

Extracting Movement-based Topics for Analysis of Space Use

G. Andrienko^{1,2} , N. Andrienko^{1,2} , and D. Hecker¹ 

¹Fraunhofer IAIS, Germany

²City University of London, UK

Abstract

We present a novel approach to analyze spatio-temporal movement patterns using topic modeling. Our approach represents trajectories as sequences of place visits and moves, applies topic modeling separately to each collection of sequences, and synthesizes results. This supports the identification of dominant topics for both place visits and moves, the exploration of spatial and temporal patterns of movement, enabling understanding of space use. The approach is applied to two real-world data sets of car movements in Milan and UK road traffic, demonstrating the ability to uncover meaningful patterns and insights.

CCS Concepts

- *Human-centered computing* → *Visual Analytics*;

1. Introduction

Movement data plays a crucial role in many different fields. Analysis of movement can concentrate on three key aspects: the moving objects, the spaces they move through, and the timing of their movement [AAB*13]. Different analysis methods are utilized depending on the intended focus. The goal of this paper is to identify patterns and trends in how objects move through a space, in order to gain a deeper understanding of how the space is being utilized. This understanding is valuable for various purposes such as optimizing space usage, enhancing safety, and informing design choices.

Our analysis aims to uncover common patterns of space utilization, including the division of space into groups of places that exhibit similar use by moving objects. We aim to compare different places, examine patterns of movements between them, and establish relationships between places, movements, and their attributes. We investigate the potential of using topic modeling methods for visual analytics of movement with a focus on space.

Traditionally, places are characterized according to points of interest they include, events that happened in places, or categories of moving objects visiting them, taking into account the timing of visits [AAFJ16]. In this work we characterize places according to their connectivity to other places, as inferred from trajectories.

There are two aspects of space use by moving objects: which places are visited and how the objects move between the places. Respectively, we encode trajectory data as abstract texts in two complementary ways. In the first representation, the texts consist of terms representing places, in the second - moves between the places. We discuss how visual analytics can support knowledge extraction by topic modelling in those texts, demonstrate our approach on example data sets, and discuss lessons learned.

2. Background: visual analytics of movement

Regardless of how the movement data is collected, it typically consists of position records in the form of $\langle \text{identity}, \text{position}, \text{time}, \text{attributes} \rangle$ [AAB*13, DBC*15, AAC*17]. Accordingly, analysis of movement data can focus on the moving objects, space, and time, taking into account the remaining aspects and attributes. Methods that focus on space are composed of three main components:

1. Defining relevant places in the space based on movement data. This involves identifying key locations within the movement data, such as frequently visited places or places of movement events, such as stops, turns, or traffic jams.
2. Characterizing places by attributes: This step involves describing the places of interest using various features, such as multivariate time series that summarize movement characteristics.
3. Uncovering relationships between places, their features, and moves: This involves identifying correlations and relationships between the places and the attributes that describe them and between the places of interest and moves.

There are several ways to divide a space into places. One way is to use pre-existing geographical or administrative boundaries. Another way is to divide the space based on specific areas, such as street segments or intersections. Sometimes a division into equal-sized rectangles or hexagonal grids is used. A better way is to use the data-driven tessellation based on the density of characteristic points of movement [AA11].

Having a discrete set of places, it is possible to aggregate trajectories by these places and, additionally, by time intervals. In this way each place is characterized by multivariate attributes and time series representing counts of visits and distinct visitors, times and durations of presence etc. Such aggregates can be computed

both for places (representing characteristics of presence in places) and for directional links, or moves between places (representing characteristics of movement between places). Once the places and moves have been characterized by attributes, various methods, such as similarity search, dimensionality reduction, and clustering, are used to uncover the relationships between places. One method that has potential for use but is still underutilized is topic modelling.

3. Related work: topic modelling

Topic modelling is a method for discovering abstract themes or topics in a collection of documents [VK20]. It is widely used in text mining and has become an important tool for uncovering hidden structures in text data. The two most commonly used methods for topic modelling are Latent Dirichlet Allocation (LDA) [BNJ03] and Non-negative Matrix Factorization (NMF) [LNC*17]. In topic modeling, each topic is expressed as a set of terms that are representative of it. The terms are assigned probabilities or weights that reflect their relative importance within the topic.

Topic modelling methods are very sensitive to their parameters such as the desired number of topics, frequency of words to be considered or ignored by the method, initialization procedure, just to name a few issues. There exist sophisticated visual analytics tools for user-steerable topic model optimization, for example [EASD*19, CAA*20].

When applied to texts, topic modeling traditionally considers each document as a bag of words, disregarding their order. However, if the order of words is important for the analysis, terms can be created from ordered pairs of words that appear in the documents as a sequence [Wal06]. This approach increases the size of the vocabulary, but it can provide more meaningful results.

Beyond text analysis, topic modelling methods can be applied to abstract documents consisting of "terms" of any nature. If, after aggregating movement data by areas, a trajectory is represented as an ordered sequence of places A, B, C, \dots , this trajectory can be treated as a document consisting of words A, B, C, \dots (if order is ignored) or words $A \rightarrow B, B \rightarrow C, \dots$ (if order is essential). In this case, the "terms" in the abstract document represent either the places themselves or the moves between places.

These two approaches have been applied to taxi trajectories in papers by Chu et al [CSZ*14] and Liu et al [LJY*19], respectively. Both papers construct a vocabulary from street segments and their ordered pairs, with the goal of finding patterns in the trajectories and focusing on the moving objects. Our work expands on these ideas with a different goal: finding patterns in space, understanding space structure, and revealing how the space is used by movers.

4. Approach

To grasp the overall structure of space, we need to divide the continuous space into a discrete set of places that is sufficiently big for uncovering essential differences in space use at a desired level of abstraction but not so large that it becomes unwieldy. Using street networks as the basis for this division is not ideal because the sheer number of street segments is overwhelming, and also because this approach does not allow for abstraction and large-scale analysis.

Therefore, we suggest dividing the space into a smaller number of places, such as Voronoi polygons, which are defined by the proximity of locations to given seeds, such as spatial densities of characteristic points of the trajectories [AA11]. This division can be refined to a desired level of detail, while still retaining the necessary information about the space.

We will demonstrate our approach using a data set of movement of 17,000 cars in Milan during a one-week period, about 2,000,000 positional records in total (Fig. 1). This data set has been extensively analyzed in multiple studies (e.g., in [AA11, AAB*13, AA13, AAR16]), so the major patterns in the data are already known. This provides us with a valuable opportunity to validate our new findings against established knowledge. Taking into account the scale of the city and its road network, we target at places of about 1km radius. After cleaning the trajectory data and dividing them into 51,498 trips, we applied data-driven tessellation procedure [AA11] and obtained 451 polygons, 385 of those were crossed by the trajectories (Fig. 1). These polygons form our set of places p_0, p_1, \dots, p_{384} . Respectively, trajectory t_i that starts in place $p_{t_i^0}$ and ends in place $p_{t_i^{N_i}}$ receives two complementary representations:

1. List of visited places: $p_{t_i^0}, p_{t_i^1}, \dots, p_{t_i^{N_i}}$
2. List of moves: $p_{t_i^0} \rightarrow p_{t_i^1}, p_{t_i^1} \rightarrow p_{t_i^2}, \dots, p_{t_i^{N_i-1}} \rightarrow p_{t_i^{N_i}}$

We treat each trajectory as a *document* (bag of words) consisting of *terms* representing either places $p_{t_i^j}$ or moves $p_{t_i^j} \rightarrow p_{t_i^{j+1}}$ depending on the chosen representation.

Let's start with the representation in the form of lists of visited places. The data set as a whole is considered as a *corpus* consisting of 51,498 *documents*, with a *vocabulary* consisting of 385 distinct *terms*. In text mining, a corpus with such characteristics is considered as a suitable subject for applying topic modelling methods. Taking into account that the documents are rather short, and the vocabulary is not very extensive, it is recommended to apply NMF [LNC*17] instead of more popular LDA [BNJ03], similarly to established practices in social media text analysis [VK20].

It is known [WVJ16, EMK*21] that the results of topic modeling can vary significantly based on the number of topics desired. To determine the optimal number of topics, we use an ensemble approach as described by Chen et al. [CAA*20]. This involves running NMF

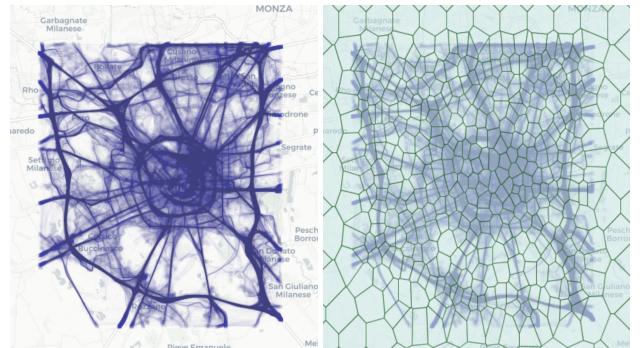


Figure 1: Trajectories of cars (left) and tessellation (right).

multiple times within a specified range of target parameters, combining all obtained topics into one table, and reducing the dimensionality of the topics using t-SNE [vdMH08]. The embedding of the NMF outputs from 11 iterations for the target number of topics ranging from 15 to 25 is shown in Fig. 2-middle. Strong clustering of topics is observed, indicating consistent results with only slight variations. To determine the target number of topics, we use color encoding in the embedding space according to the iteration number. As seen in Fig. 2-left, comparing the two extreme values of 15 (blue dots) and 25 topics (red dots) shows that the smaller number misses several clusters, while the larger number overpopulates the embedding space. Based on this, we interactively choose to acquire 21 topics (Fig. 2-right).

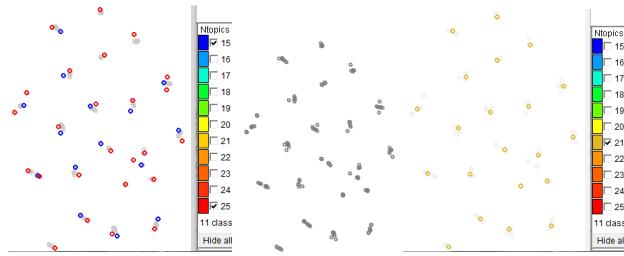


Figure 2: t-SNE embedding of NMF outputs of 11 runs (middle); runs with 15 and 25 topics are highlighted in blue and red (left); run for 21 topics is marked in orange (right).

A single run of NMF generates two output matrices: topic-term and document-topic. The first matrix represents 21 topics by assigning weights to each of the 385 terms (places). The second matrix assigns weights of the 21 topics to each of the 51,498 documents (trajectories). In the topic-term matrix for NMF, the weights represent the contribution of each term to a particular topic, reflecting the relative importance of each term in the topic. These weights are non-negative, and they do not have to sum to 1. In contrast, for LDA the weights in the topic-term matrix indicate the probability of each term given a topic, always summing to 1.

Figures 3 and 4 visualize data from the topic-term matrix on the city map. In Figure 3, each place is colored according to the dominant topic assigned to it. Figure 4 provides pie charts that show the compositions of topics in each place. It is important to note that the colors indicate the similarity of the topics based on closeness of the term weights. To assign these colors, the set of 21 topics described by vectors of 385 term weights is projected onto a 2D space using one of existing dimensionality reduction methods, namely, MDS [Kru64]. The positions in this space are color-coded using the Cube Diagonal Cut B-C-Y-R color map [BSM^{*}15, BDB^{*}16].

We stress the need of using different embedding methods for different purposes. The neighborhood-preserving t-SNE is used for selecting the optimal number of topics, while the better preserving long distances MDS is used for assigning colors to topics.

On the map displayed in Figure 4, it is apparent that some individual places, as well as larger contiguous groups of places, are primarily linked to a single topic. Conversely, there are also areas comprised of places that exhibit a combination of two or more topics. The cross-border diffusion between regions can also be seen.

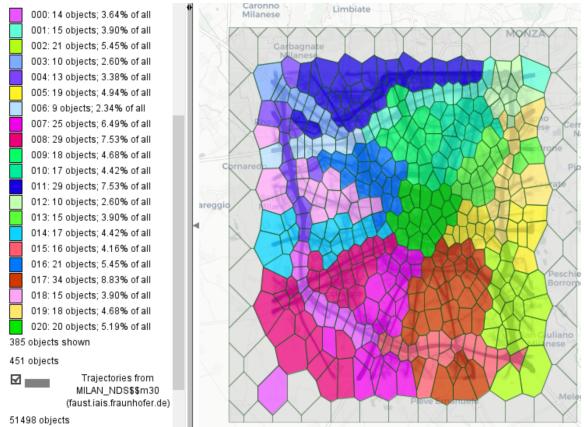


Figure 3: Dominant presence topics are shown for all places.

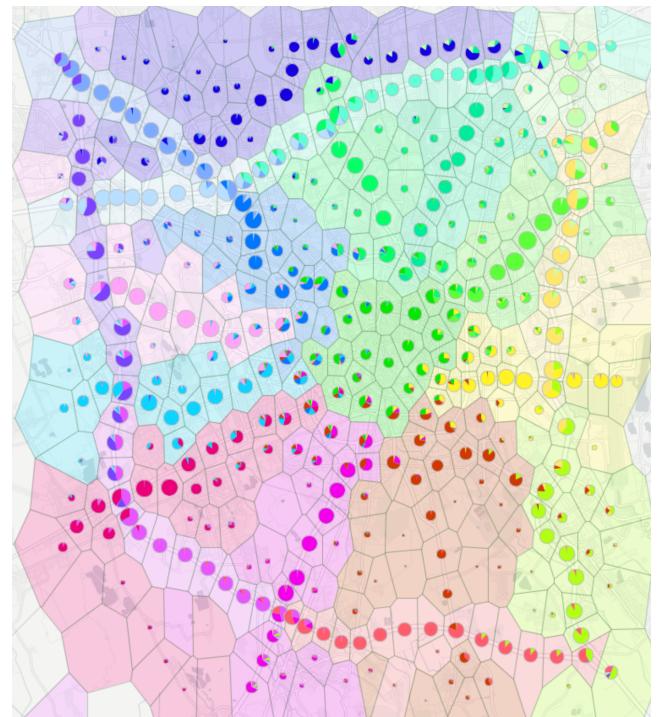


Figure 4: Composition of presence topics in places. NMF weights of topics for each place are represented by pie charts. Colors are assigned to topics according to the topics' positions in MDS embedding using the 2D Cube Diagonal Cut B-C-Y-R color map.

Similarly, we applied NMF to the representation of the data set as a *corpus* consisting of the same trajectories, with a *vocabulary* consisting of 2,156 distinct *terms* representing directed *moves* between places $p_i \rightarrow p_j$. The iterative execution of NMF with numbers of topics in the range from 20 to 35 followed by visual exploration of the embedding space suggested acquiring 30 topics. Again, two matrices are computed: assigning term weights to the topics, and assigning topics weights to documents. Figure 5 visualizes the spatial distribution of topics across the moves. Small multiples in Fig. 6 shows the footprints of all 30 topics. It is evident that a signifi-

cant number of these topics have counterpart topics that represent moves in the opposite direction. However, the colors of the contrasting topics are relatively similar, as they share a portion of their catchment areas outside the main roads, resulting in their proximity in the embedding space. The catchment areas nicely correspond to the regions of topic spread in the place topics map (Fig. 4).

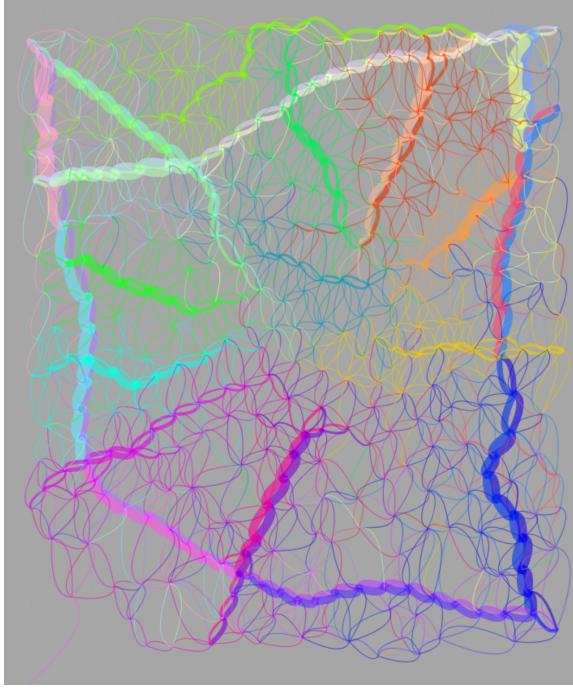


Figure 5: Move topics colored according to their similarity.

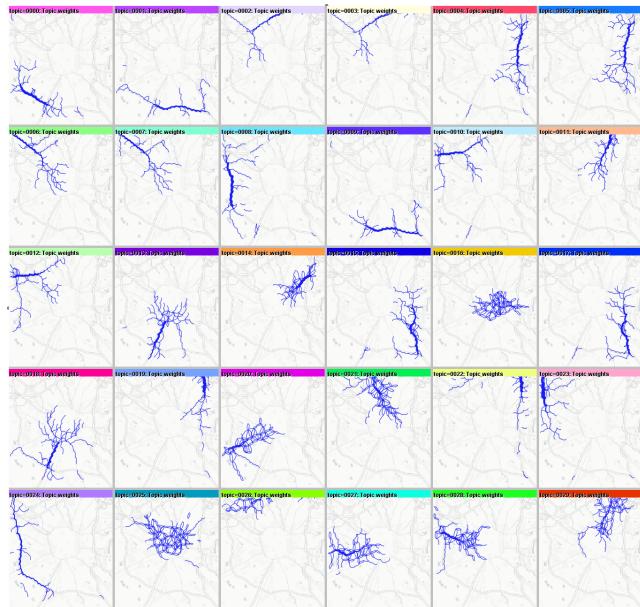


Figure 6: Small multiples of footprints of 30 move topics.

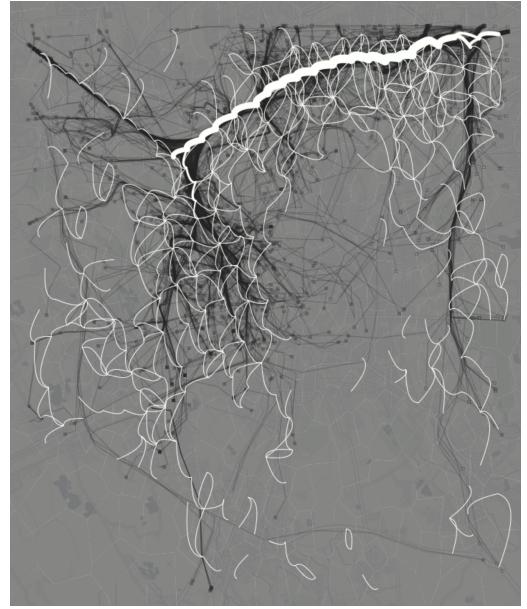


Figure 7: 1,332 trajectories associated with topic 02 and corresponding moves with their weights expressed by line width.

The interactive map allows for exploration of topics and their corresponding documents (trajectories) using dynamic query controls. As an example, Figure 7 shows the trajectories and moves having the highest weight of topic 02, meaning that topic 02 is the dominant topic among all topics for these trajectories and aggregate moves in the document-topic and topic-term matrices, respectively. The trajectories are depicted with a high level of opacity and the moves are represented by directed curves whose widths indicate the weight of topic 02. Any moves with a zero weight are omitted.

The results of the place-based and move-based topic modeling are highly consistent. Both approaches reveal similar mobility regions and catchment areas of major road network elements. City center is clearly separated from radial and belt roads and suburban areas. Our findings confirm that opposite directions along major roads frequently (but not always) belong to different topics, while this happens rarely in the city center and rural areas. This provides additional interesting insights to earlier studies such as [AAB^{*}13].

To summarize, our approach involves the following five steps:

1. Create a space tessellation at a desired level of abstraction using data-driven methods.
2. Transform trajectories into sequences of visited places and moves between places.
3. Determine the optimal number of topics for both places and moves through iterative topic modeling, projecting the topics to a common embedding space, and analyzing their distributions.
4. Analyze selected sets of topics for both places and moves.
5. Integrate and synthesize knowledge acquired at steps 3 and 4.

After conducting topic modeling for both places and moves and examining their distributions, the findings from step 3 and step 4 are integrated and synthesized to gain a comprehensive understanding

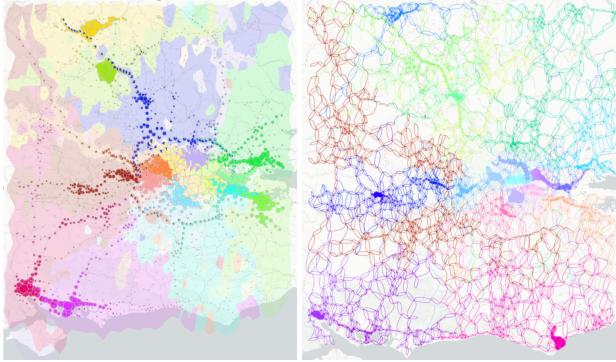


Figure 8: Dominant topics for places (left) and moves (right).

of the mobility patterns in the data set. This involves comparing and contrasting the topics of visits and the topics of moves, and using the results to gain insights into the spatial and temporal aspects of human mobility. This step provides a comprehensive view of the mobility patterns, which combines the information obtained from both the places and moves, producing a holistic understanding of the mobility patterns that is derived from the combined results.

In order to validate our approach, we applied it to several different data sets, including the UK road traffic data set [AAP^{*}21] which contains approximately 57K trajectories recorded over a 13-day period in 2017. We divided the space into approximately 1,900 places and obtained around 7,300 directed moves. After exploring different numbers of topics and examining their properties through projections, we arrived at a final result of 18 topics for places and 25 topics for moves, which are displayed in Figs. 8 and 9.

Similar to the Milan example (Fig. 4), we observe grouping of places into homogeneous regions and cross-border diffusion of place topics. However, the topics of moves behave differently. We observe little to no cases of different topics in opposite directions. Instead, we see regional patterns of dominant move topics, similarly to the topics of places. However, investigation of trajectories that are associated with the topics of moves (Fig. 9) highlights the importance of considering direction, as the same move can often be involved with significant weights in multiple topics. It is important to analyze the full spectrum of topic compositions rather than just focusing on dominant topics.

5. Discussion and conclusions

Following the prior works [CSZ^{*}14, LJY^{*}19], our study confirms that topic modelling is a powerful analytical instrument for analysis of movement data. Dual representation of trajectories as place visits and moves allows considering space use from different perspectives, opening up new avenues for exploring movement patterns by using place patterns in analysis of move patterns and relating patterns of different types [AAM^{*}21]. Due to a different analysis focus, our approach differs from earlier works in multiple aspects. Unlike previous studies, we employ a data-driven tessellation to divide space into places at the desired scale and level of detail, rather than constructing the vocabulary from the road network. We utilize the full set of places and moves for topic modeling, rather than restricting the analysis to a smaller subset of the most frequent terms.

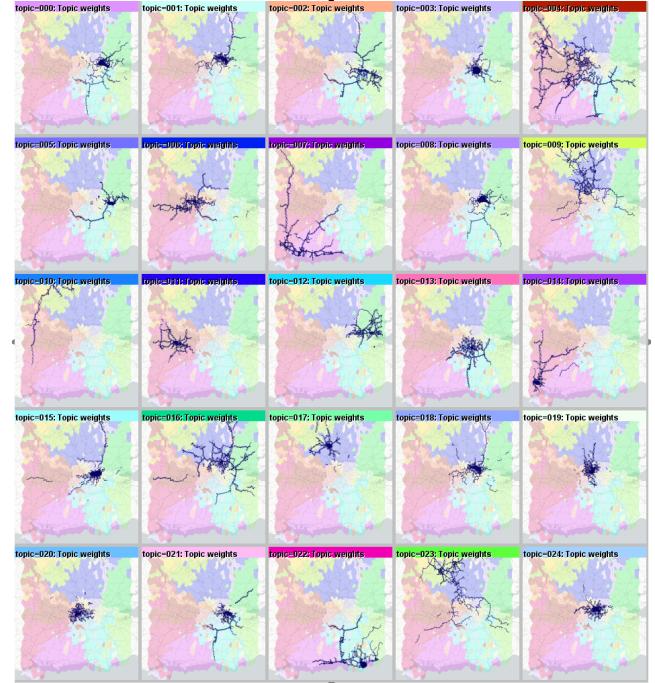


Figure 9: Topic of moves on top of dominant topics of places. Each map represents the weights of a single topic for the moves. The moves having zero weight of a topic are not visible.

A visually-driven strategy is used to determine the optimal number of topics, and new visual representations are proposed to investigate topics in space and relate them to contributing trajectories.

We learned several valuable lessons from our experiments. The results of the analysis heavily depend on the number of topics selected. It is important to note that the number of topics for moves should be higher by a factor of 1.5 to 2.5 compared to the number of topics for places visited. Dominant topics may conceal significant aspects of topic distributions, therefore analyzing the composition of topics for each place or move is crucial. The prominence of patterns is influenced by various factors such as the properties of trajectories, the scale of analysis, and the selected topic modeling parameters. Further investigation is needed to quantify these dependencies and establish guidelines for appropriate parameter settings.

Our approach can be extended to analysing changes in space use over time. To do this, it is necessary to break down the sequences of place visits and moves into subsequences corresponding to different time periods and compute the topic weights for these subsequences based on the occurring terms.

Acknowledgements

This work was supported by Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the *Lamarr Institute for Machine Learning and Artificial Intelligence* (Lamarr22B), and by EU in projects *SoBigData++* and *CrexData* (grant agreement no. 101092749).

References

- [AA11] ANDRIENKO N., ANDRIENKO G.: Spatial generalization and aggregation of massive movement data. *IEEE Transactions on Visualization & Computer Graphics* 17, 02 (2011), 205–219. [doi:10.1109/TVCG.2010.44](https://doi.org/10.1109/TVCG.2010.44). 1, 2
- [AA13] ANDRIENKO N., ANDRIENKO G.: A visual analytics framework for spatio-temporal analysis and modelling. *Data Mining and Knowledge Discovery* 27 (2013), 55–83. [doi:10.1007/s10618-012-0285-7](https://doi.org/10.1007/s10618-012-0285-7). 2
- [AAB*13] ANDRIENKO G., ANDRIENKO N., BAK P., KEIM D., WROBEL S.: *Visual Analytics of Movement*. Springer, 2013. [doi:10.1007/978-3-642-37583-5](https://doi.org/10.1007/978-3-642-37583-5). 1, 2, 4
- [AAC*17] ANDRIENKO G., ANDRIENKO N., CHEN W., MACIEJEWSKI R., ZHAO Y.: Visual analytics of mobility and transportation: State of the art and further research directions. *IEEE Transactions on Intelligent Transportation Systems* 18, 8 (2017), 2232–2249. [doi:10.1109/TITS.2017.2683539](https://doi.org/10.1109/TITS.2017.2683539). 1
- [AAFJ16] ANDRIENKO N., ANDRIENKO G., FUCHS G., JANKOWSKI P.: Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. *Information Visualization* 15, 2 (2016), 117–153. [doi:10.1177/1473871615581216](https://doi.org/10.1177/1473871615581216). 1
- [AAM*21] ANDRIENKO N., ANDRIENKO G., MIKSCH S., SCHUMANN H., WROBEL S.: A theoretical model for pattern discovery in visual analytics. *Visual Informatics* 5, 1 (2021), 23–42. [doi:10.1016/j.visinf.2020.12.002](https://doi.org/10.1016/j.visinf.2020.12.002). 5
- [AAP*21] ANDRIENKO G., ANDRIENKO N., PATTERSON F., CHEN S., WEIBEL R., HUANG H., DOULKERIDIS C., GEORGIOU H., PELEKIS N., THEODORIDIS Y., NANNI M., LONGHI L., KOUMPAROS A., YASAR A., KURESHI I.: Visual analytics for characterizing mobility aspects of urban context. *Urban Informatics* (2021), 727–755. [doi:10.1007/978-981-15-8983-6_40](https://doi.org/10.1007/978-981-15-8983-6_40). 5
- [AAR16] ANDRIENKO N., ANDRIENKO G., RINZIVILLO S.: Leveraging spatial abstraction in traffic analysis and forecasting with visual analytics. *Information Systems* 57 (2016), 172–194. [doi:https://doi.org/10.1016/j.is.2015.08.007](https://doi.org/10.1016/j.is.2015.08.007). 2
- [BDB*16] BERNARD J., DOBERMANN E., BÖGL M., RÖHLIG M., VÖGELE A., KOHLHAMMER J.: Visual-interactive segmentation of multivariate time series. In *Proceedings of the EuroVis Workshop on Visual Analytics* (Goslar, DEU, 2016), Eurographics Association, p. 31–35. [doi:10.2312/eurova.20161121](https://doi.org/10.2312/eurova.20161121). 3
- [BNJ03] BLEI D. M., NG A. Y., JORDAN M. I.: Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022. 2
- [BSM*15] BERNARD J., STEIGER M., MITTELSTÄDT S., THUM S., KEIM D., KOHLHAMMER J.: A survey and task-based quality assessment of static 2D colormaps. In *Visualization and Data Analysis* (2015), Kao D. L., Hao M. C., Livingston M. A., Wischgoll T., (Eds.), vol. 9397, International Society for Optics and Photonics, SPIE, p. 93970M. [doi:10.1117/12.2079841](https://doi.org/10.1117/12.2079841). 3
- [CAA*20] CHEN S., ANDRIENKO N., ANDRIENKO G., ADILOVA L., BARLET J., KINDERMANN J., NGUYEN P. H., THONNARD O., TURKAY C.: Lda ensembles for interactive exploration and categorization of behaviors. *IEEE Transactions on Visualization and Computer Graphics* 26, 9 (2020), 2775–2792. [doi:10.1109/TVCG.2019.2904069](https://doi.org/10.1109/TVCG.2019.2904069). 2
- [CSZ*14] CHU D., SHEETS D. A., ZHAO Y., WU Y., YANG J., ZHENG M., CHEN G.: Visualizing hidden themes of taxi movement with semantic transformation. In *2014 IEEE Pacific Visualization Symposium* (March 2014), pp. 137–144. [doi:10.1109/PacificVis.2014.50](https://doi.org/10.1109/PacificVis.2014.50). 2, 5
- [DBC*15] DEMŠAR U., BUCHIN K., CAGNACCI F., SAFI K., SPECKMANN B., VAN DE WEGHE N., WEISKOPF D., WEIBEL R.: Analysis and visualisation of movement: an interdisciplinary review. *Movement ecology* 3, 1 (2015), 1–24. [doi:10.1186/s40462-015-0032-y](https://doi.org/10.1186/s40462-015-0032-y). 1
- [EASD*19] EL-ASSADY M., SPERRLE F., DEUSSEN O., KEIM D., COLLINS C.: Visual analytics for topic model optimization based on user-steerable speculative execution. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 374–384. [doi:10.1109/TVCG.2018.2864769](https://doi.org/10.1109/TVCG.2018.2864769). 2
- [EMK*21] ESPADOTO M., MARTINS R. M., KERREN A., HIRATA N. S. T., TELEA A. C.: Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics* 27, 3 (2021), 2153–2173. [doi:10.1109/TVCG.2019.2944182](https://doi.org/10.1109/TVCG.2019.2944182). 2
- [Kru64] KRUSKAL J. B.: Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika* 29, 1 (Mar. 1964), 1–27. [doi:10.1007/BF02289565](https://doi.org/10.1007/BF02289565). 3
- [LJY*19] LIU H., JIN S., YAN Y., TAO Y., LIN H.: Visual analytics of taxi trajectory data via topical sub-trajectories. *Visual Informatics* 3, 3 (2019), 140–149. [doi:https://doi.org/10.1016/j.visinf.2019.10.002](https://doi.org/10.1016/j.visinf.2019.10.002). 2, 5
- [LNC*17] LUO M., NIE F., CHANG X., YANG Y., HAUPTMANN A., ZHENG Q.: Probabilistic non-negative matrix factorization and its robust extensions for topic modeling. In *Thirty-first AAAI conference on artificial intelligence* (2017). 2
- [vdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. URL: <http://jmlr.org/papers/v9/vandermaaten08a.html>. 3
- [VK20] VAYANSKY I., KUMAR S. A.: A review of topic modeling methods. *Information Systems* 94 (2020), 101582. [doi:https://doi.org/10.1016/j.is.2020.101582](https://doi.org/10.1016/j.is.2020.101582). 2
- [Wal06] WALLACH H. M.: Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning* (New York, NY, USA, 2006), ICML ’06, Association for Computing Machinery, p. 977–984. [doi:10.1145/1143844.1143967](https://doi.org/10.1145/1143844.1143967). 2
- [WVJ16] WATTENBERG M., VIÉGAS F., JOHNSON I.: How to use t-sne effectively. *Distill* (2016). [doi:10.23915/distill.00002](https://doi.org/10.23915/distill.00002). 2