Addressing the missing location problem in trip chain data: A POI-based probabilistic function method using open data

Peiling Wu*[1], Fariya Sharmeen[1], and Emma Engström[1,2]

[1]KTH Royal Institute of Technology, ABE School of Architecture, Stockholm, Sweden
[2]Institute for Futures Studies, Stockholm, Sweden

Short summary

Trip chaining captures integrated travel patterns besides commuting trips, uncovering activity participation. This study introduces a Point of Interest (POI)-based time-geography methodology to approximate geographic information in trip chaining in travel surveys in a computationally-efficient way. To demonstrate the applicability of this method, travel survey data from Eskilstuna, Sweden, and OpenStreetMap network data are used to achieve trip chain reconstruction with reduced computing time as compared to time geographic density estimation (TGDE). Reconstructed trip chain data provide a foundation for travel behavior pattern simulation, which is central to transport planning.

Keywords: Activity-based model, Location data imputation, Point of Interest, Travel survey, Time geography, Trip-chaining

1 Introduction

Activity-based travel behavior modeling can only explain travel choice dimensions when trip chains are complete and correct (Tajaddini et al., 2020). Trip chaining analysis, which captures interdependence of sequential trips, provides insights into real-world travel decision-making processes (Kitamura, 1984). While travel surveys serve as a primary data source for analyzing household movement patterns (Sun et al., 1998), they present inherent methodological limitations, particularly in capturing comprehensive trip chain attributes such as mode, purpose, and spatiotemporal characteristics (Stopher & Greaves, 2010). The absence of automated location extraction capabilities, unlike GPS tracking or mobile signal data, results in significant spatial data gaps that potentially compromise activity-based transport modeling accuracy (Marchal et al., 2011).

Accordingly, researchers have begun to examine data imputation methods for travel survey data. For example, inspired by clinical information correction technique in healthcare field, (Xu et al., 2002) applied hot-deck and regression imputation, which also have been applied to household travel surveys in early work. (Roux & Armoogum, 2011) used the weight-class method, and (Budhwani et al., 2023) applied the nearest neighbor (NN) hot-deck (k-NN) imputation approach to travel surveys. However, these methods are constrained by their reliance on neighboring data within the missing data category. When there a whole data category is missing, a method to proximate missing data with other existing information is needed. Probabilistic time geography density estimation (TGDE) leverages broader contextual information from adjacent trip sequences for dataset completion.

Time geography, pioneered by (Hägerstrand, 1970), provides a theoretical framework for analyzing spatio-temporal movement patterns. It enables the estimation of spatial locations for moving objects or events under space-time constraints, and it has been used across various domains, including crime pattern analysis, disease transmission studies, and human activity pattern recognition Shaw et al. (2008); Yu (2007); Chen et al. (2011), as well as traffic accident analysis Yamada & Thill (2004). In transportation, the time-geographic method for missing travel location approximation is capable to reconstruct entire mid-spot categories without direct observation Ellegård & Svedin (2012). Building upon classic time geography, probabilistic TGDE introduces an approach that generates continuous probability surfaces, representing likely object trajectories (J. A. Downs, 2010). This advancement enables the visualization of spatial probability distributions through heat map representations, offering enhanced analytical capabilities for movement pattern analysis. TGDE has been successfully extended to network-based applications, which is particularly valuable for transportation studies, where movement is constrained to predefined network infrastructure (Wood & Horner, 2018; J. A. Downs & Horner, 2012).

The network-based TGDE methodology presents a framework for interpolating missing locations in trip chains (Wood & Horner, 2018). However, its implementation faces computational challenges, particularly in the generation of probability surfaces, as it necessitates the calculation of presence probabilities for each discrete network segment (J. A. Downs & Horner, 2014). This computational intensity becomes particularly pronounced when analyzing large-scale travel survey datasets and network systems. The method's scalability is thus constrained by available computational resources, presenting a trade-off between analytical precision and computational efficiency.

Addressing this limitation, the current study aims to explore a method to identify missing location data in trip chains in travel surveys based on the time geography approach (Ellegård, 1999) in combination with open source data from OpenStreetMap. It has three specific objectives: 1. To introduce an innovative method to approximate missing trip stops through the integration of activity-based POI as potential locations, as a computationally practical approach to complete trip chains in large-scale travel survey data. 2. To demonstrate combining time-space constraints with semantic activity information to scale down the number of high potential locations by considering both the spatiotemporal feasibility of movements and the contextual relevance of destinations and activities. 3. To extend midpoint imputation to location sequence approximation by addressing the spatial uncertainty in origins and destinations report in travel survey data. We discuss the applications of this method, its strengths in trip chain reconstruction as compared to existing time geographic methods, as well as limitations and areas for future research.

2 Methodology

We develop a POI-based time geography density estimation approach to identify the most likely activity locations by comparison with semantic information in trip chains. The methodology utilizes two control points (known origin and destination locations) and temporal constraints to estimate potential activity locations (Yin et al., 2022). The network is separated into segments, and for each segment, the probability density function reflects the likelihood of it being an intermediate stop location. The density estimation considers the temporal deviation between the theoretical minimum travel times (from control points to the candidate location) and the observed travel times (to and from the actual intermediate point). This deviation is normalized by the total travel time of the trip chain, resulting in a density measure that is inversely proportional to the temporal discrepancy - higher density values indicate locations that better align with the observed travel pattern (J. Downs, 2014; Wood & Horner, 2018; Horner et al., 2012).

When TGDE is applied to large networks in real cities, the scanning of the whole network for each trip chain is computational inefficient especially considering the high computing power requirement for calculating network-based route length. Additionally, while the time-geographic method ensures location approximation from a computational perspective, multiple segments/areas may exhibit similarly high density values. To address these practical issues, our method accounts for POI information to calculate the route length in advance to reduce the computing time. The methodology also incorporates a systematic categorization of POIs based on their respective land use classifications to identify the most probable stop location by considering the type of POI in the built environment and their alignment with the purpose of trips. The constraints related to POIs and activity types contribute to improving the spatial-semantic precision of the analysis.

The network distances between control points and POIs are computed and stored systematically. C is the set of all control points, K is the set of all POIs, where $i, j \in C$ denote control points and $k \in K$ denotes a POI. The locations of control points are reported as different areas C_a where $a \in A$, the set of geographical areas. In each area, there could be multiple potential control points $i \in C_a$. To extract semantic information from POI types, the whole set K of POIs is partitioned into subsets K_b where $b \in B$ represents different activity types. For example, K_{service} includes all

POIs representing facilities such as hospitals, banks, and car wash services. During the calculation for each trip, all POIs in the subset of the reported activity of the middle point are used. The network distances are computed using Dijkstra's algorithm and stored in the matrix $D \in \mathbb{R}^{|C|\times|K|}$, where:

$$D = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,|K|} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,|K|} \\ \vdots & \vdots & \ddots & \vdots \\ d_{|C|,1} & d_{|C|,2} & \cdots & d_{|C|,|K|} \end{bmatrix} \tag{1}$$

For any trip with origin $i \in C_o$, destination $j \in C_d$, and reported activity type $b \in B$, the relevant distances can be efficiently retrieved from D, considering only the POIs in the corresponding activity subset K_b. After obtaining the distance matrix, and POI subsets, the density function is formulated as a two-step optimization problem:

1. For each potential POI k in K_a, the function $g(k)$ finds the optimal combination of origin i and destination j that minimizes the deviation between calculated and reported travel times. In travel surveys, locations are normally reported as areas such as municipalities, travel analysis zones (TAZ) or postcodes, rather than specific locations. In order to calculate the travel time, the origin and destination need to be narrowed down along a certain location within the area.

$$g(k) = \min_{i \in C_o, j \in C_d} \left(|D_{i,k}/\bar{v}_i - t(i,m)| + |D_{j,k}/\bar{v}_j - t(m,j)| \right) \tag{2}$$

where:
- $i \in C_o, j \in C_d$: Sets of potential origin and destination points within reported areas
- m: the actual intermediate point that is missing
- $D_{i,k}$, $D_{j,k}$: Network distance from origin i/ destination j to POI k
- $\bar{v}_i$, $\bar{v}_j$: Average travel speeds for the trip from origin i or to destination j
- $t(i,m), t(m,j)$: Reported travel times from origin to midpoint and midpoint to destination

The calculation of the average speed is conducted with one or multiple traffic modes reported in each trip. The fastest mode is designated as the dominant mode, while secondary modes are assumed to account for 10% of the total travel time, with the remaining time allocated to the dominant mode. The average speeds in equation (2) are calculated as:

$$\overline{v_i} = (1 - 0.1 \times n) v_{di} + 0.1 \times \sum_{si=1}^{n} v_{si} \tag{3}$$

where:
- $\overline{v_i}$: the calculated average speed of the trip starting/ending at control point i
- v_{di}: the speed of the dominant traffic mode of this trip
- v_{si}: each secondary mode in the reported modes of the trip
- n: the number of secondary modes

2. The function $\hat{f}(x)$ then selects the POI k that maximizes the overall density value, representing the most likely activity location.

$$\hat{f}(x) = \max_{k \in K_a} \left(1 - \frac{g(k)}{t_t(i,j)} \right) \tag{4}$$

where:
- $k \in K_a$: Set of candidate POIs within the activity category
- $t_t(i,j)$: Total reported travel time for the trip

The POI with the highest density value is selected as the most probable activity location for the trip. The optimal combination of origin i and destination j are also selected respectively according to $g(x)$ when there are more than one potential control points.

3 Case Study and Data Description

The POI-based TGDE method was applied to complete a travel survey in Eskilstuna municipality, located in Södermanland County, Sweden, with a population of approximately 107,000 (2024). In

2021, a comprehensive travel survey, known as Resvaneundersökning (RVU), was conducted by the Urban Development Administration of Eskilstuna Municipality in collaboration with the research firm Origo Group. The survey targeted a random sample of 5,000 citizens aged 16 to 85, of which 2,030 individuals responded. The survey was structured into three main sections: (1) background information of respondents, including age, gender, accommodation, household composition, and access to various modes of transport; (2) travel habits and attitudes, with a particular focus on changes induced by the COVID-19 pandemic; and (3) a detailed travel diary capturing the specifics of trips undertaken.

The travel diary component documented all trips made by the respondents on the day preceding the survey, with a maximum of six trips reported per individual. Detailed information was collected regarding trip origin and destination, start and end times, activities, modes of transport, and trip lengths. Origins and destinations were categorized as 'own home', 'own workplace/school', or 'another place'. The locations of home are recorded as 354 different 'Nyko's (which provides a finer geographical division within municipalities (Statistiska Centralbyrån, 2024)) and work/study place are recorded as 10 areas. Aside from spatial information for accommodation and work places, the locations of all reported 'another place's are missing in the survey. The travel survey documented 11 distinct transport modes, encompassing public transport (train, bus), private vehicles (car as driver or passenger, taxi, motorcycle), and non-motorized modes (bicycle, electric bicycle, walking, and others). Nine activity categories were recorded in the survey: education, drop/pick up, grocery, shopping, services, leisure, and visit family/ friends.

The dataset contained 1,482 trip chains. For our analysis, we focused on trip chains consisting of three points (origin, middle point, and destination), where the origin and destination locations were known but the intermediate stop location was unknown. To facilitate direct travel time calculations based on speed, we excluded trip chains involving public transportation modes. This filtering process results in a final sample of 546 trip chains for analysis. The spatial network analysis utilized the OpenStreetMap transport network with nodes and edges (latest updated in February, 2024) for distance matrix generation. 661 control points were included from the travel survey, comprising 651 residential locations and 10 workplace locations from the Eskilstuna municipality data base. There might be multiple, one or zero residential points for each nyko number. All of residential points dropped in the nyko area were included for the calculation of trips reporting the certain nyko, while the centroid of the nyko area was used as the residential location when there is no points registered in that area. All missing locations were represented by 1,849 POIs as potential intermediate stops across the city, divided in different activity categories. Additionally, the 661 control points were also included as potential stop locations as general accommodation and working places for certain activities such as drop/ pick up and visiting family/ friends.

With a limited number of stops in trip chains, the distance matrix between control points and POIs of missing locations, which is the most time-consuming part of the whole process, was precomputed. as a distances matrix $D \in \mathbb{R}^{|C| \times |K|}$, where $|C| = 651$ represents the number of control points and $|K| = 2510$ denotes the number of POIs using the Dijkstra algorithm (Schrijver, 2012). This pre-processing approach simplified the subsequent density function calculations, as the optimization procedure could be efficiently conducted by filtering the respective origin set I and destination set J with the corresponding POI set and selecting the optimized location sequence. With computing the density function of each possible origin - midpoint - destination sequence for each trip, the one with highest density value was chosen as the approximated locations for the trip chain.

4 Results and discussion

The distribution of all density values of optimal estimations are shown in Figure 1, with a mean probability of 0.749 and a median of 0.906. To show an example of how the possibilities of potential mid-points are calculated, we choose one trip chain with work place and accommodation as the origin (blue star) and destination (triangle), respectively. The agent stopped at the missing midpoint for leisure activities. The estimation yielded density values as shown in Figure 2, of which the highest density values near 1 appeared to be the approximated locations. Figure 2 also shows that the density values of all points in the activity category of the chosen trip are computed, and most possible locations are clustered.

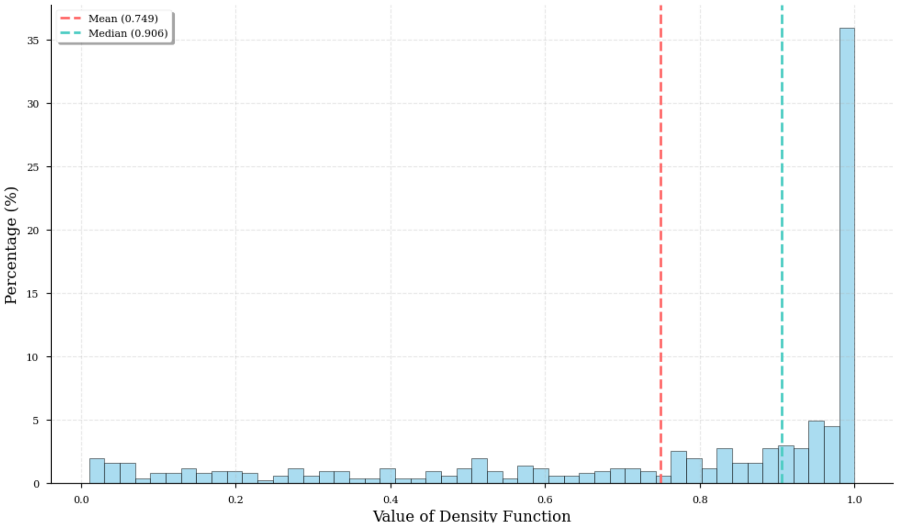

Figure 1: Distribution of Density Estimations of optimal PoIs

Computational Efficiency

This study integrated POIs as potential locations in the missing location imputation process, offering several methodological advantages. Primarily, this approach effectively constrained the number of known-unknown location pairs to a limited set, instead of scanning every segment in the network, thereby optimizing computational efficiency compared to network-based TGDE (Horner et al., 2012). The implementation of network-based distance calculations used the Dijkstra algorithm, which constitutes the majority of computational overhead. For our case, there were 12,009 edges in the Eskilstuna network, which means 12,009 times of routing distance calculation for each trip using network-based TGDE. However, the number of calculations could be substantially reduced to 21% with the limited set of POIs. The pre-computation of a comprehensive origin-destination distance matrix furthermore significantly simplified the location imputation process by eliminating redundant calculations for individual trips. With one time calculation of the route distances, we only needed to compare potential o-d pairs to select the optimized stop location for each trip.

Semantic Enhancement and Spatial Filtering

Beyond computational efficiency, our method advances upon previous time-geography applications for destination estimation through the novel integration of semantic information derived from POIs and travel survey data (Wang et al., 2024). This semantic enhancement was implemented through a sophisticated filtering mechanism that incorporated trip purposes prior to computing presence probabilities based on the agents' spatial coverage capabilities. While traditional TGDE methods scan all network sections and typically produce high probability values (approaching 1) for optimal estimates, they face challenges in discriminating between multiple high-value zones. Our approach addresses this limitation through a more refined methodology that has an advantage over conventional TGDE methods (Horner et al., 2012; J. Downs, 2014) that rely on intensity thresholds. The implementation of targeted semantic filtering substantially reduces the potential search space, resulting in enhanced precision in destination estimation.

Spatial Uncertainty Reduction

For origin and destination points, we introduced a novel method to minimize spatial uncertainty by constraining locations from reported areas to specific residential locations. This approach was particularly effective in city outskirts, where residential locations were spatially dispersed. The method utilized travel times to specific POIs as a mechanism for more accurate approximation of actual residential locations.

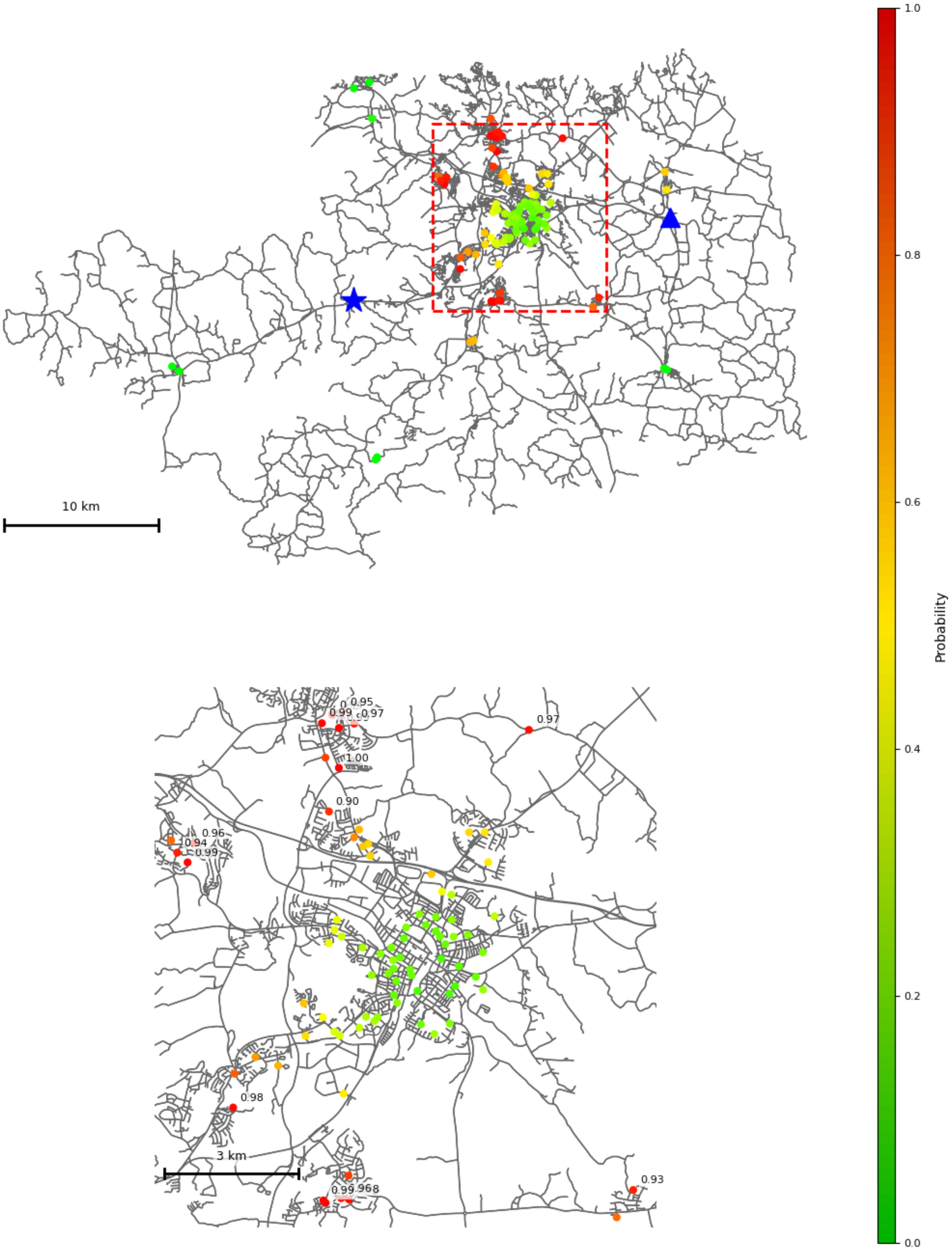

Figure 2: Example of Mid Spot Calculation of One Trip

5 Conclusions

This study presents a computationally efficient POI-based time-geography framework for estimating intermediate stop locations in trip chains, introducing a comprehensive method that synthesizes both spatial constraints and activity types with travel survey data. While it utilizes POI data from Eskilstuna municipality, the methodology's applicability extends beyond local contexts through the availability of global POI databases. Services such as Google Maps API could potentially provide comprehensive POI datasets, enabling its implementation across different geographical contexts.

This study also reveals several promising avenues for future research to enhance the prediction accuracy of intermediate stop locations. Socio-demographic characteristics, such as household composition and individual attributes, could better constrain the set of potential intermediate destinations. For example, the POI selection process could be refined by excluding educational facilities (e.g., kindergartens and preschools) for households without children. Additionally, the consideration of age and gender-specific variations in walking and cycling speeds could lead to more precise travel time estimations for non-motorized modes, as these factors are known to substantially influence individual mobility patterns. Another potential development is the implementation of spatially differentiated speed limits for travel time calculations. By distinguishing between inner-city areas, with lower speed limits and higher congestion, and city outskirts, where higher speeds are typically permissible, the method could obtain more realistic travel time estimations and enhance prediction accuracy.

In short, the presented methodology shows potential to achieve more complete representation and analysis of trip chains in real-world travel behavior studies, with broad applicability through its compatibility with widely available POI data sources. By linking trip chaining processes to geographic information, the study contributes to enable the usage of real-life travel survey data and the inclusion of time-space prism in trip chaining. It thereby enriches understanding of the impact of the built environment on travel behaviors and provides material for transport modeling and planning strategies.

Acknowledgements

References

Budhwani, A., Lin, T., Feng, D., & Bachmann, C. (2023). Assessing and comparing data imputation techniques for item nonresponse in household travel surveys. *Transportation research record*, *2677*(1), 1404–1417.

Chen, J., Shaw, S.-L., Yu, H., Lu, F., Chai, Y., & Jia, Q. (2011). Exploratory data analysis of activity diary data: a space–time gis approach. *Journal of Transport Geography*, *19*(3), 394–404.

Downs, J. (2014, 07). Integrating people and place: A density-based measure for assessing accessibility to opportunities. *journal of transportation and land use*, *7*, 23-40. doi: 10.5198/jtlu.v7i2.417

Downs, J. A. (2010). Time-geographic density estimation for moving point objects. In *Geographic information science: 6th international conference, giscience 2010, zurich, switzerland, september 14-17, 2010. proceedings 6* (pp. 16–26).

Downs, J. A., & Horner, M. W. (2012). Probabilistic potential path trees for visualizing and analyzing vehicle tracking data. *Journal of Transport Geography*, *23*, 72–80.

Downs, J. A., & Horner, M. W. (2014). Adaptive-velocity time-geographic density estimation for mapping the potential and probable locations of mobile objects. *Environment and Planning B: Planning and Design*, *41*(6), 1006–1021.

Ellegård, K., & Svedin, U. (2012). Torsten hägerstrand's time-geography as the cradle of the activity approach in transport geography. *Journal of Transport Geography*, *23*, 17–25.

Ellegård, K. (1999). A time-geographical approach to the study of everyday life of individuals - a challenge of complexity. *GeoJournal*, *48*(3), 167–175. Retrieved 2025-01-17, from http://www.jstor.org/stable/41147368

Hägerstrand, T. (1970). What about people in regional science? *Papers of the Regional Science Association*, *24*, 6–21. doi: 10.1007/BF01936872

Horner, M. W., Zook, B., & Downs, J. A. (2012). Where were you? development of a time-geographic approach for activity destination re-construction. *Computers, Environment and Urban Systems*, *36*(6), 488–499.

Kitamura, R. (1984). Incorporating trip chaining into analysis of destination choice. *Transportation Research Part B-methodological*, *18*, 67-81. Retrieved from https://api.semanticscholar.org/CorpusID:154045222

Marchal, P., Madre, J.-L., & Yuan, S. (2011). Postprocessing procedures for person-based global positioning system data collected in the french national travel survey 2007–2008. *Transportation research record*, *2246*(1), 47–54.

Roux, S., & Armoogum, J. (2011). Calibration strategies to correct nonresponse in a national travel survey. *Transportation research record*, *2246*(1), 1–7.

Schrijver, A. (2012). On the history of the shortest path problem. *Documenta Mathematica*, *17*(1), 155–167.

Shaw, S.-L., Yu, H., & Bombom, L. S. (2008). A space-time gis approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS*, *12*(4), 425–441.

Statistiska Centralbyrån. (2024). *Nyckelkodsystemet NYKO.* Retrieved from https://www.scb.se/vara-tjanster/bestall-data-och-statistik/regionala-statistikprodukter/statistikpaket/nyckelkodsystemet-nyko/ (Accessed: 2024-01-15)

Stopher, P., & Greaves, S. (2010). Missing and inaccurate information from travel surveys: pilot results.

Sun, X., Wilmot, C. G., & Kasturi, T. (1998). Household travel, household characteristics, and land use: An empirical study from the 1994 portland activity-based travel survey [Article]. *Transportation Research Record*(1617), 10 – 17. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-0032155151&doi=10.3141%2f1617-02&partnerID=40&md5=78ac8260a0ec2040d5668fbc77c5c89b (Cited by: 64) doi: 10.3141/1617-02

Tajaddini, A., Rose, G., Kockelman, K. M., & Vu, H. L. (2020). Recent progress in activity-based travel demand modeling: rising data and applicability. *Models and Technologies for Smart, Sustainable and Safe Transportation Systems*.

Wang, A.-S., Yin, Z.-C., & Ying, S. (2024). Probabilistic time geographic modeling method considering poi semantics. *ISPRS International Journal of Geo-Information*, *13*(1). Retrieved from https://www.mdpi.com/2220-9964/13/1/22 doi: 10.3390/ijgi13010022

Wood, B. S., & Horner, M. W. (2018). Aggregating mobile object trajectories: cumulative time geographic density estimation for gps data. *Transportation Planning and Technology*, *41*, 600 - 616. Retrieved from https://api.semanticscholar.org/CorpusID:115513535

Xu, M., Taylor, M., & Page, A. (2002). *A comparison of two methods for imputing missing income from household travel survey data* (Unpublished doctoral dissertation). Bureau of Transport and Regional Economics.

Yamada, I., & Thill, J.-C. (2004). Comparison of planar and network k-functions in traffic accident analysis. *Journal of Transport Geography*, *12*(2), 149–158.

Yin, Z., Huang, K., Ying, S., Huang, W., & Kang, Z. (2022). Modeling of time geographical kernel density function under network constraints. *ISPRS International Journal of Geo-Information*, *11*(3). Retrieved from https://www.mdpi.com/2220-9964/11/3/184 doi: 10.3390/ijgi11030184

Yu, H. (2007). Visualizing and analyzing activities in an integrated space-time environment: Temporal geographic information system design and implementation. *Transportation research record*, *2024*(1), 54–62.