Preface. Shifting data requirements in activity-based modeling of travel behavior - is mobile technology the panacea?

Citation for published version (APA):

Rasouli, S., & Timmermans, H. (2014). Preface. Shifting data requirements in activity-based modeling of travel behavior - is mobile technology the panacea? In S. Rasouli, & H. J. P. Timmermans (Eds.), *Mobile Technologies for Activity-Travel Data Collection and Analysis* (pp. xix-xxvi). (Advances in Data Mining and Database Management (ADMDM) Book Series). IGI Global. <https://doi.org/10.4018/978-1-4666-6170-7>

DOI:

[10.4018/978-1-4666-6170-7](https://doi.org/10.4018/978-1-4666-6170-7)

Document status and date:

Published: 30/06/2014

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Preface

Shifting Data Requirements in Activity-Based Modeling of Travel Behavior – Is Mobile Technology the Panacea?

A model can never be better than the data upon which it is based. In the context of travel behavior analysis, this commonly used statement emphasizes the critical importance of reliable and valid data to formulate, estimate, and apply models of travel demand. The ramifications of this statement become even more apparent when it is realized that the level of sophistication and complexity of activity-based models of travel demand has systematically increased across the last two decades (Rasouli & Timmermans, 2014). Whereas trip frequency and trip profile data of individual travelers were sufficient for developing four-step models, household-level activity-travel diary data were deemed necessary to estimate activity-based models of travel demand. The recent shift from cross-sectional to dynamic activity-based models increased data requirements in the sense that activity-travel diary data are required for a longer period of time. If one wishes to capture habits and repetitive patterns, data for several weeks are necessary.

Thus, to satisfy the data demand of advanced dynamic activity-based models, respondents would need to complete activity-travel diaries for many weeks. Completing such diaries is mentally very demanding and time consuming. Hence, it is not surprising that such diary panel data sets are very rare. Moreover, the few examples that exist only involve a small sample. The shifting requirements of advanced activity-based model not only require the completion of more detailed questionnaires but also these need to be completed for a much longer period of time than the usual single-day trip surveys. These shifting requirements negatively affect both the response rates and the reliability of the results. In turn, although lower response rates do not necessarily imply less valid results as long as the non-response is not systematically related with the activity-travel diary data, it is very likely that higher non-response rates may adversely affect the validity of the results. For example, due to higher response demand, individuals who are involved in more activities and trips may be less willing to participate. Similarly, to the extent that the reliability of the responses is a function of the time needed to complete the diary, more demanding surveys may negatively affect the reliability of the collected data.

Modern technology, such as GPS and smart phones, became available more than a decade ago. In the beginning, the devices were lumpy and difficult to carry. Over the years, however, the devices became increasingly smaller, improving their portability. The newest generation of smart phones is GPS-enabled and programmable, implying that the recording and uploading of travel data can be automated, reducing respondent burden. Thus, it seems that as long as respondents are willing to carry a mobile device and participate in the survey, we are reaching the point that their activity-travel diary data can be captured automatically and passively. At least, most of the activity-travel data can be collected semi-automatically, and respondents may only need to verify the data and provide additional information that cannot be derived from the instruments.

Although the geo-temporal information provided by modern technological devices is certainly supreme compared to human memory retrieval, activity-travel diary data derived from GPS traces or GSM data is also not error-free. First, some respondents may be technology illiterate and therefore may either refuse to participate in GPS-based studies or face difficulties in activating and/or uploading the data stream. Moreover, because of the accuracy, some people may object being traced on a continuous basis. Thus, like conventional travel surveys, GPS-based data collection efforts may involve sampling bias and non-response. The kind of bias and non-response likely differs from those reported for conventional travel surveys. These potential problems should be studied and remedied to the extent possible. Second, technical problems or ill-functioning devices may cause further non-response. Urban canyons may lead to signal loss or inaccurate recordings. Cold starts may imply that data are not recorded for the start of a day or trip. Limited battery life may impose constraints on the number of hours that geocodes can be recorded. Third, GPS devices basically provide geo-coded information. This information needs to be processed and converted into daily activity-travel diaries. For time and route, this process is relatively easy. However, inducing transport mode and, in particular, activity types from GPS traces is a much more challenging task. Early applications were largely based on visual inspection and simple interpretation rules. The field has recently, however, explored the use of considerably more advanced algorithms and data fusion options.

In providing a more detailed answer to the question whether modern mobile technology is the panacea for the rapidly increasing data requirements of dynamic activity-based models of travel demand, this volume brings together the work of many leading scholars from across the world. They report their latest experiences with the application of such technology and discuss their contributions to redefine the state-of-the-art and state-of-practice in this field of research and application.

The book is divided into three sections. The first section contains a set of chapters reporting experiences with the use of GPS devices and smartphones for collecting activity-travel data. The second section includes a series of chapters on methodological advances in improving our ability to impute more valid and accurate activity-travel diary data from GPS traces. The final section provides examples of the use of GPS data in transportation planning.

In their kick-off chapter to this book, Bricka, Baker, Simek, and Wood provide an overview of recent trends in the development and application of new technology to collect activity-travel data in an attempt to reduce the need to conduct intercept, in-person and mail-based, surveys. The authors both discuss self-reported data and passively or actively observed data. Historical developments with a strong emphasis on the USA, pros and cons of alternative approaches, and examples of the various approaches are described. The authors argue that technologies that generate observed travel data have the potential to revolutionize travel data collection methods.

In the next chapter, Asakura, Hato, and Maruyama discuss similar material, but now from the perspective of historical developments in Japan. In particular, they focus their attention on Probe Person (PP) survey systems, which have found ample application since the early 2000s and use GPS-assisted mobile phones connected to Internet Web diaries. Characteristics and examples of these systems are discussed. In addition, the chapter discusses a case study of a Smartphone-based PP survey system in Kumamoto, Japan. Advantages and remaining issues are discussed.

Reinau, Harder and Overgård develop a model for how to design and conduct a GPS tracking study. The model is based on the so-called V-model, a model that provides guidelines for the development of software and is well known in the field of software development. In analogue, they use their V-model

to describe the different phases of GPS data collection and discuss the different choices that have to be made during each phase. A case study illustrates an application of the model. This chapter is exemplary in presenting guidelines for the design and execution of GPS data collection. Results of prior literature reviews are reflected in the V-model and in discussions about choice of technology, recruitment, analysis of data, and other operational decisions.

Auld and Mohammadian describe a typical application of their UTRACS survey system. The system includes a demographic information questionnaire for the household, a routine activity-travel questionnaire, a periodic scheduling process survey, and a prompted-recall activity diary survey. Respondents carry a GPS device for a certain period of time. Once the device is connected to the computer, participants are invited to upload the GPS traces. These data streams are then first cleaned and processed. Activity episodes are identified on the basis of temporal and spatial thresholds. That is, these thresholds are used to decide whether the GPS traces are observed in a sufficiently compact area for a sufficient length of time to represent an activity episode. Finally, respondents are invited to check and if necessary correct the imputed activity-travel diary, and to provide additional information about the attributes of the schedules that could not be extracted from the GPS traces. Experiences with an application of the system in 2009 among 112 respondents in the Chicago metropolitan area are described. The authors report evidence of a (limited) reduction in respondent burden.

In the next chapter, we shift the focus of our attention to China. Chai, Chen, Liu, Tana, and Ma discuss collecting activity-travel diary data using GPS for smart travel planning in China. The sample consists of 709 respondents from 23 communities and 19 companies in Shangdi-Qinghe, a typical suburban new town in the northwest of Beijing. Respondents were tracked for 7 days using GPS devices. Trajectories were then visualized. A prompted recall survey instrument was used to collect the details of the trips. Thus, their study did not involve an attempt to semi-automatically impute activity-travel diaries from the GPS traces as in other studies.

Jin, Asgari, and Hossan contribute to this book by reporting their experiences with another application of GPS technology in the context of the Regional Household Travel Survey conducted by the New York Metropolitan Transportation Council between the Fall of 2010 and the Fall of 2011. The final completed data set consists of approximately 18,965 households, of which about 10% (1,930 households) participated in the GPS prompted-recall method instead of the diary-based approach. Their focus is on trip misreporting. In line with previous research, they find that, in general, the GPS-based data collection method captured more tours than the diary method. Schoolchildren of a driving age and non-working adults have the highest underreporting rate. An ordered probit model was estimated to predict tour frequency choices. The model includes a sample-indicator variable that accounts for the misreporting behavior. The estimation results indicate significant interaction effects with the GPS sample indicator variable. Driver license status, race, person type, household type, household income, and number of household vehicles showed to have significant effects on misreporting behavior.

In the following chapter, Feng and Timmermans describe the experiences and findings of one of the more ambitious projects described in this book. The study involved a one-year project, consisting of four waves of three months each. During each three-month period, a sample of respondents, located in two different regions in The Netherlands, carried a dedicated GPS device and uploaded their data to a Web server. After uploading, GPS traces were processed online to minimize respondent burden and allow respondents to immediately verify the imputed activity-travel diaries. The authors provide information

on issues related to survey design, quality control, overloading, and cross-processing, which are pertinent in large-scale GPS data collection projects. This study provides evidence that GPS data collection for longer than one week is feasible.

The applicability of modern technology for collecting activity-travel diary data ultimately depends on how successful we can correctly identify transport modes and activity types. Much progress has been made recently in developing powerful algorithms. While initial attempts were based on subjective, ad hoc rules, soon these approaches were complemented by statistically more advanced approaches, such as fuzzy sets, support vector machines, and Bayesian belief networks. In a comparative study, Feng and Timmermans (2013) found that decision tables and Bayesian belief networks outperformed several other algorithms in terms of predictive success.

Reumers, Liu, Janssens, and Wets report the result of a similar study, in which they compare state-of-the-art machine-learning methods (Multiclass Support Vector Machines, Multinomial Logistic Regression Decision Trees, and Random Forests) to classify activity type based on activity duration and activity start time. The data are based on a paper-and-pencil activity-travel diary survey and a related survey in which GPS-enabled PDAs were used. The data were collected for one week in Flanders, Belgium, in 2006 and 2007, among 2,500 households. Their results add further evidence to earlier research findings that non-linear models outperform linear models. Moreover, some algorithms have the advantage that the nature of the relationships in the data can be learned from the data.

Whereas the previous chapter considers machine-learning algorithms for detecting transportation modes, Kohla, Gerike, Hössinger, Meschik, Sammer, and Unbehaun use two multinomial logistic regression models, based on nine features derived from GPS and acceleration data. First, they estimate a regression model to detect walking trips and stages. Next, a second multinomial logistic regression model is developed to impute eight transportation modes for the remaining trip stages. Overall, the algorithm achieved a correct detection rate of 79 percent. Better results were achieved for motorcycle and moped, railway, bicycle, and walking, while lower accuracy was obtained for urban public transport. An outlook on future elaborations of the suggested detection approach is provided.

Information about the transportation network is needed to derive exact information about travelled routes. In addition, route information may be critical to successfully identify transportation modes. Blazquez and Miranda propose a real-time topological rule-based methodology that detects and solves the map-matching problem. An application to a real case scenario in Chile illustrates the success of the proposed algorithm to resolve spatial mismatches. Simulations indicate that the performance of the algorithm is sensitive to variations in buffer size. If lower spatial data quality is used, the algorithm requires larger buffer sizes to obtain the best results. Increasing the number of tested data points improves solving the map-matching problem but also increases computation times.

The majority of studies on travel data collection using modern technology is based on GPS devices. Most smartphones are currently endowed with GPS and sometimes with other sensors as well. Hence, as discussed by Asakura et al., potentially, smartphones have some advantages over stand-alone GPS devices. In that sense, it is not surprising that interest in the use of smartphones for collecting activity travel data has significantly increased lately. Ferrer and Ruiz discuss the development of an app for Android-based smartphones and report experiences of a pilot test among a small sample. In addition, they discuss the development of a neural network model to classify transport modes from the GPS traces and accelerometer data. Results indicate that the neural network is reasonably successful in identifying transportation modes.

Kazagli, Chen, and Bierlaire also discuss the applicability of smartphones. Arguments of previous chapters are elaborated. The authors provide an overview of the potential advantages and limitations of smartphones for travel data collection and mobility analysis and choice modeling. In particular, they introduce the Lausanne Data Collection Campaign that resulted in one of the biggest smartphone datasets in the world. Next, they systematically discuss the pros and cons of smartphones, leading to a set of challenges they were confronted with in their application. They provide a useful overview of state-of-the-art methodologies and some frontier developments that are needed to solve the dominant problems. Specific attention is given to the exploitation of smartphone data in route choice modeling in the context of discrete choice analysis, which indeed is a domain of application where mobile technologies offer clear advantages.

Yet another pilot study on the applicability of smartphones in collecting activity-travel behavior data, this time conducted in China, is documented by Jianchuan, Zhicai, Guangnian, and Xuemei. The pilot study was administered in Shanghai, China and involved 32 respondents. The mode identification algorithm first determines whether the trip involves walking, bicycle, or motorized transportation modes using neural networks. Next, metro is detected based on long-time data loss and relatively high average speed. The identification of trip purpose is based on the residential address and the addresses of each person's workplace and school. The percentage of correctly classified transportation modes was between 86.5 and 99 percent. Identification of trip purpose is lower, with an average of around 80 percent correct predictions.

Phone data, albeit aggregate in nature, provide another relevant source of information to examine mobility patterns. Saluveer and Ahas report their experiences in the context of perhaps the richest data set described in this volume. Their data was collected across many years in Estonia. It consists of call detail records of mobile phone providers. Empirical examples illustrate the richness of such data and the kind of analysis for which it can be used. However, the authors also discuss the methodological challenges in making the basic data ready for detailed analysis, and the limitations that are encountered in this process.

Location-based social networking is another highly interesting emerging technology to collect activity travel data. It refers to dedicated location-based social networking services that rely on GPS to locate users, and allow its members to broadcast their locations and activities through their mobile devices. Yang, Jin, Cebelak, Ran, and Walton contribute to this book by providing a very interesting chapter on the usefulness of such data to create dynamic travel demand data of high spatial and temporal resolution. They illustrate the feasibility of using these data to estimate dynamic Origin-Destination travel demand matrices for general trips. A combination of non-parametric cluster and regression analysis is used to find the functional relationship between the specific data and trip production and attraction. Next, a modified gravity model is proposed to estimate the original-destination matrix. Results look very promising.

Arellana, Ortúzar, Rizzi, and Zuñiga report an interesting and important application of information technology. They discuss how in-vehicle GPS technology can be used to collect large amounts of data to accurately measure level of service of public transport. In particular, the authors suggest a procedure to derive relevant level-of-service performance indicators, such as waiting and travel times and their variability, at any level of spatial and temporal aggregation, for a dense bus network in Santiago de Chile. Their procedure is based on easy-to-use, freely available software. An illustration of the proposed procedure is provided.

Concas, Barbeau, Winters, and Georggi describe another highly interesting and timely application of modern technology in transportation demand management. Their chapter discusses the application of a GPS-enabled mobile phone system to collect travel behavior data. The context is differences in hourly rental rates for car-sharing users as part of the University of South Florida (Tampa, Florida) car-sharing program. The application easily allows investigating changes in activity-travel behavior (in particular peak-hour travel) in response to variable pricing. A model is developed to estimate rental start-time probability density functions. Findings suggest that the system can be used to influence the timing of trips.

Hanson and Hildebrand document another relevant study on the use of GPS devices. They used vehicle-instrumented passive GPS units and a prompted recall with Geographic Information Systems to collect travel diary data. The latter instrument was used to identify activity purpose. The percentage correct identification was high, but the authors qualify this finding by arguing that identification might be easier in rural areas. The aim of their study was to better understand driver travel behaviour of a convenience sample of 60 older rural drivers in New Brunswick, Canada. Data collection involved, on average, 5.3 days. The generated data were used to analyze the impact of license restrictions on older drivers, the opportunities of these drivers to satisfy their needs without a car, and exposure analysis by road class.

While most studies using modern technology for data collection have focused on cars or multiple transportation modes, Venter, Minora, Shukrani, and du Toit summarize a project on the use of GPS in a multi-method approach to explore environmental factors affecting walking patterns in South Africa. In common with most studies, they first derive quantitative measures of walking activity from multiday GPS traces of a sample of 174 respondents in three case study areas in Pretoria, South Africa. The results indicate the amount of walking that takes place across a range of neighbourhood types and the routes that are involved. Next, these quantitative findings were further interpreted using open-ended qualitative research methods to obtain richer insights into the motivations behind the observed behavioural patterns. In particular, the authors were interested in examining the extent to which these motivations are related to the built environment. Personal security and fear of crime turned out to be critically important motivators.

Transportation management initiatives have become increasingly important and complementary to infrastructure investments in mediating negative impact of transportation on the environment. Meloni and Sanjust describe a fascinating case study in Cagliari, Italy aimed at reducing car use. The study uses a personal active logger for the collection of individual activity-travel patterns before and after the implementation of the program. In particular, in 9 different waves, 109 individuals in the Cagliari Metropolitan area were invited to record their activity-travel patterns before and after the submission of personalized travel plans. One-week activity-travel diaries were collected to identify any travel behaviour changes. An app that was installed on a smartphone with a built-in GPS was used to track a participant's daily routes. A sequence of pull-down menus was used to allow participants to provide details of their activity-travel diaries. Two versions were used: one real-time and one in which the activities were reported by phone at the end of the day. Results indicated that despite the greater effort involved in real-time compilation the information collected by the active logger was more accurate and also generated more behavioural change.

In the final chapter of the book, Zhong Zheng and Suhong Zhou describe an interesting case study using taxi GPS data. It is an interesting example in the sense that data that are available anywhere are used for research purposes. A disadvantage of taxi data is that they cover only a subset of all travel that may not be representative of overall patterns. However, keeping this in mind, the analysis of taxi data offers rich opportunities to study daily rhythms in travel patterns. In their chapter, the authors describe the spatial and temporal distributions on one-day taxi trips in Guangzhou City, China.

Keeping these studies in mind, the question is whether this modern technology is the panacea for addressing typical problems in conventional travel surveys. The rich experience reported in this volume suggests that the answer depends on the aim of the study. Nevertheless, some general conclusions can be drawn. First, although GPS-devices are not error free, their accuracy in recording route information and the timing of activities relative to classic travel surveys is undeniable. It implies that the new technology is ideal for collecting data about route choice and the dynamics in route choice over time. Smartphones tend to be less accurate than dedicated GPS devices and have the problem of battery drainage. However, if the purpose is route choice, the frequency of the recording does not need to be very high as map matching and other methods can be used to detect the route and complete missing links. The need for validation may also be less.

Second, in principle, smartphones have some clear advantages over dedicated GPS devices: communication can be automatically organised, it is easy to combine GPS traces with short questions, either to annotate the travel and its stages or to provide additional information. However, if transportation modes and/or activity types should be imputed to reduce respondent burden, it is critical to record GPS traces with a high frequency so that acceleration can be measured accurately. Consequently, batteries will drain fast at the current state of technology. This means that additional software to optimize battery use is needed. For example, it is possible to activate the GPS only if the individual moves. Alternatively, imputation should rely more on other sensors or an extra battery should be carried. Without such very sophisticated solutions, it is fair to say that smartphones do not provide an operational solution yet as an alternative for collecting activity-travel diary data. Dedicated GPS devices tend to provide more accurate data that can be recorded for a longer period of time.

Third, although some reduction in respondent burden can be achieved, one should not forget that getting knowledgeable about the technology and the time required to validate the data should be taken into account. Thus, reduction of respondent burden may only be significant if the survey is implemented over a longer period of time, and then only if the accuracy of the imputation is high so that the number of corrections that needs to be made and the amount of missing information that needs to be provided are relatively small.

Finally, although much progress in the semi-automatic imputation of activity-travel patterns has been accomplished, a fully automatic process is very difficult to achieve. A fully automatic process implies that valid activity-travel diaries can be derived from the GPS traces without respondents needing to verify the imputed diaries. Although with the current state of algorithms, detection of transportation modes is high, it is not perfect. The current detection of activity types is inherently more difficult. Prompted recall instruments are advisable, but it should be realised that respondents also make errors. In the end, therefore, some visual inspection of mapped GPS traces may be required.

Overall, then, the design of a technology-based data collection project should be dependent on the specific intended use of the data and the accuracy that is deemed necessary. Choice of technology and most operational decisions depend on subjective trade-offs between respondent burden, minimization of error in in-take questionnaire, imputation of activity-travel data, and prompted recall. Strict management and monitoring is essential, particularly in multi-week data collections.

This volume describes the state-of-the-art in a rapidly developing field of research and application. The next years will see many new applications of increasing size and duration. Algorithms to impute transportation modes have become very mature. In particular, learning-based algorithms that allow a

nonlinear relation between input data and the classification of interest should be chosen because evidence has accumulated that these algorithms outperform simpler methods. Once battery problems have been solved, smartphone apps are preferable in many studies, but it still may take some years before such longer-lasting batteries are available. In the meantime, clever hybrid approaches and coupling of technology should be developed and tried.

Soora Rasouli

Eindhoven University of Technology, The Netherlands

Harry Timmermans

Eindhoven University of Technology, The Netherlands

REFERENCES

Feng, T., Timmermans, H. J. P. (2013). Comparative evaluation of algorithms for GPS data imputation. In Proceedings of WCTR. Rio de Janeiro, Brazil: WCTR.

Rasouli, S., & Timmermans, H. J. P. (2014). Activity-based models of travel demand: Promises, progress and prospects. *International Journal of Urban Sciences*, 18, 31–60.