
end

2.2 Mutual Information (MI)

Mutual information (MI) is a measure of the amount of information that two variables share. In information theory, MI is defined as the reduction in uncertainty about one variable (X) given knowledge of another variable (Y). In contrast to Pearson correlation, the MI score is ideally suited to capture non-linear dependence between random variables. Mathematically, the MI between two discrete random variables X and Y is defined as: $MI(X;Y) = H(X) - H(X|Y)$, where $H(X)$ is the entropy of X, which measures the amount of uncertainty in X, and $H(X|Y)$ is the conditional entropy of X given Y, which measures the remaining uncertainty in X when Y is known.

The rationale behind utilizing the mutual information of mobility is simple: the premise is that in the presence of emergency events such as a natural disaster, all types of mobility are affected. As such, variation in mobility will be manifested in various measured indices of mobility. By utilizing mutual information across different geographical scales and spatio-temporal extents, we hope to establish and characterize mutual dependence of mobility indices and fuse movement data that may be different in coverage, uncertainty, and bias.



(a) Lake Fire.

(b) Silverado Fire.

(c) Geographic scale, spatial and temporal extent in relation to fire and mobility data.

Figure 1 POI location and OSM grid representation of the study areas. For demonstration purpose only the zoom level 13 is illustrated, denoting the coarsest level of detail.

2.3 Tensor Decomposition

Tensors are multi-dimensional arrays and, as such, require multi-dimensional methods of analysis. Tensor decomposition allows us to uncover hidden latent factors (clusters of behavior) in multi-dimensional data. There are different types of tensor decomposition, including Tucker, CANDECOMP, and Tensor-Train [7]. Of particular interest to this study is the PARAFAC2 factorization [4], because it allows to jointly model data arrays of different sizes (for instance, where the spatial extent of data varies), by aligning them across a shared dimension (e.g. time). The multiset data can be decomposed as follows: $\mathbf{X_k} = \mathbf{U_k S_k V^T}$, where R is the number of components derived from the decomposition, $\mathbf{U}_k \in \mathbb{R}^{I_k \times R}$, $\mathbf{S}_k \in \mathbb{R}^{R \times R}$, and $\mathbf{V} \in \mathbb{R}^{J \times R}$. The PARAFAC2 decomposition is fitted via an alternating direction method of multipliers (AO-ADMM) [9] available through MatCouply Python package [8]. Since MI scores are always non-negative, we impose non-negativity constraints on the uncovered decomposition components (factors).

3 Case Study

3.1 Study Area

This study focuses on two major wildfires in California (Figure 1a, 1b) that happened in 2020: Lake Fire, which burned around 31,000 acres in Angeles National Forest in Los Angeles County from August 12 to September 28 and Silverado Fire, which burned around 13,000 acres in Orange County from October 26 to November 7, 2020. The data for this study was collected specifically for two months before and after the ignition date of the two wildfires: Silverado Fire (August 26 - December 26, 2020) and Lake Fire (June 12 - October 12, 2020).

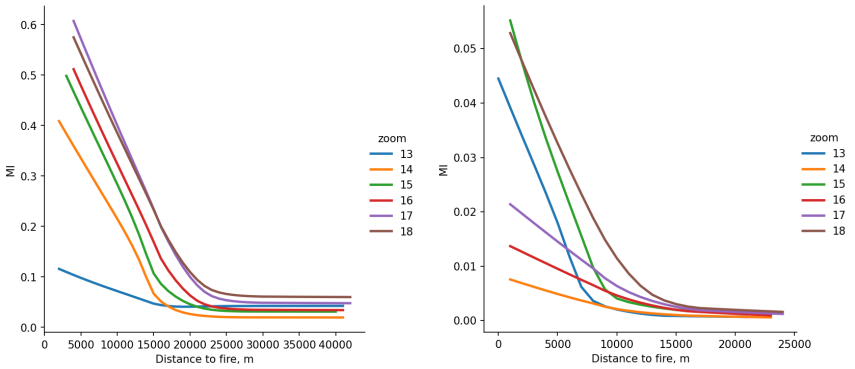
3.2 Data

SafeGraph published several mobility data products during the COVID-19 pandemic through their Data for Good Initiative. One of such products is *Weekly Patterns*, which reports raw visitor counts at the points of interest (POI) level daily. Mapbox provides gridded data, representing the amount of mobility, at OpenStreetMap (OSM) tile level (with OSM zoom level 18, finest resolution available, corresponding to a square grid of 100m $\times$ 100m). The data is aggregated and delivered daily, and is available in the form of an activity index, ranging from 0 to ∞, where higher index values denote higher levels of mobility.

To investigate the mutual information at various levels of aggregation using the proposed framework, we utilize OSM zoom levels 13–18, where zoom level 13 corresponds to a 1:70,000 screen scale (village level), and level 18 corresponds to a 1:2,000 scale (buildings/trees levels)[2]. OpenStreetMap tiles are regular square tessellations that remain uniform across remote and isolated areas where wildfires occur. Thus we can ascertain mobility at different spatial scales, while minimizing modifiable areal unit problem [2]. To delineate the study area, we create a bounding box from a $10km$ buffer around each of the wildfire perimeters and filter the mobility data to this extent. We hypothesize that direct impact zone of wildfires will be pronounced the most in close vicinity to the fire.

3.3 Data Tensorization

The mutual information values are shaped into a multiset data $\mathbf{X_F}$, where $\mathbf{X} \in \mathbb{R}^{I \times J}$ is a matrix with spatial extent on the rows (I), temporal extent on the columns (J), and F denotes a fire name (in Algorithm 1). The bins for spatial extent are calculated starting from the centroid of the fire perimeter, and are incrementally increased by $1km$, resulting in a progressively expanding geographical area around the fire. The distance of $1km$ provides a balanced binning for remote areas, when mobility data is sparse. The temporal extent (T_j) is measured in the number of days before and after the fire. For instance, if $j = 3$ we have a period of 6 days, starting 3 days before fire and terminating 3 days after the fire ignition. These bins are progressively increased by the increment of 3 to the total of 21 bins (i.e. 62 days or roughly two months). Thus, the final dimensions of the multiset data are as follows: $\mathbf{X_{lake}} \in \mathbb{R}^{43 \times 21}$ and $\mathbf{X_{silverado}} \in \mathbb{R}^{25 \times 21}$. Since the fire perimeters vary in size and shape, buffering and discretizing the study area will result in different number of spatial bins ($\mathbf{X}$ rows): 43 rows for the Lake fire and 25 rows for the Silverado fire.



(a) Lake Fire. **(b)** Silverado Fire.

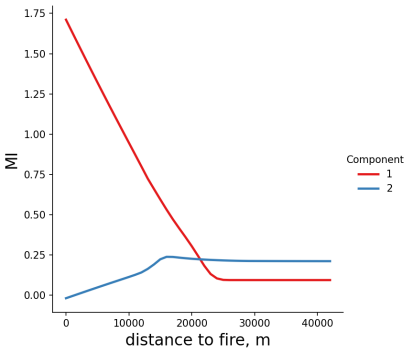
Figure 2 Mutual information score curves for different levels of analysis (zoom levels) and geographical extent (radii from the fire).

4 Results

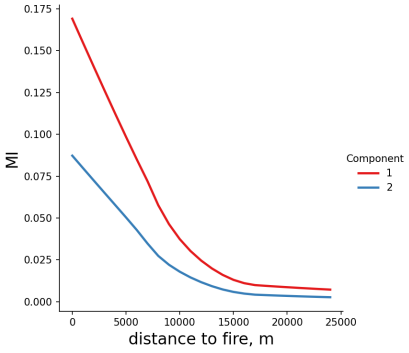
The fitted curves of the matched mutual information scores are plotted against the distance to the fire in Figure 2. For both fires the mutual information score decreases across various zoom levels as the distance from the fire increases. This is logical: as we include more spatial

[2] For more details see `https://wiki.openstreetmap.org/wiki/Zoom_levels`

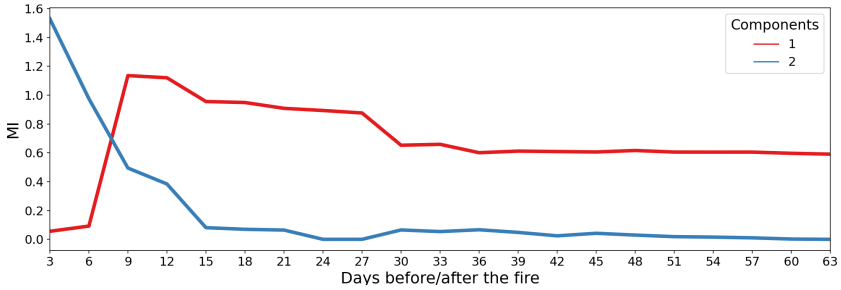
units into our area of interest, we are also capturing daily urban mobility signal and noise. Since Lake Fire perimeter acreage is higher, the curves plateau around $20km$ from the fire (Figure 2a). One important difference between the two sets of curves is the magnitude of the fitted curves (noted in the different y-axis) which is largely due to drastically different number of POI at the two locations: so much so that the the MI scores are dominated by daily mobility, and not wildfire related mobility. This is supported by lack of relationship between temporal extent and MI scores.



(a) Spatial view of uncovered decomposition components for Lake Fire.



(b) Spatial view of uncovered decomposition components for Silverado Fire.



(c) Temporal view of uncovered decomposition components for both fires.

Figure 3 PARAFAC2 modeling results.

PARAFAC2 decomposition allows us to analyze both fires jointly, identifying distinct movement patterns (Figure 3) that are shared across two wildfires. A model with two decomposition components ($R = 2$) fits the data very well (99% fit). *Component 1* (denoted in red) shows fire-related mobility signal and plots MI scores against the spatial extent (Figure 3a, 3b). As we increase the geographic extent, the mutual information decreases for both fires (with more rapid decrease for sparse Lake Fire data), pointing to the higher dependence of mobility indices in close proximity to the fire event. On the temporal view for *Component 1* (Figure 3c) we notice that the MI scores are highest for the observation period of 9-12 days before/after the ignition date of the fire, declining gradually. This might be an indication that we need a relatively extended period of time to establish the effect of the wildfire. *Component 2* (denoted in blue) shows daily mobility associated with nearby urban cores at two locations. For Lake Fire (Figure 3a) the closest urban area, Lancaster is located approximately $15km$ to the Northwest of the fire (peak for blue line). For Silverado Fire (Figure 3b) the urban area borders with the fire perimeter on the Southeast, and as such, coincides with direct impact zone of the wildfire. On the temporal view for *Component 2*, the

MI scores fall abruptly, approaching zero around 15 days before/after the fire. This is logical, as we increase the observation period for two urban areas, there is no added information about the fire event.

5 Conclusion

This paper demonstrated how the proposed framework can be used to fuse the data on mobility at different spatial and temporal scale and establish relationships between mutual dependence of mobility indices around disruptive events. The mutual information score coupled with tensor decomposition is able to identify two clusters of behavior, which were otherwise not traceable in raw mobility signals (e.g. SafeGraph visitor counts or Mapbox activity index). The framework presented in this paper can be easily scaled up to incorporate more locations, different event types (hurricanes, floods, etc.) and event duration, and mobility indices. Future work will compare the methods laid out in the paper to other clustering techniques for studying aggregate human movement.

References

1 Somayeh Dodge. A data science framework for movement. *Geographical Analysis*, 53(1):92–112, 2021. doi:10.1111/gean.12212.

2 A Stewart Fotheringham and David WS Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A*, 23(7):1025–1044, 1991. doi:10.1068/a231025.

3 Michael F Goodchild. A GIScience perspective on the uncertainty of context. *Annals of the American Association of Geographers*, 108(6):1476–1481, 2018. doi:10.1080/24694452.2017.1416281.

4 Henk AL Kiers, Jos MF Ten Berge, and Rasmus Bro. PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 13(3-4):275–294, 1999. doi:10.1002/(SICI)1099-128X(199905/08)13:3/4\%3C275::AID-CEM543\%3E3.0.CO;2-B.

5 Mei-Po Kwan and Tijs Neutens. Space-time research in GIScience. *International Journal of Geographical Information Science*, 28(5):851–854, 2014. doi:10.1080/13658816.2014.889300.

6 Evgeny Noi, Alexander Rudolph, and Somayeh Dodge. Assessing COVID-induced changes in spatiotemporal structure of mobility in the United States in 2020: a multi-source analytical framework. *International Journal of Geographical Information Science*, 36(3):585–616, 2022. doi:10.1080/13658816.2021.2005796.

7 Evangelos E Papalexakis, Christos Faloutsos, and Nicholas D Sidiropoulos. Tensors for data mining and data fusion: Models, applications, and scalable algorithms. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2):1–44, 2016. doi:10.1145/2915921.

8 Marie Roald. MatCoupLy: Learning coupled matrix factorizations with Python. *SoftwareX*, 21:101292, 2023. doi:10.1016/j.softx.2022.101292.

9 Marie Roald, Carla Schenker, Vince D Calhoun, Tulay Adali, Rasmus Bro, Jeremy E Cohen, and Evrim Acar. An AO-ADMM approach to constraining PARAFAC2 on all modes. *SIAM Journal on Mathematics of Data Science*, 4(3):1191–1222, 2022. doi:10.1137/21M1450033.