Field notes from the travel app frontier:

Build Me, Break Me, Impute Me

Danielle McCool

We are stuck with technology
when what we really want is just stuff that works.

Douglas Adams

ISBN:	978-94-6537-113-9
Printed by:	Ridderprint BV, www.ridderprint.nl
Cover by:	Sharkblood
Digital version:	available at https://dspace.library.uu.nl/.

Field notes from the travel app frontier:

Build Me, Break Me, Impute Me

Veldaantekeningen van het reis-app front:

Bouwen, Breken, Imputeren

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. ir. W. Hazeleger,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen op

vrijdag 30 januari 2026 des middags te 2.15 uur

door

Danielle Marie McCool

geboren op 8 april 1985
te Santa Fe, New Mexico, Verenigde Staten

Promotoren:

Prof. dr. J.G. Schouten

Prof. dr. P. Lugtig

Beoordelingscommissie:

Prof. dr. S. van Buuren

Prof. dr. ir. D.F. Ettema

Prof. dr. E.L. Hamaker

Prof. dr. F. Keusch

prof. dr. A.G. de Waal

Contents

Acronyms

CAWI Computer Assisted Web Interviewing

CS Candidate Specificity

Dist ↓ Distance underestimation

Dist ↑ Distance overestimation

DTW Dynamic Time Warping

DTWBI Dynamic Time Warping-Based Imputation

DTWBMI Dynamic Time Warping-Based Multiple Imputation

DTWBMI-HI Dynamic Time Warping-Based Multiple Imputation - High Information

DTWBMI-LO Dynamic Time Warping-Based Multiple Imputation - Low Information

GNSS Global Navigation Satellite Systems

GPS Global Positioning Satellite

kNN k-Nearest Neighbor

LI Linear Interpolation

MB Match Buffer

MI Mean Imputation

MICE Multiple Imputation by Chained Equations

N Imps Number of Imputations

NSIs National Statistical Institutes

ODiN Underway in the Netherlands

OSM Open Street Map

RMSE Root Mean Square Error

RoG Radius of Gyration

SBTS Smartphone-Based Travel Surveys

SN Statistics Netherlands

SSI Smart Survey Implementation

TD Travel Distance

TDS Travel Diary Study

TDTR Top-Down Time Ratio

TP Travel Period

TP ↓ Travel Period underestimation

TP ↑ Travel Period overestimation

TP Acc. Travel Period Accuracy

TS Total Stops

TW Time Window

TWI Time Window Imputation

Part I

An introduction to smart surveys and location data

1

Introduction

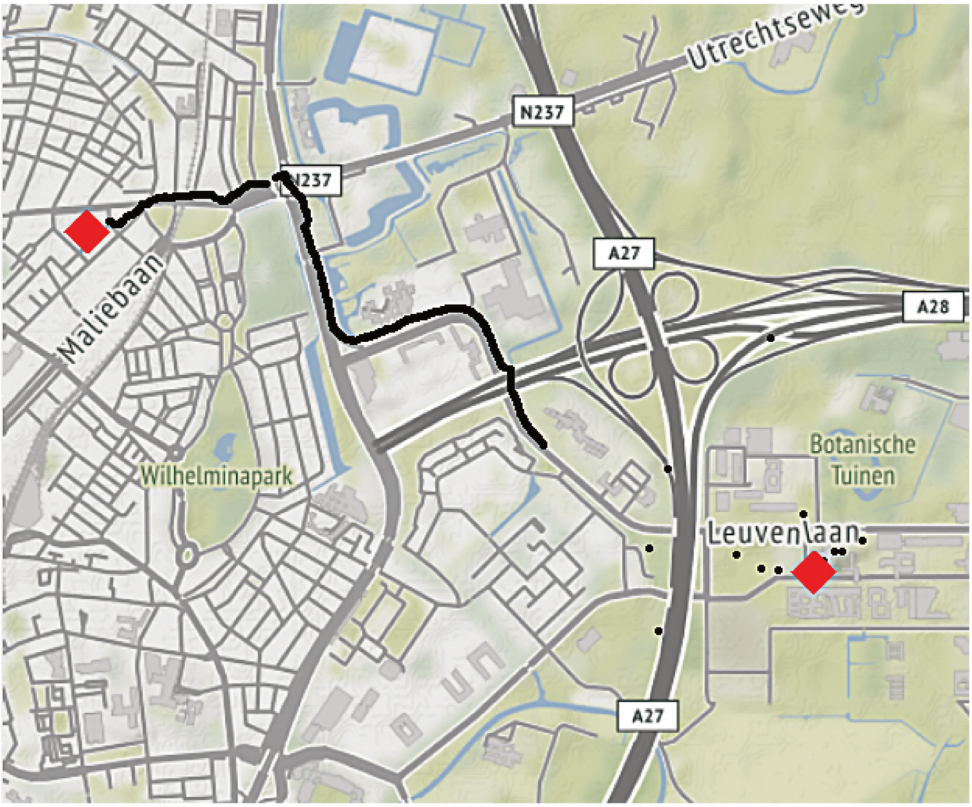


Figure 1.1 *Measuring mobility behavior with smartphone sensors is tricky due to the gaps in the data. This is one common example of missing data called a "cold start," where the boxes indicate my start and end locations. By the time the app started tracking my location, I was already halfway home.*

Despite over a decade of attempts at using smartphone-based applications for measuring national travel behavior, so far none have been robust enough to stick. The single most important factor isn't respondent willingness or lack of institutional investment: it's missing data.

Some of the missing data are the "fun" type, where you almost can't help but to see all the ways to correct it from the moment you look at it, because it's very intuitive. Like in Figure 1.1: drop a few clearly erroneously reported locations, use the roads and some common sense, and you've solved the problem. Supposing one opts out of addressing it, the consequences also seem like they would be relatively mild: maybe a slight underestimation of travel distance if we draw a line between where the gap started and where it ended, rounding off the corners, so to speak.

Unfortunately, the extent of the problem is much larger, and the consequences more dire. It was during the first Mobile Apps and Sensors in Surveys conference back in 2019 when the light first went on for me, that all of us working with passively-

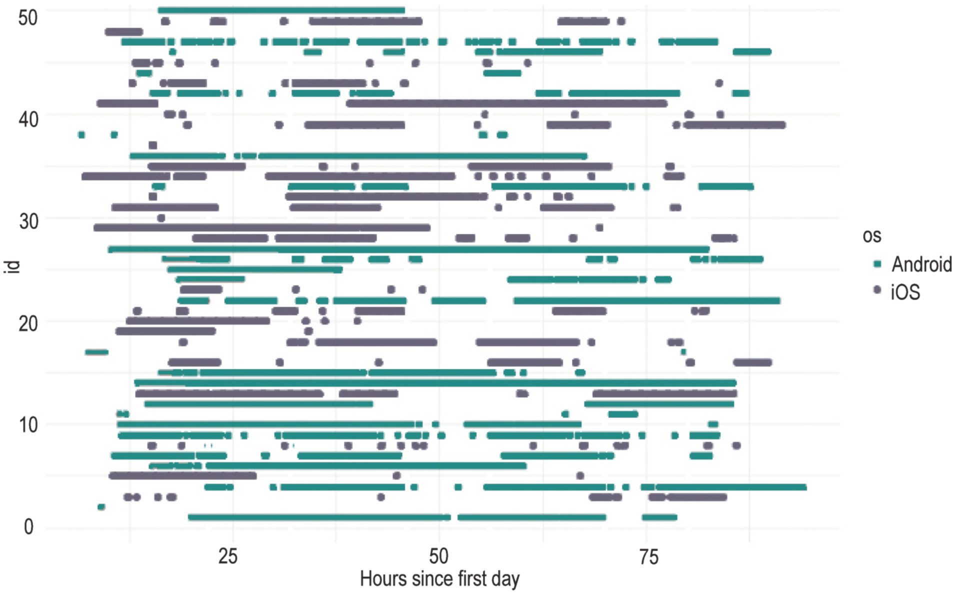


Figure 1.2 *Data (dis-)continuity for a random sample of respondents to the 2018 travel app study. Contiguous lines show times for which location measurements were successfully recorded. Breaks in these lines represent gaps of various lengths in the data from that respondent.*

acquired sensor data were dealing with the same thing. I floated the theory later that year at the European Transport Conference to a group of researchers specifically focused on measuring human mobility with apps, and got the same response. The missingness was pervasive, silent, and seemingly unavoidable when using respondents' own devices.

Rather than looking like Figure 1.1, the data that came in across all our studies looked much more like 1.2: more likely to be missing than to be present when considered in the context of a week-long field test.

It would be easy for a skeptic to look at that fact and suggest that *maybe* Smart Surveys aren't actually the future, but they'd be wrong. Nothing has done as much to clarify this fact for me as reading the reports from the sessions at the second International Conference on New Survey Methods in Transport, held in 1983 in Hungerford Hill, Australia (Ampt et al., 1985):

> Emerging from the conference discussions, six key issues were prominent; the gap between the state-of-the-art, as espoused by researchers, and the state-of-practice, as employed by practitioners; the differences in results obtained when employing different survey methods to collect essentially the same data; the emerging use of the telephone for the conduct of interview travel surveys; the interaction between survey methods and demand modelling and the disparate levels of sophistica-

> tion in the two areas; the role of microcomputers in the design, conduct, and analysis of transport surveys; and the need for practitioners to engage in controlled experimentation, so that methodological advances may continue to be made in the field of transport surveys. (p3)

As it stands now, travel diaries are a cornerstone of mobility research, of course. They provide detailed insights into individual travel behavior, which is crucial for urban planning, transportation policy, and infrastructure development. Long-running longitudinal panels such as the Dutch National Mobility Panel are currently used to calibrate countrywide transportation models – they are the gold standard to which the other auxiliary data are fit. Seen in this regard, the list of six key issues seems very remote.

And yet, here and now, forty-some-odd years in the future, Smart Survey researchers find ourselves asking questions which, if not word-for-word identical to those posed back in 1983, are alike in spirit. If you were to go back and address the researchers attending these sessions, informing them that the gap between state-of-the-art and state-of-the-practice had been so sufficiently addressed that the then-new travel survey methodologies had become the gold standard, I think they would be unsurprised. If you then (gently) let them know that it would take over a decade to iron out some of the more difficult wrinkles, just in time for web-administered versions to begin to supersede them, I think some might throw in the towel. Innovative research leans on people who believe that they are one paper away from solving The Big Problem.

Actually, if you asked the researcher working adjacent to a methodologist involved with innovative technologies, they would probably tell you that the methodologist seemed to have the opposite of the Midas touch: everything that they worked on was the equivalent of turning gold into lead. It wasn't like they went in with starry-eyed dreams that were never going to materialize — everything looked very plausible on paper, after all — but the results only ever generated more questions than answers.

The truth is that technological innovation is often a journey of decades. Groundbreaking technologies begin as ambitious experiments that look like very plausible opportunities to expand the future of some field, but that subsequently fail to meet initial expectations. Really, these early failures *aren't* failures. They serve the critical function of laying the foundations for the technology to make it into production in the future. The only failure is one of assumptions: that the street-ready date was next year versus ten years on. Frankly, you'd be forgiven for thinking that, because that's also how these things are sold, because we're none of us very patient, and we mistakenly believe ourselves that we're only one bug away from solving the whole thing. Take for example, ARPANET.

Though it would eventually lead to the thing that we call the internet, the first time that researchers connected the first four computers together over 2000 feet instead of the usual 50 feet, it all but collapsed in on itself from errors. Despite the fact that the packet-switching was clearly doing the thing it was intended to do, and quite well, it became obvious that long distance connections were simply going to be too noisy to enable them to rely on transmission without errors (Lam, 1974). The error checking they implemented as compensation led to revision after revision that eventually became the set of protocols we use today as the basis for the internet.

1.1. Smart surveys

Smart Surveys integrate traditional survey methods with data collected from sensors and apps on smartphones and other smart devices. This approach combines the strengths of self-reported data (e.g., being able to assess attitudes and motivations, or offer interpretations that require the respondent's involvement), with added objectivity and precision from sensor data (Struminskaya et al., 2020). The opportunities are clear: the data are richer, more finely grained, and more detailed; optimally, it should reduce measurement error by eliminating recall and social desirability bias; it's more efficient both for the user, who can let passive data collection take on their role as respondent, and for the researcher, who can distribute apps and collect data quicker and at reduced expense.

Smart surveys are at the frontier of survey methodology, and although we've now established protocols that allow them to work across 2000 feet, they're not yet equal in maturity to the networks they run on. But, like ARPANET's packet-switching, it's obvious that they do the core elements of what they were designed to do quite well, whether that's optical character recognition of text on a receipt (Schouten et al., 2025), providing suggestions for activities as users search (Fritz et al., 2025), or measuring someone's precise location (Gootzen et al., 2025). This makes sense, because these use smart *features* that have themselves had decades to reach maturity. This is what makes them valuable to incorporate as data collection tools.

It's also what makes them very complicated. Sensor data exists for its own purposes, and researchers seeking to harness the data for their own purposes have found that they are putting in a lot of legwork to align the resulting data with their own aims. The findings from the second ESSnet-funded project on the topic, Smart Survey Implementation, noted a "Smart versus Clever" divide between the features as designed and their capacity to improve rather than complicate the process for respondents (Fritz et al., 2025). By one name or another, all these surveys are encountering the same challenge: our capacity to make use of the new features is lagging behind our capacity to put them into production. Put this way, it becomes clear that this is an opportunity for innovation.

Back in 2018, Statistics Netherlands embarked on its journey with smart surveys with an app we called "Tabi[1]." The Tabi app, which is detailed thoroughly in Chapter 2, was developed as part of a student's Bachelor's thesis and designed to be installed on a respondent's phone to capture geolocation data. This data was fed back to users, who were asked to provide annotations of the things that a traditional diary

[1] After the Japanese sandal

survey would have asked them, like the mode of transportation used in the trip, or their purpose for undertaking it. For an auteur effort, it worked surprisingly well and, although the challenges were evident when trying to make the New Data do all the things the Old Data did inherently, the richness of the New Data also spoke for itself (see Figure 1.3).

1.2. An improvement over traditional methods

For some, the existence of a new technology alone is sufficient reason to justify its usage, while others require some deficit with the current approach. Conveniently, Travel Diary Studys (TDSs) face several challenges that hinder their effectiveness.

Participants often fail to report all their trips, leading to underreporting and incomplete data (Richardson et al., 1995; Sammer et al., 2018). The opportunity to record a respondent's location independently of their capacity to be both invested in the survey and also accurately reproduce their daily activities can allow us to identify missing data that we would otherwise be invisible to researchers.

Additionally, participants have become increasingly reluctant to participate in surveys overall, leading to reduced response rates.(Stedman et al., 2019). When the survey is *also* burdensome, as travel diaries tend to become (Richardson et al., 1995), the combined non-response and break-off results in particularly low response rates, which are continuing to decrease. In order to continue to generate sufficiently large sample sizes, the net must be cast still wider, which gets expensive.

Smart surveys have the potential to revolutionize mobility studies by providing more accurate and comprehensive data on travel behavior. They can capture detailed information on travel patterns, modes of transportation, and trip purposes, which is essential for developing effective transportation policies and infrastructure.

Smart surveys offer significant advantages over traditional diary surveys in measuring human mobility. The challenge is in harnessing those advantages despite the obstacles. In the next section, we describe the specific problems that smart surveys and travel apps can encounter, setting the stage for what this thesis aims to accomplish.

1.3. Current issues with smart surveys and travel apps

This introduction opened with the claim that the primary issue holding Smart Surveys back from wider usage was the problem of missing data. This is true, but it describes a symptom rather than a cause. Is it possible to identify a single underlying cause?

Looking at the problem holistically, one might say that the underlying cause is due to a misunderstanding of what we are looking for in complete data. This follows from

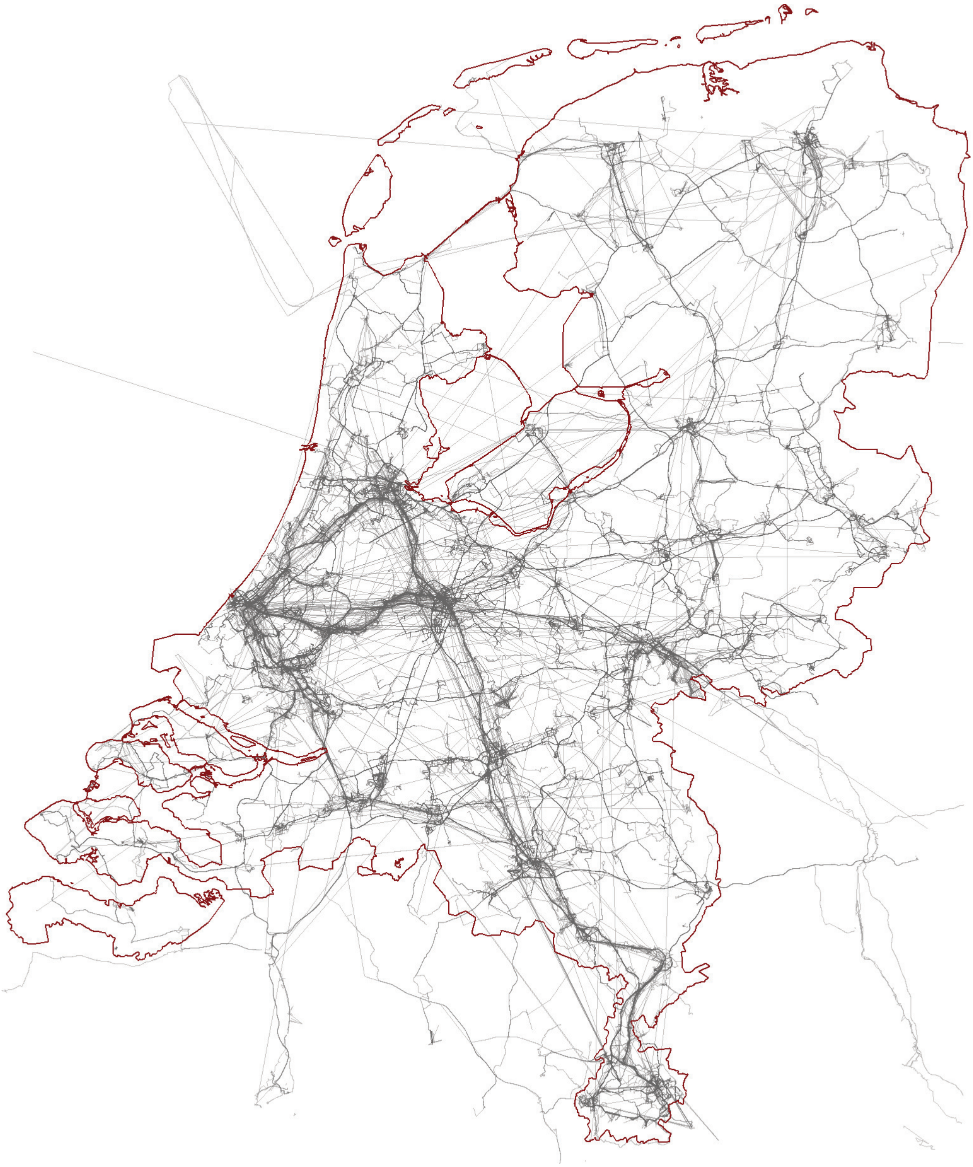


Figure 1.3 *All trajectories from the 2018 Tabi app*

the fact that the data captured from device sensors are rarely directly comparable to the data we are used to seeing. There is a fundamental incompatibility for near-continuous data to be complete at the level that we expect, because we are only ever sampling from it.

In the case of the travel survey, the raw data are functionally incomprehensible barring a comprehensive geographic knowledge of lat/lon pairs. To derive meaningful insights with the data, for example producing estimates of how far the average person travels in a given day, researchers must apply a whole host of algorithms in the interpretation of the raw data just to arrive at the starting point for a traditional survey. This necessity dramatically increases the complexity and requirements for data analysis.

Traditional diaries are themselves not inherently algorithm-free, of course. Given a list of all the trips on a given day for a particular respondent, you would at the very least have to sum them to arrive at a total, followed by consideration of selection bias before calculating a mean across persons. This process, too, has become gradually more complex over time all with the goal of improving our estimates. For example, if a respondent said they took the train between Station A to Station B, we could replace their record of how many kilometers they traveled with our own cross-referenced from a list of train station distances, which is something that the Dutch Travel Survey Underway in the Netherlands (ODiN) has done since 2018 (CBS-CvB, 2018). Assuming the respondent traveled as expected, this would be near enough perfect accuracy, whereas a respondent's own estimate would likely be quite poor as they were not themselves driving the train.

While algorithms have always played a role in deriving these estimates – that's why statistics exists as a field – the fact that the data that we take in is in such a different form than the data we would like to get out means that we've dramatically increased the requirements here.

Making all of it much more difficult is the fact that not all algorithms involved in a study are under the researcher's control. The device itself plays a massive role in data collection, with each device's operating system dictating how and when a location can be acquired, or how long an app will be allowed to stay open if it's determined not to be actively in use. The combination of these two highly-variable factors creates a complex situation for missing data. Complicating this further is the fact that many of the algorithms applied to the raw data, for example, segmentation of the geolocations into track/stop events, are themselves still in the development phase. The back-and-forth between tweaking algorithms and dealing with device-specific limitations creates a noisy environment for accurate mobility analysis.

1.3.1. Missing data

One of the most significant issues with smart surveys and travel apps is missing data. This can occur due to various reasons, such as:

- **Satellite-based technical issues**: The quality of GPS data can be compromised by environmental factors, such as tall buildings or dense foliage,

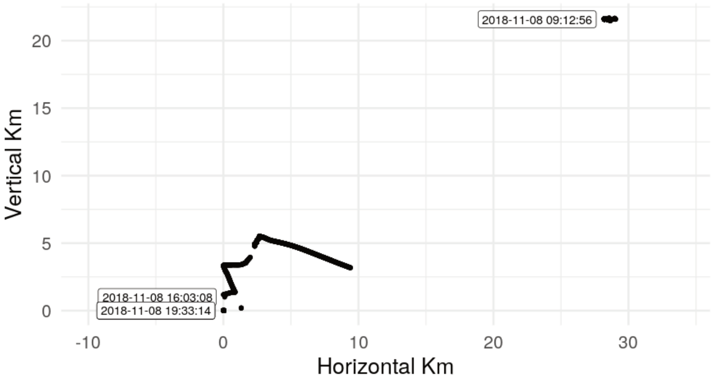


Figure 1.4 *A typical example of missing data due to software-based technical issues where the missingness is conspicuous due to the spatial displacement.*

leading to signal loss or inaccurate location tracking. Studies have shown that even short periods of signal loss can result in significant biases in mobility metrics, such as travel distance and movement events (Karaim et al., 2018).

- **Software-based technical issues**: Battery drain, software malfunctions, or OS functionality (such as ”doze-mode”) inhibiting the app’s ability to request locations or process data can lead to gaps in data collection (González-Pérez et al., 2022; Pejović, 2025).
- **User non-compliance**: Participants may forget to activate the app or fail to carry their devices with them during trips. This non-compliance can result in missing data for entire trips or segments of trips, leading to incomplete mobility records. Additionally, participants may not follow the app’s instructions correctly, leading to further data gaps.
- **Privacy concerns**: Some participants may be reluctant to share their location data, leading to incomplete or missing data. Privacy concerns can also result in participants turning off location services or providing inaccurate information, further compromising data quality.
- **Data transmission issues**: In some cases, data may be collected but not successfully transmitted to the server due to network issues or device limitations. This can result in missing data that is difficult to recover, as the data may be lost permanently if not properly stored locally on the device.

Imputation techniques, such as Multiple Imputation by Chained Equations (MICE) are therefore crucial as tools to solve a host of different issues at the same time. MICE is widely used across various disciplines to handle missing data, as it allows for the estimation of missing values based on the patterns in the available data. This is a predictive approach that helps not only in filling gaps, but also ensures that errors are correctly propagated. Imputation technologies like MICE have been praised for their ability to handle complex datasets and provide reliable estimates, which makes them a good choice for our use case.

While missing data is a significant challenge, it is not the only one. Other issues include participation willingness, and conceptual challenges. While many participants are still willing to permit location tracking for scientific purposes, there has been a decline over the last years in the willingness of participants to be tracked (Lunardelli et al., 2024). Additionally, conceptual challenges, such as defining what constitutes a stop or a track, or how to accurately secment movements, continue to pose issues. These challenges highlight the complexity of using smartphone-based applications for travel behavior measurement and the need for comprehensive solutions.

1.3.2. Integration issues

Beyond the problem of the missing raw data lies a more subtle challenge: integration. Once you've collected the smartphone location data (gaps and all), you're faced with the task of translating it into something comparable to traditional travel measures.

Traditional TDSs and Smartphone-Based Travel Surveyss (SBTSs) operate on fundamentally different principles. One provides discrete, pre-categorized events as interpreted by the respondent; the other offers a continuous stream of coordinates requiring extensive algorithmic processing to extract meaningful travel behavior. This difference isn't just inconvenient, it creates structural incompatibilities that affect how we analyze and interpret mobility patterns.

For instance, identifying a ”trip” in smartphone data requires algorithmic decisions about when movement begins and ends – decisions that might not necessarily align with how respondents would classify their own journeys in a traditional diary. For example, a person biking to a train station where they wait 15 minutes before getting on the train might rightly classify the biking leg as ending when they get off their bike, but label the ambiguous second leg as beginning immediately afterwards, whereas an app would likely begin the train-mode leg when the train itself began to move. Neither interpretation is inherently correct, but the difference in behavior leads to quantifiable differences between modes.

These integration challenges compound our missing data problems. When gaps occur in smartphone data, the algorithmic interpretation of surrounding data becomes even more uncertain, as we don't really have an appropriate standard. Did this 30-minute gap contain a trip? If so, where did it go and why? To answer the question on the basis of self-reported information or retrospective surveys would first require a comprehensive understanding of the differences in measurement attributable to each of the two modes.

Even when working with relatively complete data, for example after imputation, the differences in data structure make direct comparisons between traditional and smartphone-based measures problematic. This complicates validation efforts and makes it difficult to establish continuity with historical mobility data – a crucial consideration for longitudinal mobility studies and transport planning.

1.4. The big problem

We know that the potential is there for travel apps, and smart surveys in general, to create a massive step forward in reducing measurement error. If we fix the bugs, *the concept* is very straightforward, even if the process is not: if you can accurately record someone's location, that's 90% of the job of the survey. From there you can work out the address, the distance traveled, the route, the times, all of it, and with very little input necessary from the respondent. And that's important because people, generally speaking, don't want to fill out surveys. So either they don't do it, and they're non-respondents, or they do it but stop partway, and they're break-offs, or they say they will do it, and then realize they didn't do it, go back to try to remember the day, and just make up an average day, and they're adding measurement error.

The promise was compelling enough that originally, it seemed like smart surveys were going to win a higher response rate as respondents weren't as turned off by by the burden. Although still up for debate, the last several years have provided some evidence to the contrary (Fritz et al., 2025). Even without an improvement in response rates, they *can* allow us to reach people more easily and keep them on for a longer period of time at little cost to either respondent or researcher. This scalability is a decisive advantage, even if the initial implementation has been more challenging than anticipated.

This pattern of promising technologies facing early limitations before achieving their potential before methodological refinement is a common thread in innovation. The shift from traditional sequencing methods to Next-Generation Sequencing in genomics offers an instructive parallel. Sanger sequencing worked perfectly well for small-scale DNA analysis but proved impractically slow and expensive for sequencing entire genomes (Gomes & Korf, 2018). The development of parallel sequencing methods didn't immediately produce better results - in fact, because they used shorter segments of DNA, they had much higher error rates (Molnar, 2012). But the fundamental advantage of scalability eventually transformed the field, dramatically reducing costs and enabling applications impossible under the old paradigm.

What many of us suspect is that we're doing this because in 20-30 years, this is likely to be the only way to gather data. So, while there are other benefits in the meantime, like the availability of actual route information, which would allow models that currently use data from self-report travel studies (like the Dutch National Mobility Model, see (Smit et al., 2021)), to integrate at a more granular level than with Object Destination matrices, as they can have a one-to-one relationship with road sensors, the true benefit is our capacity to build these models *at all* in the near future. And the corollary to this is that, given that the current development of smartphones suggests trends that will result in data that are more piecemeal rather than less in the near future, development of strategies to account for this missingness are truly essential.

Like Next-Generation Sequencing relative to Sanger sequencing, the smartphone-based travel survey is (currently, at least) less reliable than its progenitor, but its capacity for continuous, longitudinal data collection at population scale represents a similar paradigm shift. The question isn't whether they'll replace traditional methods, but when and how we'll solve the methodological challenges currently limiting their potential. This thesis addresses one critical piece of that puzzle - the missing data problem that currently stands as the primary obstacle to realizing the full promise of smartphone-based travel surveys.

1.5. Current limitations and gaps of Smart Survey research

Much of the existing research has focused on addressing individual elements of the problem, such as developing algorithms for stop detection or mode prediction using complete datasets. To do so, these studies often rely on optimal conditions, such as using developer-unlocked devices or restricting data to complete sets. And naturally, where algorithms to address missing data have been developed, it has been under similarly optimal situations, such as having a years' worth of data for each participant, or by limiting the correction mechanism to the context of gaps small enough that they can be filled with algorithms based on external data. What is missing is a comprehensive understanding of the effects of missing data and practical solutions that can be applied to realistic datasets.

1.6. Research Questions

To address these issues, this thesis will explore the following research questions:

- What are the reasons for missing data in smart surveys and travel apps? Which issues are due to hardware, software, or other factors, and which are likely to persist into the future?
- How significant is the problem of missing data? What are the consequences for making independent estimates and comparing data longitudinally?
- Can we create a complete dataset that provides a foundation for accurate mobility analysis? Is this dataset better than having no data?

1.7. Aims and Outline

The progression of technology adoption follows a pattern that mirrors this thesis itself. What begins as a straightforward question – ”Can we use smartphones to collect travel data?”– inevitably leads into increasingly complex technical territory. It reminds me of a recent conversation about automatic windshield wipers. The concept is simple: detect whether the car is moving, activate wipers. And yet implementing this seemingly basic function required decades of engineering to handle the countless edge cases.

This pattern of increasing complexity upon moving from concept to implementation characterizes both technological evolution and the structure of this thesis. Each chapter follows logically from the previous one, but each step requires more specialized knowledge and increasingly sophisticated solutions as we tackle progressively thornier aspects of the missing data problem.

We begin with foundations. **Chapter 2** introduces the Statistics Netherlands travel app and explores its possibilities and challenges through a field test involving 674 participants. This chapter examines technical performance, response rates, and data quality, providing a first assessment of smartphone technology's feasibility in mobility research. It establishes the terminology and core concerns framing the thesis, presenting a broad overview of the challenges we face.

Having concretely established the reality of the missing data problem, **Chapter 3** asks what is really the only logical followup question: when should we be concerned about the levels of missing data in our surveys? This chapter introduces critical metrics for assessing mobility data, including sparsity calculations and the concept of Radius of Gyration (RoG), It's ultimately a very practical chapter in this way, serving as a guide for analyzing the data resulting from an app-based TDS. It outlines several essential algorithms for parsing raw location data, including stop detection and the Top-Down Time Ratio algorithm for segmentation. Whereas a traditional survey relies on respondents to interpret their own activities, smart surveys must accomplish this algorithmically—a seemingly simple task that reveals layers of complexity in implementation.

Finally, it evaluates the ”do nothing” approach to handling missing data, which has maintained its status as most common method of addressing these real-world challenges. The fact that interpolation alone performs quite poorly and results in biased results sets the stage nicely for the chapters that follow, which propose some new methodology.

Chapter 4 represents the methodological heart of the thesis, where we dive deep into the technical weeds. Here we introduce Dynamic Time Warping-Based Multiple Imputation (DTWBMI) as a novel approach to filling long gaps in human mobility trajectories. The apparent simplicity of the original question – ”How do we fill missing data?” – gives way to a complex methodological investigation requiring sophisticated time series analysis techniques. This chapter develops and tests two variants of the method (Dynamic Time Warping-Based Multiple Imputation - High Information (DTWBMI-HI) and Dynamic Time Warping-Based Multiple Imputation - Low Information (DTWBMI-LO)) through extensive simulation studies, comparing their performance against other approaches, such as Linear Interpolation (LI). Like the engineers fine-tuning the automatic windshield wipers for the infinite variety of real-world conditions, we're developing specialized solutions for the particular challenges of smartphone-based mobility data.

Finally, **Chapter 5** applies these methodologies to real-world data, demonstrating their practical utility and limitations. In it, we use the findings from the previous chapters to provide a roadmap for researchers working with smartphone-based travel survey data, particularly when dealing with the inevitable problem of missing data. This chapter brings our technical journey full circle, returning to practical applications after having been informed by the specialized knowledge developed along the way.

Throughout this progression, the thesis mirrors the technological development process itself: moving from identifying general problems to developing increasingly specialized solutions, each building upon the last. The work becomes progressively more technical as we go, but this increasing specialization is precisely what enables practical advances in how we collect and analyze mobility data. Just as each refinement in automatic windshield wiper technology addressed another edge case, each methodological development in this thesis tackles another aspect of the missing data challenge, bringing us incrementally closer to reliable smartphone-based travel surveys.

Part II

On identifying and addressing missing data

2

An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges

Abstract

Advances in smartphone technology have allowed for individuals to have access to near-continuous location tracking at a very precise level. As the backbone of mobility research, the Travel Diary Study, has continued to offer decreasing response rates over the years, researchers are looking to these mobile devices to bridge the gap between self-report recall studies and a person's underlying travel behavior. This article details an open-source application that collects real-time location data which respondents may then annotate to provide a detailed travel diary. Results of the field test involving 674 participants are discussed, including technical performance, data quality and response rate.

Parts of this chapter have been published in McCool, D., Lugtig, P., Mussmann, O., & Schouten, B. (2021). An app-assisted travel survey in official statistics: Possibilities and challenges. Journal of Official Statistics, 37(1), 149–170. https://doi.org/10.2478/jos-2021-0007 (McCool et al., 2021).

2.1. Introduction

Understanding the true underlying movement behavior of persons in a given geographic area is a key component in the foundation of national infrastructure decisions. Institutions responsible for generating official statistics have designed streamlined instruments to enable the collection of important travel behavior metrics. Most organizations currently implement some form of TDS, in which participants record a series of trips and stops over a specified time period. When these diaries are completed within probabilistic samples, the aggregate results can be used to model travel demand between regions, generate statistics on transportation modes, or monitor the uptake of green incentives such as telecommuting.

Usage of TDS in official statistics to create a granular picture of individual travel behavior over time has spanned more than half a century. Modes of administration have evolved with the times, from face-to-face interviews in the 1950s transitioning gradually into mail and telephone survey instruments in the 1980s and 90s, followed by a transition to web-based methodology in the early twenty-first century (Adler et al., 2002; Arentze et al., 2005; Axhausen, 1995). Although this mode evolution has led to both reduced costs as well as increased ease of administration, reliance on respondent recall for generation of the diary has remained constant.

Researchers have long been aware of the tendency of this method to lead to trip underreporting (Clarke et al., 1981; Richardson et al., 1995). Recent studies comparing concurrent GPS and recall methodologies have demonstrated that reliance on recall methodology produces underreporting of short trips, differences in reported trip departure times, overestimation of trip length, and underestimation of vehicle miles of travel (Bricka et al., 2009; Carrion et al., 2014; Forrest & Pearson, 2005; Kelly et al., 2013; Stopher & Shen, 2011; Wolf et al., 2003). More generally, the response rates for TDS have have been decreasing steadily over the decades. Bricka et al. (2009) noted disproportionate non-response for large households, low-income households and younger adults, and Ogle et al. (2005) showing similar non-response trends within households making the fewest and most trips. Although presumably less burdensome, studies using standalone GPS devices have presented similar non-response challenges, but do seem to increase uptake among younger participants (Bricka et al., 2009). Managing lowering response rates requires increasing the cost per respondent of an already expensive design, which has prompted researchers to find more cost-effective ways to access the information.

Smartphone-based travel studies have been proposed as a solution for addressing issues of cost and decreasing response rate among younger households. Smartphone penetration over the last decade has neared saturation, with recent numbers from Statistics Netherlands indicating that over 90% of the Dutch population owns a mobile device (Centraal Bureau voor de Statistiek, 2019). Additionally, Roddis et al. (2019) found that respondents rated interaction with an app more enjoyable than either a traditional user-completed travel diary or a personal log, rating it as less burdensome, and both Roddis et al. (2019) and Safi et al. (2017) demonstrated that smartphone-based apps provided higher-quality data in comparison with recall-based TDS. These potential advantages have led to the introduction of multiple app-based travel diaries (Berger & Platzer, 2015; Cottrill et al., 2013; Greaves et al., 2015; Lynch et al., 2019; Prelipcean et al., 2018). To date, however, these studies have yet to address the unique challenges involved with the fieldwork within the general population, nor has the impact of a large-scale implementation been assessed. To that end, the primary goal of this research is to present a realistic assessment of a smartphone-based travel study within a national sample, showcasing problems at the different component levels of Total Survey Error. For fitness in general population surveys, important requirements are acceptable recruitment rates, low drop-out, low in-app missing data, and high in-app data quality. This article explores these features. Apart from these more methodological requirements, there are complex logistical and procedural requirements. Although these are referenced here, they are not focal in this exploration.

Introduction of the app-based TDS is not a silver bullet. There are known issues arising from Global Navigation Satellite Systems (GNSS) measurements themselves, ranging from the problems of urban canyons to the length of time required to establish an initial signal (Park et al., 2014). Additionally, although processed Android location data can be as accurate as +- 10m, this accuracy varies across different devices (Liu et al., 2017; Menard et al., 2011). Even when the accuracy is acceptable, technical issues with the applications themselves can lead to completely missed travel behaviors (Roddis et al., 2019). Independent of technical issues, semantic issues present an additional hurdle. The transition to automated stop identification from respondent's subjective interpretation of their travel behavior has proven a difficult task, and the reduction in burden expected from travel model identification is yet to surface (Prelipcean et al., 2016; Yang et al., 2016; Zhao et al., 2015). Importantly, apps must also remain user-friendly, or risk early dropout (Assemi et al., 2018). In this study, we aim to investigate the feasibility of an app-based TDS to be robust enough against these errors to be usable in official general population surveys of travel behavior.

2.2. Statistics Netherlands Travel Application overview

The initial objective of this study was to develop an application that would be able to assess the potential of smartphone technology in mobility research. To this end, the app was designed both to collect the data of interest as well as metadata and user input that would allow assessment of the data quality. In order to collect and represent back to the participant their mobility data, the app needed to provide latitude and longitude updates frequently enough to reliably determine location, separate these measurements into a series of moving periods (tracks) and stationary periods (stops). This would then allow the user to enrich the data with auxiliary day-level and trip-level information.

The Travel App System is comprised of a front end and a back end. The front end consists of the SN Travel App, which collects the location data, resolves stops and tracks, and exposes these to the user for annotation purposes. Both the raw location data and the resolved data are stored locally in a SQLite database on the mobile device. The back end consists of an API written in GO that performs the data ingestion and transformation into a PostgresSQL database that ultimately receives and stores the data (see 2.1).

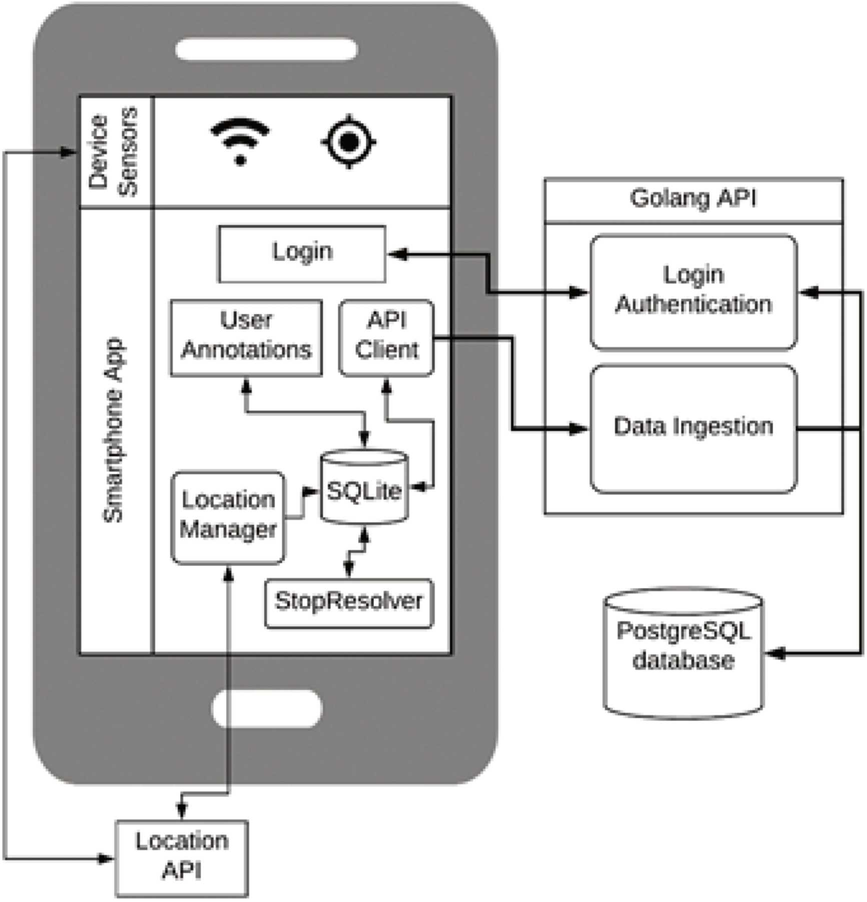
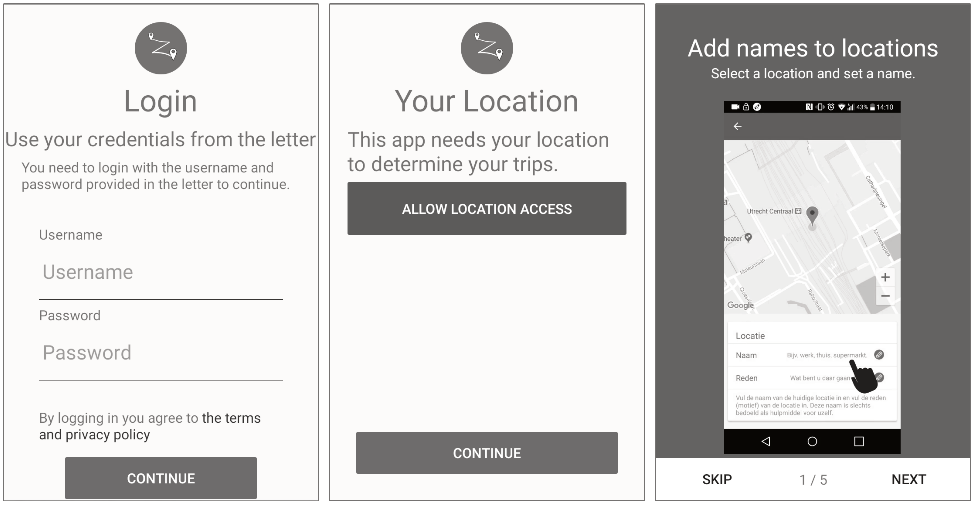


Figure 2.1 *Technical implementation and integration of back end and front end*

2.2.1. Statistics Netherlands Travel App

In order to be able to deploy equivalent algorithms to both Android and iOS versions, the client was developed in C# using the Xamarin framework. This framework provides compilation to Intermediate Language which is Just-in-Time compiled to native assembly on Android devices, and Ahead-of-Time compiled into native ARM assembly code for the iOS build. The application was developed Open Source and hosted on a publicly accessible collaborative code repository in order to facilitate distribution and address potential privacy concerns. Full code for backend and frontend is available at https://gitlab.com/tabi.

User Interface Users who download the application to their smartphone are requested to log in using the credentials received in the invitation letter received in the mail. Upon successful registration of their device on the server, users are asked to enable location permissions on their device. Specifics of this permission request differ across devices, although the UI prompt does not. Following this, regardless of permission status, users are provided with a brief video tutorial explaining proper use of the application. Users are shown how to navigate between stops and tracks within a day, adding annotative information, as well as how to pause and resume the application's location-tracking behavior (see 2.2).



Login

Use your credentials from the letter

You need to login with the username and password provided in the letter to continue.

Username

Username

Password

Password

By logging in you agree to the terms and privacy policy

CONTINUE

Your Location

This app needs your location to determine your trips.

ALLOW LOCATION ACCESS

CONTINUE

Add names to locations

Select a location and set a name.

Locatie

Naam Bijv. werk, thuis, supermarkt.

Reden Wat bent u daar gaan

Vul de naam van de huidige locatie in en vul de reden (motief) van de locatie in. Deze naam is slechts bedoeld als hulpmiddel voor uzelf.

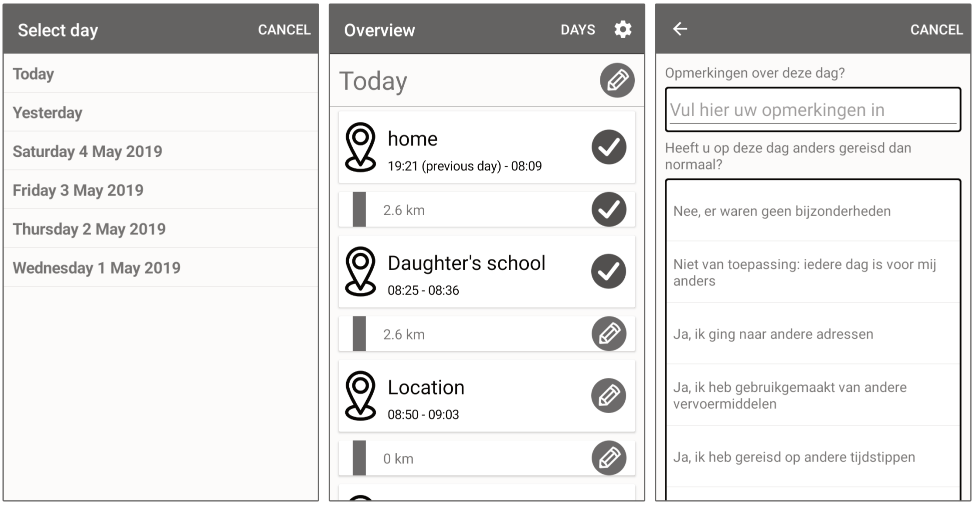
SKIP 1 / 5 NEXT

Figure 2.2 *Interface for initial login, permissions, and use instructions*

While the application is running, an icon is visible in the system bar. When the user accesses the notification drawer, a message is visible alerting the user that the app is running in the background to track movement.

The user interface for adding annotations is organized by day, with days beginning at midnight. A user opening the app while it is running in background mode will be taken to a list of the current day's stops and tracks, called the Day Overview. From the Day Overview, it is possible to return to the calendar to see a list of available days, to drill down into each stop or track, or to answer day-specific questions (see Figure 2.3.)

Clicking on a stop opens a map with a point and surrounding radius representing the user's registered location for that time period. Clicking on a track will provide a map with two points representing start and stop locations and a blue line representing their movement trajectory. This map can be manipulated by zooming or panning in order for the user to determine with greater accuracy where they were. The stop



Today
Yesterday
Saturday 4 May 2019
Friday 3 May 2019
Thursday 2 May 2019
Wednesday 1 May 2019

Today
home
19:21 (previous day) - 08:09
2.6 km
Daughter's school
08:25 - 08:36
2.6 km
Location
08:50 - 09:03
0 km

Opmerkingen over deze dag?
Vul hier uw opmerkingen in
Heeft u op deze dag anders gereisd dan normaal?
Nee, er waren geen bijzonderheden
Niet van toepassing: iedere dag is voor mij anders
Ja, ik ging naar andere adressen
Ja, ik heb gebruikgemaakt van andere vervoermiddelen
Ja, ik heb gereisd op andere tijdstippen

Figure 2.3 *User Interface for travel log display and day-level questions*

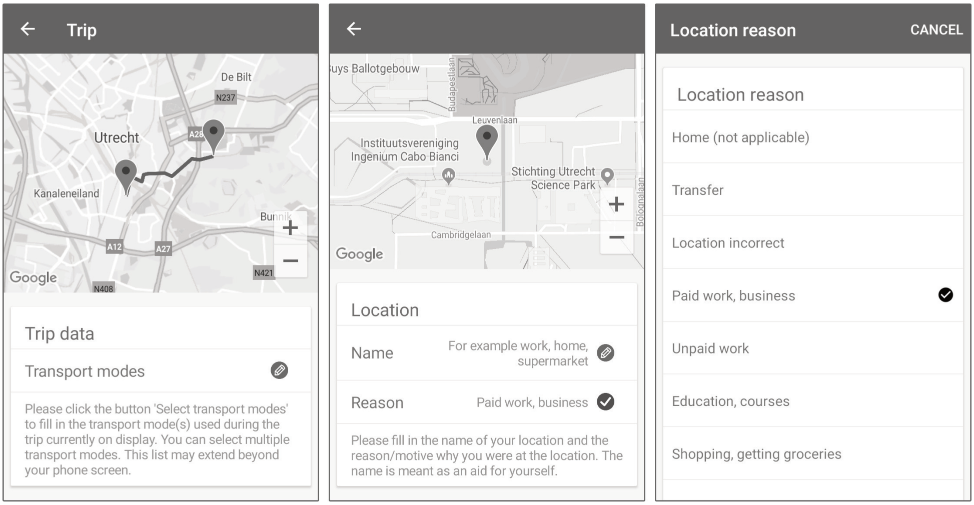
menu requests users to record a name and reason for the stop. Figure 2.4 shows the track menu, which requests that users enter the mode of transportation used in the trip leg.

The user is presented with differential icons indicating whether the annotation item has been completed. A user who completes a day by assigning motives and mode annotations to each stop and track on a day as well as completes the day-level annotation questions will see a check mark for the day.

2.2.2. Algorithms and implementation

The application collects and saves raw data from an operating-system-specific location API implementation. For iOS, this is Core Location. For Android, Google Play services' Google Location Services API is preferred. When it is not available, the native android location API is called in its place. All three APIs return similar information. Although the methods that produce the location information are proprietary, it is known that these location APIs aim to produce accurate location results by combining information from GNSS, local Wi-Fi signals and cell phone tower signals. Android devices offer a setting external to our application where the user may prefer to use only GNSS signals, or to turn on 'high-accuracy mode' which provides a location derived from the combination of Wi-Fi, Bluetooth, cell tower and GNSS signal information. Devices running iOS as an operating system also offer a setting to disable GNSS, although it is accessible only through a series of menus. Beginning with iOS 11, Apple devices offer the user the option to share location information with the requesting application either only when the app is open and running in the foreground, or at all times.

The format and content of the coordinates returned by the location APIs vary by the polled system. Latitude and longitude are universally returned, although the



Trip

Trip data

Transport modes

Please click the button 'Select transport modes' to fill in the transport mode(s) used during the trip currently on display. You can select multiple transport modes. This list may extend beyond your phone screen.

Location

Name — For example work, home, supermarket

Reason — Paid work, business ✓

Please fill in the name of your location and the reason/motive why you were at the location. The name is meant as an aid for yourself.

Location reason CANCEL

Location reason

Home (not applicable)

Transfer

Location incorrect

Paid work, business ✓

Unpaid work

Education, courses

Shopping, getting groceries

Figure 2.4 *User Interface for trip and stop detail annotation*

number of significant digits varies. Altitude is reported if and only if a connection is established between the device and a GNSS system. The accuracy returned by the Google Location Services API is defined as the radius of 68% confidence in meters. The interpretation of the horizontal accuracy variable returned by CoreLocation is currently undocumented by Apple.

High/Low tracking The application running on the mobile phone requests information from the OS-specific API at regular intervals in order to generate a pattern of movement. When the device is determined to be not in motion, the application requests a location update once per minute in order to preserve battery life. When the device is in motion, the location request is submitted once per second. On Android, the application also accepts other location updates that were requested by different applications with the goal of providing the highest possible accuracy at no increased cost to battery life.

An algorithm was developed in order to determine when a device should move between the two tracking profiles. This algorithm has two parameters: time and distance. The implementation is similar in the Android and iOS implementations. A listener is engaged to receive location updates, all of which are saved in the location repository on the device. Each updated location is checked against the time parameter to determine whether a distance check should be performed. Once the time parameter is exceeded, the distance from the immediately preceding location is calculated using the Haversine formula. This calculated distance is compared to the distance parameter to determine whether or not the user is moving. In the event that the reported accuracy of the location is larger than the distance parameter, this number is used instead of the distance parameter in order to offer adjustment for the potentially erratic behavior in a situation in which multiple low-accuracy locations are returned because of poor signal availability. If the calculated distance exceeds this number, the application activates high tracking mode and begins requesting location updates at the rate of one per second. If it does not, low tracking mode is activated and location updates are requested once per minute.

Stop detection/resolution The stop detection algorithm developed within this application functions similarly to the high/low-tracking algorithm. The algorithm identifies a stop when the device reported location is within a given radius for a particular length of time. By altering these two parameters, respectively radius and duration, it is possible to adjust the sensitivity with which a set of locations can be consolidated into a single stop. In Section 3 Field Test, we describe the design of our study in which we varied these attributes across participants. A simulation study conducted post-hoc on the field-test data in which we clustered participants' locations in each combination of radius and duration parameters identified minimal differences in the number of identified stops (Kiillaars et al., 2019).

While all location updates were stored in the repository, only those locations where the reported accuracy was less than 80 meters were used for the separate step of stop determination. The stop detection algorithm consists of four steps. First, the timestamp of the last known stop is requested from the internal location repository. Second, the set of locations with timestamps greater than or equal to the last known stop are returned. Third, the set of all positions are divided into groups based on the distance between the positions; this step involves calculating the distance between each new proposed location and the set of all locations identified as belonging to a stop. If this distance is less than the radius parameter, the location is incorporated into the existing stop. If the distance is greater than the radius parameter, it becomes the first location in a proposed new stop. Fourth, the time elapsed between the first and last location in the proposed stop is calculated and compared against the duration parameter. If the elapsed time exceeds this value, the mean latitude, longitude and the beginning and ending timestamp are registered in the local database.

Adjacent stops are merged when the distance between the two stops' average position is less than one hundred meters. Following the stop resolution phase, stop visits can be processed. Stops may be places to which users return, such as a house or workplace. A stop visit is therefore a singular instance of having been at a stop.

Tracks Tracks were defined as the set of locations between stops. Start time, beginning latitude/longitude and ending latitude/longitude are saved in the local repository. Track length was calculated as a summation over individual distances between consecutive locations after filtering out coordinates with an accuracy greater than 100 meters.

2.2.3. Other data collected

Data were collected at various intervals. Upon registration with the server, device information was recorded in the database, including phone manufacturer, model, OS and version number. As described in the preceding section, location data were collected at time intervals of varying length and processed client-side into tracks and stops.

Users were requested to annotate all tracks with the mode(s) of transportation utilized in the movement. Various common options were provided for the user in a drop-down select menu in addition to an open text field to accommodate less common forms of transportation. Users were requested to annotate stops with the motive of the stop. Users were provided with a list of many common motives, such as home, work and shopping. Additionally, users could submit their own motive or mark the stop as incorrect. Stops could be given a name for ease of recall.

To facilitate detection of unexpected events, users were requested to provide feedback on three day-level questions: 'Did you have your phone with you at all times today?', 'Was today a normal day for you?' and 'Do you have comments on the day?'.

2.3. Field test

A field test was designed and conducted in late 2018 in order to address two primary research questions:

1. To what extent are persons willing to register an app and provide a week of time-location sensor data?
2. What is the quality of the resulting sensor data and additional survey data?

To facilitate comparison with existing mobility research, the TDS, conducted by the Dutch Ministry of Infrastructure, was selected as the basis for the pilot study (Centraal Bureau voor de Statistiek, 2022). ODiN is designed as a Computer Assisted Web Interviewing (CAWI) diary study, where respondents keep track of all their trips, including start- and endpoints, and times for a specific day of the week.

2.3.1. Participants

The field-test survey was sent to 1902 sample persons, half of whom were randomly sampled from the Dutch population register and half of whom were sampled from the pool of previous ODiN respondents whose surveys had been conducted in the months of September and October 2018. The target population consists of people 16 years and older living in non-institutionalized households.

2.3.2. Methods

Incentive stratification A random stratification of the sample was made for incentive conditions. Persons were randomly assigned to one of three incentive conditions: €5 + €5 + €5, €5 + €0 + €10 and €5 + €0 + €20 and received a brief with the corresponding information. All incentives were paid in the form of gift cards mailed to recipients. All sample persons received an unconditional €5. One third of the sample was promised a split sum of €10, with €5 conditional on registration and €5 conditional on seven days of recorded travel data. One third was promised €10 conditional on seven days of travel data. One third received €20 conditional on seven days of recorded travel data. In the following, we will omit the reference to the unconditional €5 and denote the incentive conditions by $5+5$, $0+10$ and $0+20$.

Stop detection stratification Sampled persons were additionally allocated different stop detection parameters at random. Participants were assigned a duration, d, of either 2-, 3-, 4- or 5-minute intervals and a radius, r, between 60 and 100 meters inclusive, representing the maximum distance from a central point of a stop. These parameters were implemented to influence the relative strictness of the algorithm to automatically identify a stop, with lower values representing looser criteria and higher values representing more stringent requirements. These parameters were implemented in the back end of the app and were unknown to participants.

2.3.3. Materials

The sample was sent an invitation letter by mail with app login information and both a QR code and URL leading to the SN Travel App landing page. The invitation letter arrived on November 2nd and persons were given the opportunity to respond until December 15th. On the landing page, persons were given background information and a brief explanation of the study and the app, and were directed to the appropriate application store based on device operating system (Android or iOS). On November 16th, those persons yet to register with the app were sent a reminder letter in the mail containing the same login information, QR code and URL. On November 23rd, respondents who had not reached seven days of travel data were sent a motivation letter. All letters, the app, and the landing page contained information on how to contact Statistics Netherlands if there were questions or difficulties.

2.3.4. Analysis plan

As the primary goal was to establish the feasibility for future widespread implementation, we aimed to investigate initial uptake, study dropout and data quality. Initial uptake analyses were stratified by the demographics available in the Dutch population register, as well as by incentive condition and previous participation in TDS status. Dropout analyses were stratified by these same variables and additionally included information on the type of mobile device used by the participant and its operating system. Quality of the data was judged by frequency of collection and alignment of summary measures for field test data and ODiN responses on the subset of participants who were involved in the September 2018 data collection.

2.4. Results

2.4.1. Registration and response

Of particular interest was whether use of an app-based TDS could lead to acceptable participation rates, and to determine which factors would lead to either nonresponse or participation. Additionally, this study sought to determine the extent to which nonresponse and dropout were selective by using administrative data available on all Dutch residents of all potential respondents.

The following sequential stages were required for full participation in the study: receiving and reading the invitation letter, downloading the application, registering within the app, accepting location permissions on the device, not closing the app for the full seven days, and providing annotative data. Unit response was considered to coincide with the third step, device registration.

Of the 1902 respondents who were sent a letter, 674 registered a device using the login information they received, leading to a total unit response rate of 35.4%. Previous ODiN participants had a response rate of 44.4% compared with the newly-acquired sample's response rate of 26.5%. This reflects a slightly lower response rate than the 31.0% and 27.9% obtained for ODiN in 2018 and 2019 respectively (Centraal Bureau voor de Statistiek, 2022; Centraal Bureau voor de Statistiek (CBS) & (rws-Wvl), 2020). As shown in Table 2.1, the three incentive conditions 0+20, 0+10 and 5+5+5 achieved response rates of 39.7%, 36.4% and 30.1% respectively.

The measure of non-registration represents an overestimation of nonresponse. However, distinguishing non-contact, refusal and non-eligibility from other factors was possible only for those persons who independently contacted Statistics Netherlands. Contact was received from people who attempted to participate in the study, but were unable to download and install the application for various reasons. Unregistered is, within this study, a close analogue of nonresponse. Most who fail to register a device are likely to be traditional nonresponders.

Table 2.1 *Device registration (unit response) rate by sample source and incentive condition.*

	Unregistered		Registered	
	n	%	*n*	%
Sample source				
ODIN respondents	529	55.6	422	44.4
Newly acquired	699	73.5	252	26.5
Incentive				
5 + 5	443	69.9	191	30.1
0 + 10	403	63.6	231	36.4
0 + 20	382	60.3	252	39.7
Total	1228	64.6	674	35.4

2.4.2. Nonresponse

Device registration status varied across known demographic variables from the Dutch population register, which is a governmental database containing administrative information on all persons registered as living in the Netherlands. Younger persons were more likely to register a device than older persons. Immigrants were less likely to register a device than those originally of Dutch origin, and first generation immigrants were less likely to register a device than second-generation immigrants. Persons with college degrees or higher were more likely to register a device than those with elementary or vocational school degrees.

Divorced and widowed persons were less likely to register a device in comparison with persons who were never married. However, single-person households were less likely to respond than family households with children. Home-owners were more likely to register a device than were renters. People with higher household incomes were also more likely to register their device. See Table 2.2.

Most transportation-related characteristics that could be drawn from the Dutch population register, including possession of a car, moped or lease car, were not significantly related to registration status. Possession of a driver's license, however, was significantly related to registration status, with those in possession of a driver's license more likely to register a device than those not in possession of a license.

Geographic variables, including address density and province- or city-related variables were not significantly related to device registration status. A full table of response across all available variables can be found in Table A.1 in Appendix A.

2.4.3. Drop-out

Technical difficulties contributed to ambiguity in the identification of dropout as participants reported an inability to send data following successful device registration. Of the 674 participants who registered a device, 98 were never able to send GPS data. Participants who contacted Statistics Netherlands with this problem were instructed to upgrade their OS version, reinstall the application, or install on a different device. Each additional registration was recorded separately. In total, the 674 participants completed 748 registrations, of which 706 were unique configurations (differing model or OS version), and the remaining 42 were reinstallations. In total, 136 unique-to-user device configurations produced no GPS data. Table 2.3 shows the distribution of lack of GPS data across OS and OS version. The likelihood that an Android device would send GPS location data at least once generally increased across operating system versions. Although we found that iOS devices were more likely never to send data, we did not observe differences across versions of iOS. Distinguishing in a determinative way those users who expressly denied location permission when prompted in the OS from those who experienced technical issues with location provision or for whom other settings in their mobile device disallowed location provision was not possible with the available data.

In addition to dropout over time within the seven day period, we identified a pattern of dropout within a day in which the app would not send location reports during

Table 2.2 *Device registration by sample characteristics.*

	Unregistered		Registered	
	n	%	*n*	%
Age categories				
[15,30]	240	58.0	174	42.0
(30,50]	343	58.7	241	41.3
(50,70]	437	66.5	220	33.5
(70,96]	202	83.8	39	16.2
Origin				
Dutch	961	62.5	577	37.5
Non-western	130	76.5	40	23.5
Western	131	69.7	57	30.3
Generation				
First	149	78.8	40	21.2
Second	112	66.3	57	33.7
Marital status				
Married	624	63.7	355	36.3
Never married	410	60.8	264	39.2
Divorced	121	71.2	49	28.8
Widow/widower	67	91.8	6	8.2
Education				
Vocational	155	72.4	59	27.6
Elementary	63	78.8	17	21.2
Secondary	282	60.1	187	39.9
Graduate	74	49.7	75	50.3
University	136	48.2	146	51.8
Unknown	512	72.9	190	27.1
Household type				
Single	244	72.6	92	27.4
Partners	420	64.2	234	35.8
Partners, child	473	60.4	310	39.6
Single parent	79	69.3	35	30.7
Other household	6	66.7	3	33.3
Has drivers license				
No	281	73.4	102	26.6
Yes	941	62.2	572	37.8
Home ownership				
Own	790	61.6	492	38.4
Rent, corporation	275	73.1	101	26.9
Rent, other	123	66.8	61	33.2
Unknown	34	63.0	20	37.0
Total	1222	64.5	674	35.5

Note. Omits 6 non-responders lacking register data. All χ^2 differences significant, p < .01.

Table 2.3 *OS and version by data status*

	No locations		Locations	
	n	%	*n*	%
Android				
<6.0	5	18.5	22	81.5
6	5	10.6	42	89.4
7	3	3.5	83	96.5
8.0	9	5.0	170	95.0
8.1	1	4.8	20	95.2
9	0	0.0	2	100.0
iOS				
<11.4.1	10	25.0	30	75.0
11.4.1	7	22.6	24	77.4
12.0	4	22.2	14	77.8
12.0.1	40	24.4	124	75.6
12.1	21	23.1	70	76.9
Total	105	14.9	601	85.1

the full 24 hours, but would continue to send data either later in the day or on a subsequent day. While it is not possible to distinguish intentional closing and reopening of the application from the data, the patterns identified in the data lend support to the idea that this effect comes from the behavior of the OS. In order to investigate this phenomenon and determine its effects, we created a measure called gap time. Gap time refers to the length of time between two subsequent location reports from a single device. A device functioning properly and without additional restrictions imposed by the operating system of the device should have gap times of approximately one second while in motion, and of one minute while stationary. After removing duplicate records, on average, we identified a mean gap time of 18.8 seconds during trips and a mean gap time of 47.5 seconds during contiguous non-trip activity.

At face value, receiving an update during trips at half the rate and updates during stationary periods more often than requested seems to indicate that very little data is lost. However, if we instead consider the maximum gap time per day per user, a different picture emerges. Figure 2.5 shows that approximately 30% of trips and 65% of non-trip activity contains gap times greater than one hour. Any statistics calculated on the basis of available data, therefore, must consider this fact, as aggregate summative statistics will likely be underestimated if we assume that gap times are likely to cover periods of meaningful data. Although determining the length of an acceptable gap time to consider coverage to be complete may depend on end goals, choosing a maximum gap time no larger than fifteen minutes yielded an average of 12.8 hours covered per user-day.

Complicating the problem, the gap times are not evenly distributed throughout the day. In fact, there is evidence that these gap times represent app or device fall-off, leading to a greater proportion of missingness later in the day than in the beginning

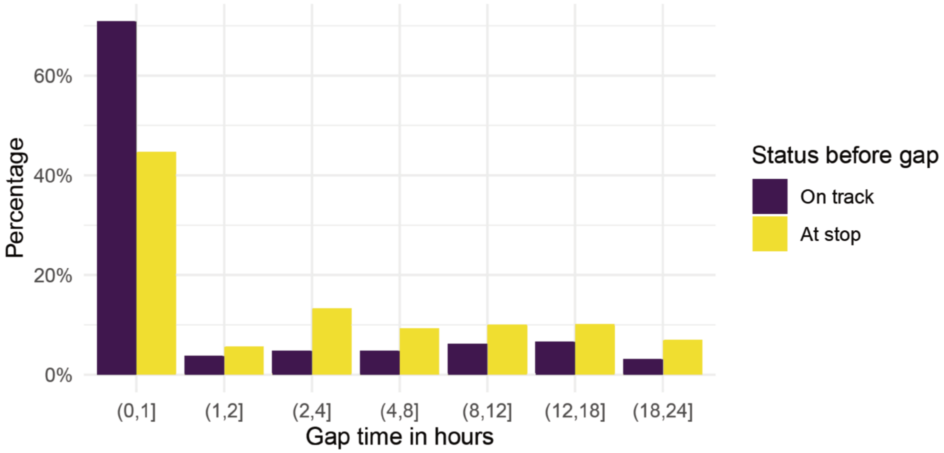


Figure 2.5 *Maximum gap time per user day*

Table 2.4 *Descriptive statistics of trips per day by survey*

Survey	Respondent Days	Mean	Median	Max
ODiN	321	3.54	3	12
Current Study	1353	5.13	4	33

of the day. Figure 2.6 shows the hour at which contact is lost and the hour at which we again start receiving data from the device, following from a gap time of at least thirty minutes.

2.4.4. Comparison to Traditional Diary

Data from the ODiN travel diary were made available to facilitate comparison between the app-based and travel diary methods. Only the data for respondents who participated in both this study as well as the ODiN study were used for comparison. Three measures were selected for comparison between the two methodologies: number of trips within a day, trip length and trip distance. ODiN respondents were asked to self-report information for a single, specific day that was assigned randomly. Trip-level measures were summed per day to establish number of trips per day, total trip length and total trip distance within a day. Trip-level characteristics were calculated per trip and then averaged for the user.

The total number of trips within a day were compared between the ODiN data and this study. As shown in Table 2.4, the median number of trips as determined by the SN App mechanism was four as compared to the median of three in the ODiN data. Additionally, we see a distribution with a much longer tail from the SN App data in Figure 2.7. Some portion of this increase likely represents a desirable outcome for our study, in which we are capturing short trips known to pose problems in self-report measures.

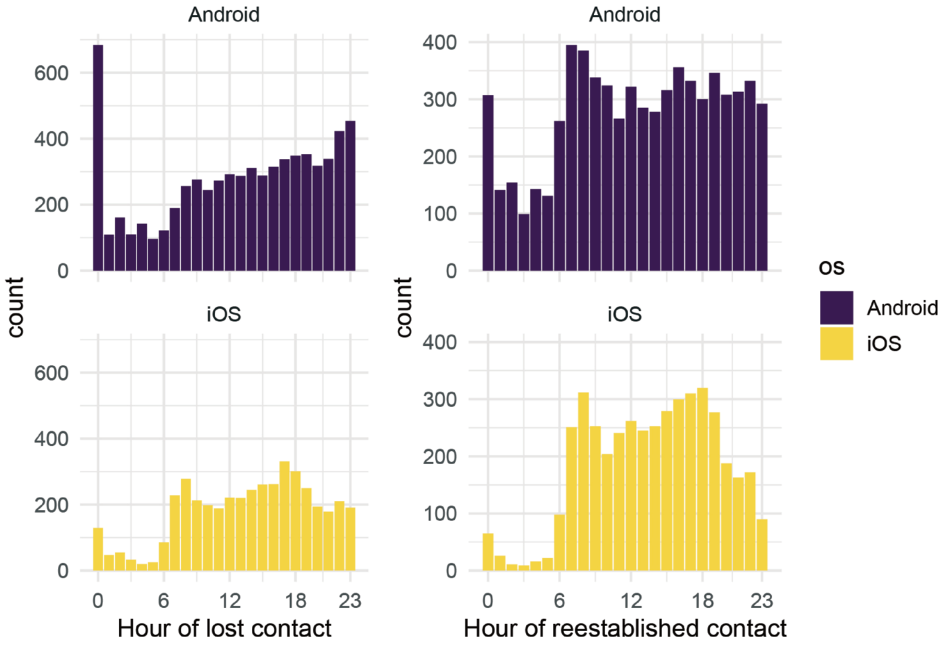


Figure 2.6 *Hour of resumed contact after >= 60 minutes of no device activity*

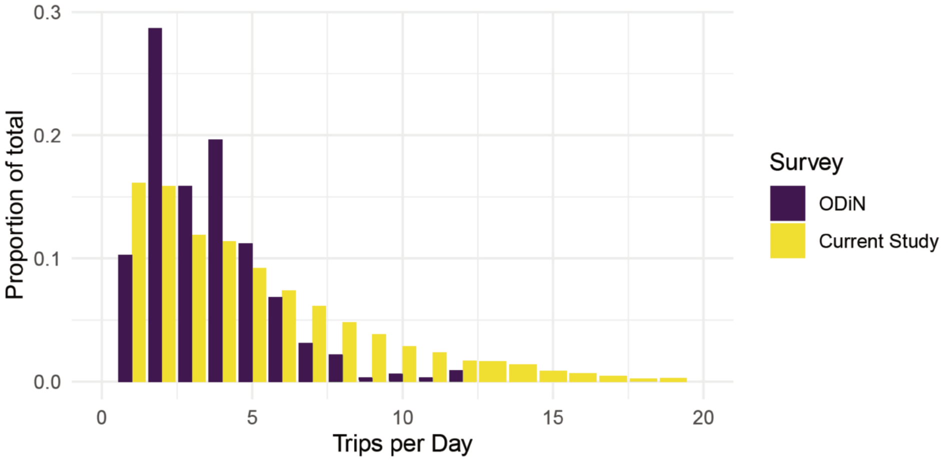


Figure 2.7 *Comparison of number of trips per user day*

Table 2.5 *Descriptive statistics for travel time (hours) per day by survey*

Survey	Respondent Days	Mean	Median	Max
ODiN	321	0.59	0.33	8.00
Current Study	2327	1.14	0.51	20.26

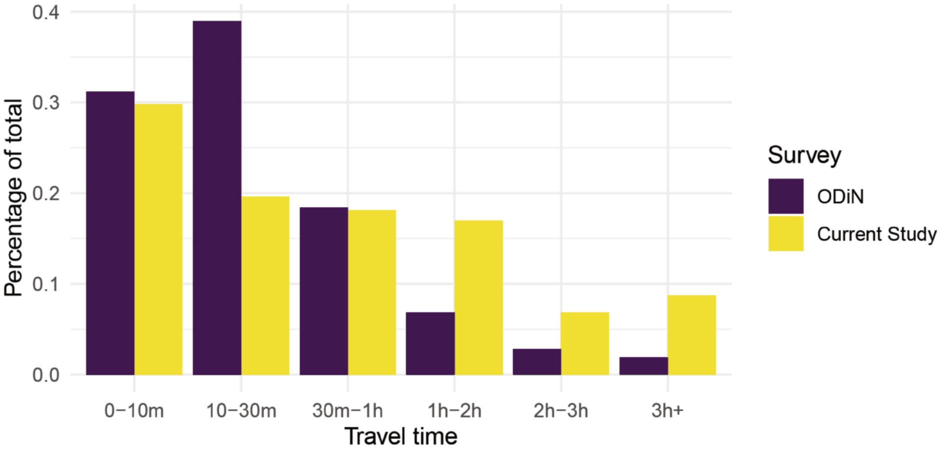


Figure 2.8 *Travel time per day by survey*

Additionally, our study records a longer time spent in transit per day than did ODiN. The average time spent in travel is higher for the SN app data, as shown in Table 2.5. Figure 2.8 shows recorded total travel time per day within both the current study and ODiN. The SN App recorded more time spent in all categories above one hour, which is in alignment with our hypothesis that automation would capture travel behavior at a more granular level. However, the combination of an increased number of trips per day and an increased time spent in travel could additionally be due to mechanisms within the tracking application that are too sensitive to movement, inflating both counts by including trips that would generally fall outside the purview of travel behavior, such as trips from an office workplace to a canteen in the same building. These differences may also be related to natural month-to-month discrepancies in travel behavior.

Curiously, although we demonstrate both more trips per day and more time spent traveling, aggregated distance within a day is notably shorter in this current study as compared to the ODiN data. Table 2.6 shows descriptive statistics for total distance traveled in a day within both studies. ODiN respondents reported a median distance per person-day of 32.2 kilometers compared to the median of 8.47 kilometers as tracked within this study. Figure 2.9 shows the differential distribution, with over 40% of person-days within the current study summing to fewer than 5 kilometers in comparison to approximately 15% within ODiN. It is likely that the missing data problem contributes to this difference. For example, consider a situation in which

Table 2.6 *Descriptive statistics for travel distance (KM) per day by survey*

Survey	Respondent Days	Mean	Median	Max
Current Study	2290	34.68	8.47	1366.41
ODiN	319	288.22	32.30	12884.00

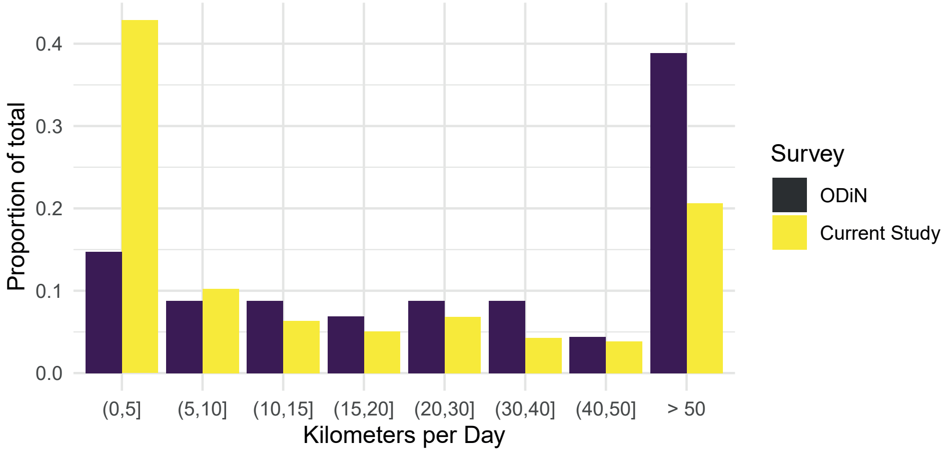


Figure 2.9 *Kilometers per day by survey*

the app successfully initiates tracking a commuter successfully in the morning, but which loses contact in the course of the day, dropping the likely return trip in the evening.

2.4.5. Conclusions

The primary goal of this research was to determine whether or not direct implementation of smartphone travel diaries was feasible in large scale own-device studies. In order to be considered feasible, we outlined requirements that we must have an acceptable response rate and the data collected from the device must be of high enough quality to allow for important metrics to be calculated reliably.

Response rates for this study were similar to those of the travel diary study ODiN. Although prior to 2018, response rates were upwards of 50%, this was largely a function of the CATI and CAPI contact initiatives following initial nonresponse, which were very costly. Since 2018, these modes have been phased out, leaving only the CAWI mode, which lowered response rates to 31.0% and 27.9% respectively for 2018 and 2019. The 26.5% response rate for this study's new responders is similar, but in place of a single day's data, they are responding to our request for a week's worth of data.

Respondents also proved willing to provide annotative data to the passive traces. Although this currently takes the form of labeling stops for purpose and tracks for transportation mode, these are variables which could conceivably be calculated from the data itself as the accuracy of mobile-device GPS units approaches centimeter-level accuracy (Dabove & Di Pietra, 2019; Humphreys, 2018). This frees researchers to reduce respondent burden by either relying on the passive data itself or adapting a verification approach where respondents either confirm the predicted mode or correct it.

Importantly, this study identifies a major hurdle for researchers wishing to transition to app-based TDS implementations. There is ample evidence within our data to suggest that concerns over dropout and missing data are inherently device-related. A recent survey of the state of the art of current smartphone-based travel apps demonstrated that across 22 different apps, device characteristics such as operating system and phone manufacturer impacted both quantity and quality of collected data (Harding, 2019). Researchers in this field have remained hopeful over the years that improved smartphone technology would address current limitations with missing data, low accuracy, urban canyons and battery life (Allström et al., 2017; Berger & Platzer, 2015; Cottrill et al., 2013; Geurs et al., 2015; Gong et al., 2014; Greaves et al., 2015; Verzosa et al., 2017). Although these predictions are likely warranted as they pertain to the inexorable march of technology, it does not necessarily follow that the resultant data will be any better as new technological issues will arise to replace the old. Consider battery life: in order to keep pace with increasing battery requirements, Android has implemented mechanisms to close apps without informing the user, leading to gaps and unintentional drop-out in the data (Petter et al., 2019). For the same battery concerns, iOS has taken over strict control of the location management system, restricting the frequency that locations can be polled, leading to the problem of cold starts in the data.

Researchers ultimately have little control over the device upon which the application will be installed. Google, Apple and the various device manufacturers are unlikely to be open in releasing details describing the precise functioning of their location systems. Additionally, as versions move quickly and much changes between iterations, a system that functions well one year may require an intensive change in the following year in order to continue to function, and the data that are generated may ultimately be quite different as well. It may be that researchers must, at least for the time being, continue to involve participants in assessing their passive data with active control questions.

It is easy to view this as a negative if we view it in the context of the relatively clean and complete data that comes back from written survey instruments. Researchers have been promised higher-quality data that strictly improve upon the TDS, but the current study demonstrates that any increase in quality is by no means free. Instead, these data must be judged on their own terms. The richness is not only useful, but key to its use. Identification of the issues underlying missingness and measurement is an important first step. The next step must be to find ways to compensate for the technological issues as we do for other methodological issues, and here the unique characteristics of the data itself may prove invaluable: the longitudinal nature of the data collected may allow recovery of missing track segments within a person’s own travel log, the spatial nature of the data may allow us to correct for measurement error and the sheer size of the data may open up new methods of inference to us. Technological challenges will persist, necessitating development of robust methodology for identification of the true underlying behaviors, but the data are sufficient to provide grounds for good inference if researchers can move past these initial steps.

3

Maximum interpolable gap length in missing smartphone-based GPS mobility data.

Parts of this chapter have been published in McCool, D., Lugtig, P., & Schouten, B. (2022). Maximum interpolable gap length in missing smartphone-based GPS mobility data. Transportation. https://doi.org/10.1007/s11116-022-10328-2 (McCool et al., 2022).

Abstract

Passively-generated location data have the potential to augment mobility and transportation research, as demonstrated by a decade of research. A common trait of these data is a high proportion of missingness. Naïve handling, including listwise deletion of subjects or days, or linear interpolation across time gaps, has the potential to bias summary results. On the other hand, it is unfeasible to collect mobility data at frequencies high enough to reflect all possible movements. In this chapter, we describe the relationship between the temporal and spatial aspects of these data gaps, and illustrate the impact on measures of interest in the field of mobility. We propose a method to deal with missing location data that combines a so-called top-down time ratio segmentation method with simple linear interpolation. The linear interpolation imputes missing data. The segmentation method transforms the set of location points to a series of lines, called segments. The method is designed for relatively short gaps, but is evaluated also for longer gaps. We study the effect of our imputation method for the duration of missing data using a completely observed subset of observations from the 2018 SN travel study. We find that long gaps demonstrate greater downward bias on travel distance, movement events and radius of gyration as compared to shorter but more frequent gaps. When the missingness is unrelated to travel behavior, total sparsity can reach levels of up to 20% with gap lengths of up to 10 min while maintaining a maximum 5% downward bias in the metrics of interest. Temporal aspects can increase these limits; sparsity occurring in the evening or night hours is less biasing due to fewer travel behaviors.

3.1. Introduction

Sensor data has the potential to introduce new depths to travel survey by reducing response burden, human error, and subjectivity. In transportation planning, health studies and ecological research, data from GNSS is now well into its second decade of use. Collection of individual traces using respondents' own mobile devices rose with the ubiquity of sensor-equipped mobile device. Research shows that this individual-level sensor data has the capacity not only to collect mobility data, but to provide a basis for behavioral interventions (Batool et al., 2022; Cellina et al., 2019). Despite this fact, most publications represent field tests and introductory apps (Allström et al., 2017; Chambers et al., 2017; Marra et al., 2019; McCool et al., 2021). In fact, a recent SWOT analysis of SBTS found that the majority of research teams discontinued their applications following the initial project (Pronello & Kumawat, 2021). Only recently have SBTS advanced to second rounds of data collection, or undertaken research beyond their own feasibility (Axhausen et al., 2020; Molloy et al., 2020; Patterson et al., 2019).

Because SBTS aim to collect data both at a high frequency and over a lengthy period, almost all studies encounter problems with missing data, often wholly outside the control of any involved party (Gadziński, 2018; Harding et al., 2021; Wang et al., 2018; Xie et al., 2020). While many researchers opt to remove cases or periods of time containing sparse data, this both risks biasing results and functionally reduces sample sizes in a field where participation is already limited (Körner, 2012; Wang et al., 2019). Researchers who aggregate the data to remove the spatiotemporal nature discard the potential benefits inherent in this novel method of data collection, and may still introduce bias (Baratchi et al., 2014). Robust methods of handling missing data arise independently from any of six different fields, (statistics, machine learning, transportation, engineering, geoscience or computer science) (Chen et al., 2016; Harrison et al., 2020; Servizi et al., 2021; Shen et al., 2014). Each field may have its own terminology and set of underlying assumptions, leaving researchers in search of best practices to parse through disparate partial solutions.

Crucial to avoiding the propagation of biases from the raw data is quantifying the impact of the missing data. Quantification can be thought of as a multi-step process. First key outcome measures should be identified, as these guide the decision making process. Second it is possible to establish a relationship with some mechanism available in the incomplete data, such as the amount of missing data or the length of the gap. A third step would allow trip characteristics to develop this relationship further. Step four is the estimation of the missing measures of interest, which must be based on the relationships uncovered in previous steps, given the selected complete data. Finally, we may allow for steps three and four to vary across individual features.

In this paper, we detail steps one through three. The remainder of the introduction discusses causes and existing solutions for missing data, while Section 3.2 outlines methods for describing the extent of missingness in the raw data and pre-processing it for analysis. Section 3.3 explores the relationship between bias in mobility metrics and missing data by inducing missingness into complete mobility data. Finally, Section 3.4 discusses the results of the simulation, and suggests methodology for establishing the limits at which simple mechanisms for addressing gaps in the data begin to fail.

3.1.1. Causes of missingness

Causes of missingness in passively-collected location data vary in both cause and relative impact (Hecker et al., 2010; Shen & Stopher, 2014). Mechanisms due to signal loss, such as blocked line of sight, or the ”cold start” problem, often produce small gaps. Others may produce longer gaps, obscuring one or more trips within a day, by causing the device to stop transmission, such as battery drain, termination of the app by the device or user, powering off of the device, or entering into hibernation mode. Some of these may take days to resolve, leading to wholly missing days. Lastly, device incompatibility and respondent willingness can lead to missingness at the user-level.

The simplest cause of signal loss is interrupted line of sight. In a study comparing GNSS-generated trajectories with users' recorded travel diaries, 15.7% of the GNSS trajectories contained at least one instance of blocked line of sight (Stutz, 2019). This tends to be related to particular location-related circumstances, e.g., traveling through a tunnel or underpass. In these cases, there will be a loss of signal transmission, leading to no data being recorded for the length of time during which the GPS satellite and the phone are unable to establish a connection. If the gap is linearly interpolated by connecting the coordinates immediately preceding and following this gap, this assumes a path that is perfectly straight in the intervening time. As tunnels and underpasses are often constructed with the shortest distance in mind, the expected impact on distance or number of discrete travel events is minimal, but any deviation of the true path from this linearity will bias estimates based on the underlying behavior. On the other hand, a true straight path may show little to no bias, even if the signal is blocked for a much longer period, as might be the case with a train traveling through a tunnel.

A secondary physical cause is the so called ”urban canyon,” which may occur in areas where many tall buildings are situated close together, and have the potential to block line of sight (Chen et al., 2010). Similarly, urban environments can contain ”black holes,” or areas in which all entering trajectories may disappear, as documented by Hong et al. (2015). Unlike missing data caused by a short tunnel, travel behavior within urban canyons or black holes is not restricted to a generally straight path. A slight downward-biasing of distance and radius of gyration would be expected in data that contain missingness due to urban canyons, and stops within the built-up area may be lost entirely. Urban canyons can also lead to noisy data, when reflected satellite signals cause erroneous triangulation.

The cold start problem is another common cause of signal loss, with one study demonstrating that 27.5% of all trajectories contained a cold-start period (Stutz, 2019). Because someone's location is usually determined by Wi-Fi sensors when indoors, and by GPS outdoors, the start of a trip is often missed during the hand off at the boundary. The process of identifying a sufficient number of GPS satellites in order to accurately record a position is not immediate and can take anywhere from 20 seconds to 12.5 min to provide an initial position (Langley, 2015). Someone can range quite far within this time period, and, unlike in the tunnel situation, is unlikely to be following a strictly linear path from the point at which they left the building, to the point at which the signal is regained. A downward bias of estimates of distance traveled and radius of gyration in calculating naive statistics with linear interpolation is expected. When the underlying travel behavior represents a round trip that is sufficiently short, an entire trip may be lost.

While map features and data characteristics may allow for distinguishing the causes of short gaps, the generation mechanisms behind long gaps are more difficult. A common pattern of missing data occurs when the device itself ceases acquiring data on behalf of the application. This can be as a consequence of user intent – the user has shut off their phone or closed the application. This can also be unintentional, and part of the design of the operating system. Android operating systems and iOS operating systems alike both have introduced measures to limit battery drain when a phone user is not directly engaged with the device, called variously Doze mode or Hibernation (Bähr et al., 2022; McCool et al., 2021). Additionally, some phone manufacturers include versions of the Android operating system that will aggressively kill apps, preventing them from running in the background (Zhou et al., 2020). Consequences vary with respect to the length of the gap and the cause of the gap. A lengthy period of missing data followed by a period of activity may indicate that the device was in hibernation or doze mode. Yoo et al. (2020) report higher levels of sparsity in the nighttime hours, likely due to this OS behavior. In this case, the impact on distance, spread of activity (measured as RoG), and discrete trip events is likely to be minimal. However, there is little to distinguish this from the case in which the app has been closed, either by the user or the OS, and subsequently reopened by the user. This has the potential to obscure extensive travel behavior, and may produce large downward biases in distance traveled and RoG, and will often miss trips and stops. The longer the gap, the more influential it is likely to be.

Devices that cease recording data due to battery discharge may demonstrate similar data patterns, but may be identifiable if the app records battery life history. This is more likely to occur during a trip, leading to missing trip ends. As the location updating process tends to be in itself draining on the battery, longer trips are both more likely to use battery, as well as prohibit charging for longer periods of time. If data collection begins again, this is likely to occur at a known stop, such as home or work.

These situations lead to gaps in a user's location history that can be quite large, ranging from hours to days or weeks. Lacking a model for predicting and contextualizing the interim period, accurately accounting for the bias becomes impossible.

Table 3.1 *Existing solutions for missing data*

Study	Gap[1]	Assumptions	Auxiliary info
Barnett 2020	L	Recorded behavior $\simeq$ missing behavior	Longitudinal data
Bierlaire 2013	S	Trajectories follow maps	Transportation network data
Bihrmann 2015	L	Geographic variables not of interest	-
Huang 2020	B	Trajectories represented by key metrics	Dense spatial overlap
Li 2021	S	Trajectories follow maps	Map data
Liu 2021	S	Gaussian gaps, strictly MCAR	Longitudinal data
Meseck 2016	S	Missing assumed stationary	-
Nawaz 2020	B	Spatial aggregation	Longitudinal data
Prelipcean 2015	B	Recall data are accurate	Travel survey
Schuessler 2009	S	Traj. follow maps	Transportation network data
Zhao 2021	L	Recorded behavior $\simeq$ missing behavior	Longitudinal data

Note. [1] Applicable to gaps of length L = long, S = short, B = both

3.1.2. Previous studies

Previous studies have proposed addressing missing data within the SBTS in certain ways. Prelipcean et al. (2015) used self-completion trip diaries to fill gaps in passively-generated data to establish a joint ground truth. Meseck et al. (2016) filled each gap with the median location of the preceeding twenty coordinates. Bihrmann and Ersbøll (2015) performed multiple imputation on aggregate measures of interest. Huang et al. (2020) implemented fuzzy c-means imputation on missing taxi Global Positioning Satellite (GPS) data to construct missing segments within the trajectories. Barnett and Onnela (2020) and Zhao et al. (2021) sampled from existing trajectories to fill in gaps. Schuessler and Axhausen (2009), Bierlaire et al. (2013) and Li et al. (2021) use map-matching methods to improve sparse data. Others, such as Nawaz et al. (2020) and Liu and Onnela (2021), use other methods of probabilistically establishing what occurs within gaps. Table 3.1 provides an overview of some recent methodologies.

While methods have been proposed for managing these gaps, nothing is currently available as a benchmark to researchers looking to assess the extent and composition of their own missing data, in order to guide the choice of when and how to apply these methods (Hwang et al., 2018; Yoo et al., 2020; Zhao et al., 2018). We fill this gap by simulating missing data with different characteristics based on real travel survey data. In the simulations, we vary gap length and density. Doing so, it is possible to set lower thresholds for when missing data becomes problematic. This critical first step enables selection of an imputation mechanism on the basis of research goals and data availability.

In this paper, we evaluate the method of linear interpolation for addressing gaps of varied sizes, under the assumption that certain features may define gaps where it is likely that users have followed a mostly linear path. We distinguish this from map-matching methods, which may be applied in a subset of these situations, but which are more complex, and often unavailable for pedestrian or bike routes.

3.2. Methods

When location data are collected, they are sampled from an underlying continuous trajectory. Two consecutive sampled points will be separated both by distance and by time. The shorter the time interval between the sampled points, the more accurately the continuous trajectory is approximated. A consequence of this discretization of the continuous trajectory is that all location history data contain missingness due to the nature of sampling. This limits the extent to which GPS traces can be categorized either as wholly complete or wholly missing. Instead we propose a metric to establish the impact of potential missing information between successive points.

3.2.1. Sparsity

By discretizing a respondent's total observation time, $\mathcal{T}$, into a number of same-length intervals of length τ, τ becomes the temporal resolution of our missingness analysis. τ must be chosen to reflect the goals of the eventual analysis, reflecting an interval that is short enough to preclude missing impactful behavioral changes, but long enough to encompass the sufficient sampling interval.

The discretization of $\mathcal{T}$ into intervals of length τ leaves us with $\frac{\mathcal{T}}{\tau} = T$ intervals, 1, ... , T.

Each interval in a user's trajectory can be assigned an indicator representing presence, $r_t = 1$, or absence, $r_t = 0$, of at least one record during the time period. The proportion of $r_t = 0$ relative to T provides a measure of sparsity with respect to τ, parameterized as q in Equation 3.1.

$$q = \frac{1}{T}\sum_{t=1}^{T}(1 - r_t) \tag{3.1}$$

We can extend this measurement of sparsity across persons, units of time (e.g. days, weeks), and states (e.g. traveling, stationary). We speak of N persons, $i = \{1, 2, \dots, N\}$. Each person i has data occurring in J_i units of time, $j = \{1, \dots, J_i\}$, and K_{ij} states, $k = \{1, \dots, K_{ij}\}$. Each state k contains T_{ijk} intervals, $t = \{1, \dots, T_{ijk}\}$. Let r_{ijkt} represent a binary indicator of any record for discrete time period t in state k in time interval j for person i. This leads to the following full equation for sparsity shown in Equation 3.2.

$$q = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{1}{K_{ij}}\sum_{k=1}^{K_{ij}}\frac{1}{T_{ijk}}\sum_{t=1}^{T_{ijk}}(1 - r_{ijkt}) \tag{3.2}$$

3.2.2. Segmentation

Location data are generated as a time-stamped sequence of points which reflect a geographical position at a moment in time. These coordinates reflect someone's

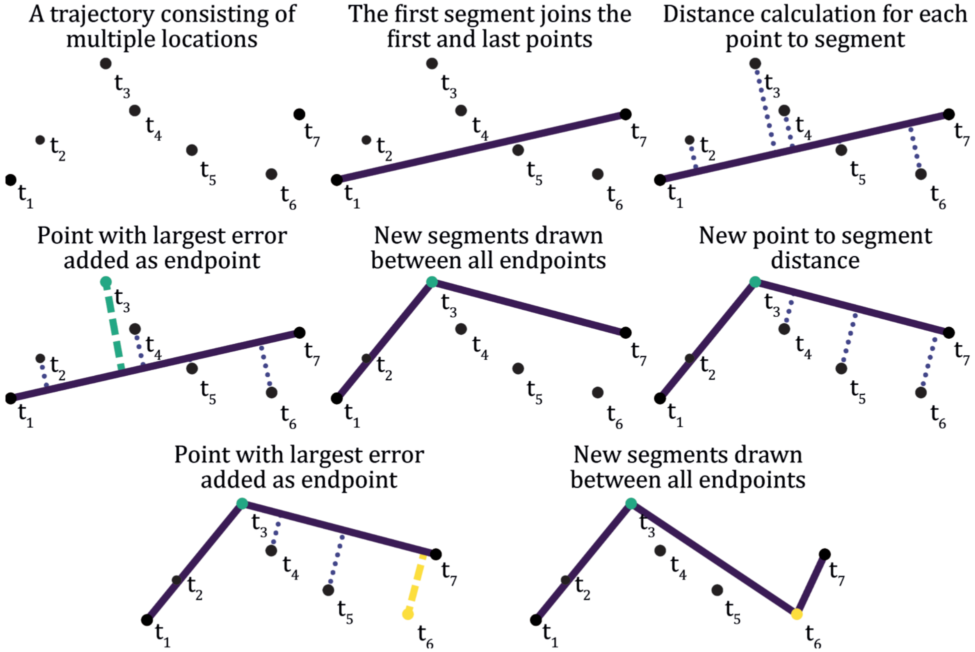

A trajectory consisting of multiple locations
The first segment joins the first and last points
Distance calculation for each point to segment
Point with largest error added as endpoint
New segments drawn between all endpoints
New point to segment distance
Point with largest error added as endpoint
New segments drawn between all endpoints
t_1
t_2
t_3
t_4
t_5
t_6
t_7

Figure 3.1 *Ramer-Douglas-Peucker Algorithm*

continuous location history with a set of discrete points which make calculations on underlying trajectories costly and conceptually difficult. If the data can be reduced in size without sacrificing information, it is possible to simultaneously reduce both the computational complexity as well as the number of assumptions that must be made about individual locations. One method of data reduction is by partitioning time-stamped trajectories $\{p_0, p_1, ..., p_i, ..., p_n\}$ into straight line segments $\overline{p_0, p_i}$ $\overline{p_i, p_n}$ that sufficiently represent the path (Lee & Krumm, 2011). These line segments can then be used for calculations on properties of the underlying trajectory such as speed or distance. They can additionally be more easily compared against other trajectories consisting of line segments in order to identify similar paths and come with the added benefit of reducing measurement error common to GPS navigational systems.

Methods of segmentation vary, but most follow along with the well-known Ramer-Douglas-Peucker algorithm's method of creating new segments based on the magnitude of discrepancy between the proposed segment and the point that it should represent (Douglas & Peucker, 1973; Ramer, 1972). Figure 3.1 illustrates the original Ramer-Douglas-Peucker algorithm in which endpoints are successively introduced within the trajectory at the point with the largest perpendicular euclidean distance. This simulation study implements the Top-Down Time Ratio algorithm outlined in (Meratnia & By, 2003). This is an extension of the Ramer-Douglas-Peucker algorithm in which endpoints are selected on the basis of largest spatial euclidean distance. A segment is generated between the first location in a trajectory and the last location in the trajectory. Each recorded location point between the two segment ends is given a proposed new point along the generated segment, with respect to the elapsed time. The distance is calculated between the recorded point and this pseudo-sampled point. The point lying furthest from its segment is selected to form a new segment end, whereupon the process begins again. Algorithm 1 describes the process as implemented in this study.

The algorithm is iterative and if allowed to run indefinitely will create $N-1$ segments for N points. In order to be useful it must be given a stopping criteria such as the number of segments or the maximum error distance between the recorded points and the adjusted points that lie along the line segment. The smaller the error, the more information is preserved, and the more segments are created. Too many segments will reduce capacity for later imputation. It is therefore important to choose stopping criteria such that we find a balance between the two opposing aims.

3.2.3. Metrics of interest

The relative impact on commonly used mobility metrics was used as an evaluation criteria for outcome comparison. Trajectories were decomposed into stop and move sections using an implementation of a rule-based stop classifier, as described in (Montoliu et al., 2013). As implemented within this study, no upper limit was set for dwell time as this was unnecessary given the maximum trajectory length of 24 h. (See Appendix B, Algorithm 6 for complete details.) Individual subsequent stops were merged into single stops if their centroids were less than 100m distant in order to reduce the number of incorrectly differentiated stops.

Distance metrics were calculated using the Haversine method (Robusto, 1957). To arrive at total distance, the distance between all segment endpoints was calculated for the entire trajectory. Moved distance involved summation of distances between all segment endpoints when segmentation was performed on move events only. Similarly, total move time was established by summation of elapsed time for each move event segment.

RoG is calculated as the root-mean-square time-weighted average of all individual locations during an individual's 24-h period, as shown in Equation 3.3. It was necessary to weight by time due to the unequal frequency of location collection, since the higher density movement trajectories would otherwise inflate estimates of the metric.

$$\sqrt{\frac{\sum_j w_j \times dist([\overline{lon}, \overline{lat}], [lon_j, lat_j])}{\sum_j w_j}} \tag{3.3}$$

Where $\overline{lon}$ and $\overline{lat}$ are defined respectively as $\frac{\sum_j w_j lon_j}{\sum_j w_j}$ and $\frac{\sum_j w_j lat_j}{\sum_j w_j}$ and w_j is a weighting element, representing half the time interval during which a location was recorded $w_j = \frac{t_{j+1} - t_{j-1}}{2}$

Algorithm 1 Top-Down Time Ratio extension of Ramer-Douglas-Peucker algorithm

```
Input: $l$ = locs[1, ..., N], $d$ = max allowed distance
Output: locs′[1, ..., N] with segment id annotation
function topDownTimeRatio($l$, $d$)
    Annotate $l$[1] with
        segmentstart $\leftarrow$ true
    Annotate $l$[] with
        segmentid $\leftarrow 1$
    function iterate($l$)
        for all segmentid do
            $\Delta_{segtime} \leftarrow l[N]_{time} - l[1]_{time}$
            $\Delta_{seglatdiff} \leftarrow l[N]_{lat} - l[1]_{lat}$
            $\Delta_{seglondiff} \leftarrow l[N]_{long} - l[1]_{long}$
            for all $l$[i] in segmentid do
                $\Delta_{partialtime_i} \leftarrow l[i]_{time} - l[1]_{time}$
                $TimeRatio \leftarrow \Delta_{partialtime_i} / \Delta_{segtime}$
                $l[i]_{lon'} \leftarrow l[1]_{lon} + \Delta_{seglondiff} \times TimeRatio$
                $l[i]_{lat'} \leftarrow l[1]_{lat} + \Delta_{seglatdiff} \times TimeRatio$
                $l[i]_{err} \leftarrow$ haversineDist($l[i]_{lon,lat}$, $l[i]_{lon',lat'}$)
            end for
            if any $l[]_{\text{err}} > dist_{max}$ then
                $s \leftarrow$ index of max($l[]_{err}$)
                Annotate l[s] with
                    segstart $\leftarrow$ true
                Annotate l[]with
                    segmentid $\leftarrow$ cumsum$(segstart)$
                iterate(l[])
            end if
        end for
    end function
end function
```

3.3. Relating bias and sparsity

3.3.1. Motivating example

As a motivating example, we consider the data collected from a 2018 field test of the SN travel app. This field test concerned 1902 sample persons aged 16 and older. The sample was evenly divided between a new random sample taken from the Dutch population register and a secondary group of respondents randomly sampled from participants who had participated in the study ODiN in the two months preceding the field test. ODiN is an online-only study of individual mobility in the Dutch population (Centraal Bureau voor de Statistiek (CBS) & (rws-Wvl), 2020).

Both groups of respondents were contacted via post with a request to download the application onto their personal mobile devices, register using the enclosed personal username and password, and record seven days of movement behavior. Full details on app methodology and data structure are available in McCool et al. (2021).

While the app was running on the phone, it captured a participant's location once per second while the person was determined to be in motion, and once per minute while the person was determined to be stationary. This determination was based upon an algorithm that assessed whether or not the displacement between recorded intervals exceeded thresholds indicating movement behavior.

Collectively, a total of 2087 person days were recorded amongst 576 participants. The mean length of participation was 13.3 days, and the average number of hours with location information in a day was 8.2 h. The large percentage of missing data rendered calculation of the summary statistics of interest, such as number of trips and distance, difficult without careful consideration of the underlying mechanisms leading to the missing data.

Missing data in the Statistics Netherlands travel app

Because gap times differ, the choice of τ impacts assessment of overall data sparsity. The distribution of sparsity within the data set was evaluated with τ set variously to be 1 min, 5 min, and 15 min. Because the sampling interval while stationary was set to 1 min, reducing our temporal interval to the width of the sampling interval does not allow for the same level of discrimination between persons. A temporal resolution of 15 min may be too large to preclude non-negligible travel behavior. Figure 3.2 shows the distribution of q_i in the full data set under these three temporal resolutions. With a τ of 1 min, a very small percentage of our data would achieve a sparsity <.5. However, the difference between temporal resolutions of 5 and 15 min is less pronounced. When selecting for complete data, we do so on the basis of 5 min intervals.

Selection of complete data

A subset of data were selected where the $q_{ij}|\tau_5 < .05$ for a contiguous 24 hour period. Figure 3.3 provides a graphical breakdown of the exclusion steps leading to this selection. In total, 185 persons representing 584 complete 24-h periods remained. As we intend for the simulation study with induced missingness to be

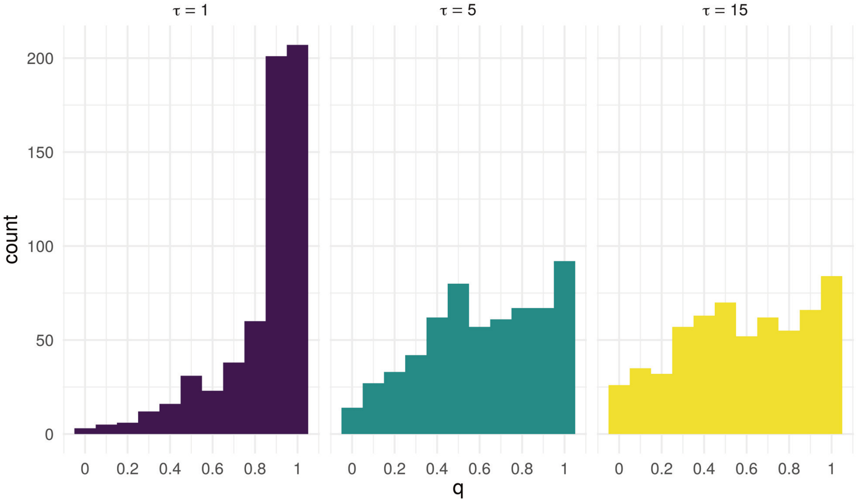


Figure 3.2 *Sparsity across* τ, $\mathcal{T}$ *< 7*

generalizable to those persons with true missingness, we tested whether the persons with complete data were likely to be representative of the group as a whole. For this analysis, we make use of the fact that we have independently recorded data on travel behavior from the ODiN study for many individuals. It was possible to link 354 of the 360 from the ODiN sample that had provided at least some data.

Of these 354, 114 recorded at least one complete day within the app (group CD), and 240 did not (group NC). Groups were compared on three measures of interest from the proposed simulation study using a 2-sample permutation test and 10^4 iterations. Group CD recorded less active travel time (μ = 77.0, σ = 58), than group NC (μ = 93.9, σ = 77), *p* $= .02$. Travel distance was similar between group CD (μ = 44.4, σ = 62.2) and group NC (μ = 54.7, σ = 71.5), $p = .17$, as was mean number of trips for group CD (μ = 3.5, σ = 2.3) and NC (μ = 3.3, σ = 2.0), $p = .47$. There may be some indication of differential travel behavior between groups CD and NC, with the group with more complete data more likely to have mobility behavior that reflects more time spent with the app actively recording locations once per second.

Selection of an error parameter for segmentation

A simulation study was conducted on the subset of complete set of data, $q_{ij} < .05$ in order to determine the relationship between the selected error level of the stopping parameter and the distance covered. Baseline comparison was to an unfiltered error parameter of 1 meter. Error conditions ranged between 1 and 150 meter tolerance. The data were either unfiltered, mean filtered, or median filtered prior to segmentation, as described in (Lee & Krumm, 2011).

Results from the simulation study demonstrate a relationship between total distance and maximum error that is dependent upon the number of move states that a per-

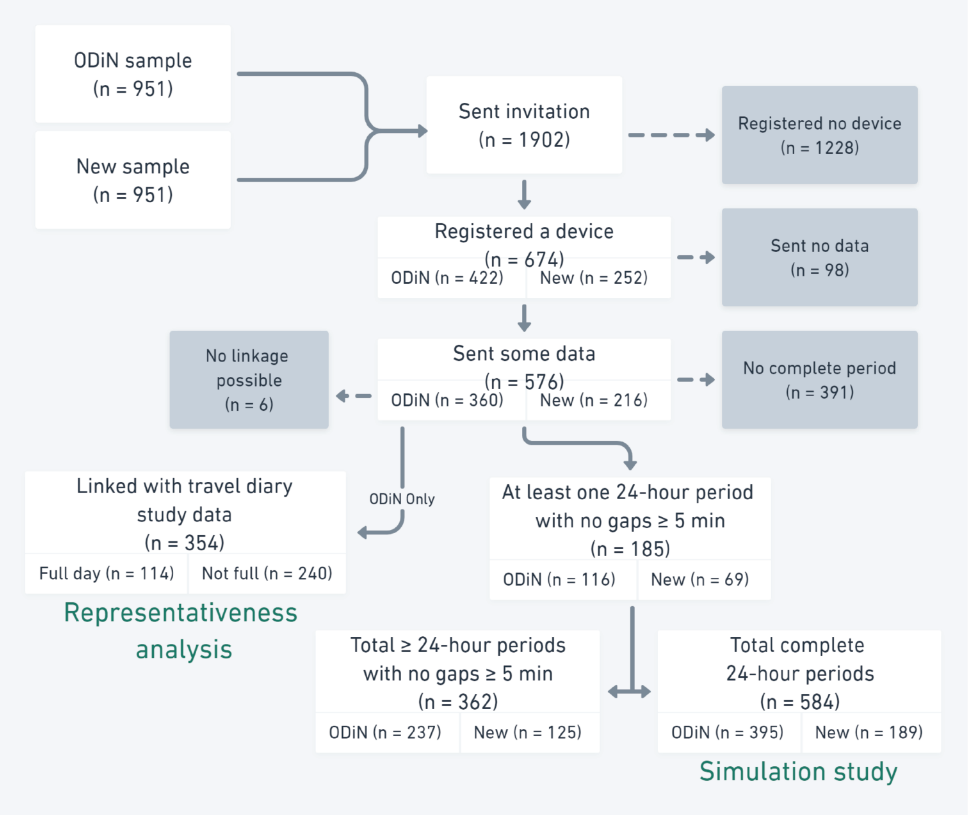


Figure 3.3 *Steps leading to selection of complete data*

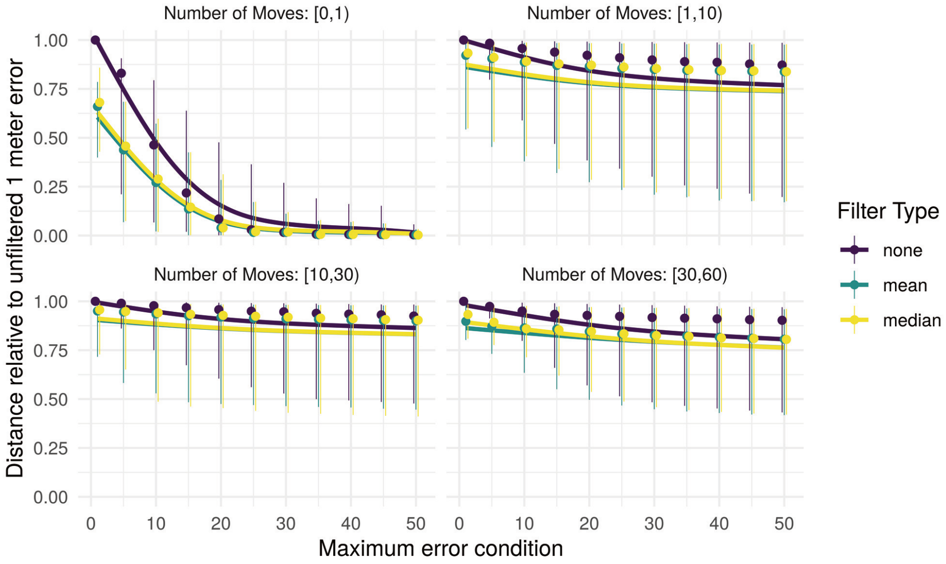


Figure 3.4 *Distance comparison for max error tolerances across differing number of true moves*

son has entered. While there is a clear negative relationship between the maximum error and the relative total distance, it is non-linear. Figure 3.4 demonstrates this complexity. We expect a very small amount of true movement when a person is stationary, so an appropriate error parameter in this case is one that reduces the relative distance to zero. An ”elbow” at the error tolerance of 20 meters in the stationary condition indicates a bottoming-out of noise-reduction. Higher error parameters would reduce this number further, but at the cost of perhaps erroneously reducing the distance during true movement behavior. The median relative distance in the true movement cases at an error parameter of 20 meters is approximately 90%, which aligns with previous findings (Palmer, 2008; Ranacher et al., 2016).

3.3.2. Simulation study design

Data

The set of data $q_{ij}|\tau_5 < .05$ was divided into individual 24 h periods for improving comparison between users. A user with one four-day contiguous set afterwards had four sets of 24 h, with any remainder discarded. The 24 h period began from the first measurement for which all subsequent measurements in the period had no gaps greater than 5 min. The raw data were subsequently cleaned according to the stop detection protocol implemented in the original mobility app. Data were retained if the estimated accuracy provided by Android or iOS was under 80 meters. It was necessary to select a sufficiently large accuracy in order to incorporate data acquired via Wi-Fi triangulation on iOS as this defaults to 65 meters. Selecting a suitably low accuracy effectively removes cell tower-based locations and locations for which there are an insufficient number of navigational satellites visible to establish a reliable position. Before introduction of missing data, $\bar{q} = .001$, with a range of 0–$.03$. 187 users were retained, representing 362 contiguous periods, broken into a total of 584 24 hour periods. The mean number of periods per user was 3.12, with a range of 1–25.

Simulating short gaps

In order to assess the impact of increasing levels of sparsity generated by small gaps, the first study introduced missingness to the data (completely) at random. This represents a situation in which the missingness was not functionally related either to mobility or the user. Each period was divided into 288 five-min time intervals (i.e $\tau = 5min$). Sparsity was introduced at ten percent intervals, ranging from $q = 0$, where no data were removed, to $q = .9$ in which 90% of the five minute intervals were excluded. For each period, this process was repeated 20 times at each q to allow for different portions of the data to be removed. This led to 180 versions of each set with varied patterns of missingness. Each version was linearly interpolated across gaps and segmented, followed by calculation of the outcome measures. Algorithm 2 describes the steps in detail.

Algorithm 2 Algorithm for short gap simulation study

```
for all Set[n] do
    for all $q \in \{0, .1, .2, ..., .9\}$ do
        for all $i \in \{1, ..., 20\}$ do
            Sample without replacement $q$ proportion of intervals and exclude
            Set[n] $\leftarrow$ interpolate(Set[n])
            Set[n] $\leftarrow$ resolveStops(Set[n])
            Merge adjacent stops with centroids less than 100m distant
            Calculate aggregate measures on moves and stops
            if number of move states $> 0$ then
                for all  move states $\in$ Set[n] do in parallel
                    topDownTimeRatio(move states)
            end if
            topDownTimeRatio(Set[n])
            Record aggregate measures
        end for
    end for
end for
```

Simulating long Gaps

Functionally, short gaps reduce the overall density of trajectories while maintaining overall mobility characteristics. Long gaps at the same overall level of sparsity should induce a different pattern. Long gaps at increasing levels of sparsity are likely to remove whole trips and thereby decrease movement distance and RoG, which should meaningfully distort travel metrics at lower q than short gaps.

A simulation study was designed in order to test this assumption. q was induced at ten percent intervals, as in Section 3.3.2. Instead of removing data in 5-min intervals at random, a starting point was selected in the data, after which locations were removed in 2.4 h intervals, representing one tenth of a 24 h day and thus corresponding to each level of q. In order to investigate the temporal characteristics, this process was carried out 24 times for each data set, selecting a starting point for each of the recorded hours. For iterations $i > 2$, some q would reach the end of the 24 h period, in which case further removal started from the beginning of the 24 h period. This process is detailed in Algorithm 3.

Algorithm 3 Algorithm for long gap simulation study

```
for all Set[n] do
    for all $hour \in \{1, ..., 24\}$ do
        for all $q \in \{0, .1, .2, ..., .9\}$ do
            Remove $q$ proportion of intervals starting from hour
            if intervals to be removed extend past data set end then
                Remove remainder from beginning of data set
            end if
            Set[n] $\leftarrow$ interpolate(Set[n])
            Set[n] $\leftarrow$ resolveStops(Set[n])
            Merge adjacent stops with centroids less than 100m distant
            Calculate aggregate measures on moves and stops
            if number of move states $> 0$ then
                for all  move states $\in$ Set[n] do in parallel
                    topDownTimeRatio(move states)
            end if
            topDownTimeRatio(Set[n])
            Record aggregate measures
        end for
    end for
end for
```

Short gap sensitivity analysis

Analysis of the results from the short gap and long gap simulation studies indicated that gap length was important independent of the total sparsity in the data. We hypothesized that it should be possible to use the time length of the gap in order to discriminate between two types of missing data: those that can be ignored and solved with linear interpolation, and those that cannot be, terming them ”short” and ”long” gaps, respectively.

A third simulation study was conducted in order to provide a more in-depth look at the variation when gap lengths vary from 1 min to 15 min. For each complete data set, gaps were created between 1 and 20 min in length in increments of 1 min, in each of the 24 h in the data set. The range of q calculated after removal fell between $.02$ in the case where 1 min was removed from the start of each hour and $.33$, where 20 min were removed from the start of each hour. This is further described in Algorithm 4.

Algorithm 4 Algorithm for short gap sensitivity simulation study

```
for all Set[n] do
    for all $m \in \{1, ..., 20\}$ do
        Remove $m$ min from the start of each of 24 hours
        Set[n] $\leftarrow$ interpolate(Set[n])
        Set[n] $\leftarrow$ resolveStops(Set[n])
        Merge adjacent stops with centroids less than 100m distant
        Calculate aggregate measures on moves and stops
        if number of move states $> 0$ then
            for all  move states $\in$ Set[n] do in parallel
                topDownTimeRatio(move states)
        end if
        topDownTimeRatio(Set[n])
        Record aggregate measures
    end for
end for
```

3.4. Results

3.4.1. Simulation study

Across all three simulation scenarios, we investigated the impact of linear interpolation as a method for addressing gaps in the data. Outcome variables of interest were chosen to represent variables commonly used in mobility research and included total distance, total travel distance, number of stop and move instances, RoG, and total move time. The method of calculation for these metrics is described in Section 3.2.3.

Because the absolute metrics may differ up to an order of magnitude between persons, comparisons were performed by evaluating the percentage difference of the metric in the interpolated data set as compared to the metric in the complete data set. Figures 3.5, 3.7, 3.6, and 3.8 group these percentages by box plot across the differing levels of sparsity in order to provide a robust summary of features of the distribution. The box spans between the upper and lower quartile values of the statistic, and the horizontal lines extends from these values through to 1.5 times the inter-quartile range from the median. The median is represented by the horizontal line through the box. Values extending beyond the horizontal lines are represented by individual points. Tables and figures evaluating percentage difference in movement behavior excluded participants registering no moves in their complete data, and all figures and tables excluded persons registering less than 200 meters of total movement behavior.

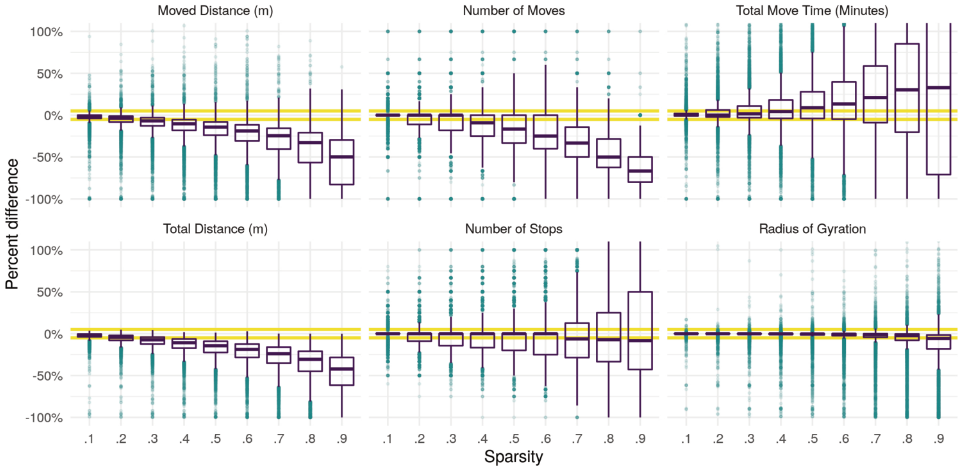


Figure 3.5 *MCAR short gap analysis*

Table 3.2 *MCAR short gap: Median (%) absolute differences by induced sparsity*

q	Travel Distance (km)	RoG	Stops
0.1	-0.4 (-1.2%)	0 (0%)	0 (0%)
0.2	-1 (-3.8%)	-0.4 (0%)	0 (0%)
0.3	-1.7 (-6.8%)	-1.4 (-0.1%)	0 (0%)
0.4	-2.6 (-10.4%)	-3.2 (-0.1%)	0 (0%)
0.5	-3.6 (-14.3%)	-6.8 (-0.3%)	0 (0%)
0.6	-4.9 (-18.9%)	-13.7 (-0.6%)	0 (0%)
0.7	-6.3 (-24.6%)	-26.3 (-1.1%)	-1 (-6.2%)
0.8	-8.4 (-32.7%)	-54.1 (-2.3%)	-1 (-7.1%)
0.9	-11.6 (-49.8%)	-138.8 (-5.9%)	-1 (-8.3%)

Short gaps

The data with induced missingness were compared to the complete data set in order to evaluate the impact on metrics of interest. Table 3.2 and Figure 3.5 show the decrease in moved distance and number of stops with increasing sparsity. At 30% sparsity, the mean distance retained is almost 90%, and the median distance retained is 93%. Only as sparsity levels exceed 60% does the median distance lost reach 20%. Similarly, filling gaps through linear interpolation fails to impact the median number of stops until sparsity reaches 50%. In fact, these short gaps become problematic only when they become long gaps, as two or more adjacent short gaps merge into one.

Figure 3.5 shows a relationship between distance metrics and sparsity that may be predictable in aggregate. Individual response demonstrates a considerable amount of variance – while it may be possible to predict the percentage of distance lost,

Table 3.3 *MCAR long Gap: Median (%) absolute differences by induced sparsity*

q	Travel Distance (km)	RoG	Stops
0.1	0 (0%)	0 (0%)	0 (0%)
0.2	-0.2 (-0.7%)	-0.1 (0%)	0 (0%)
0.3	-2.6 (-12.7%)	-1.6 (-0.1%)	0 (0%)
0.4	-6.4 (-36%)	-74.8 (-4.9%)	-1 (-14.3%)
0.5	-10.6 (-51.4%)	-205.4 (-14.3%)	-2 (-28.6%)
0.6	-15.1 (-74.3%)	-404.5 (-23.6%)	-3 (-42.9%)
0.7	-19.6 (-100%)	-711.9 (-52.9%)	-3 (-60%)
0.8	-22.8 (-100%)	-1304.9 (-94%)	-4 (-69.2%)
0.9	-25.8 (-100%)	-1903 (-99.2%)	-5 (-78.7%)

number of travel behaviors, or total transit time based on the available data and sparsity, the uncertainty as q exceeds $.3$ in a naïve prediction would lead to confidence bands extending, in some cases, from a 50% increase to a 99% decrease.

Long gaps

The results of the second simulation study, designed to investigate whether or not the same method of linear interpolation worked for gaps of increasing length, can be seen in Figure 3.6 and a brief summary of median results can be found in Table 3.3. An increase in q to .3 accompanies a downward bias of 12.7% in median travel distance. Median RoG remains relatively stable through a $q = .4$, whereas the number of recorded stops becomes unstable with $q > .3$.

Figure 3.6 demonstrates wide variability in response from individuals across all metrics. Removal of 50% data contiguously could remove the entirety of one respondent's travel behavior within a day, while leaving all travel behavior intact for another.

Figure 3.7 shows the relationship between increasing sparsity by inducing multiple smaller gaps versus increasing sparsity through increasing the length of a single gap. In all situations, long gaps demonstrate a more extreme departure from the ground truth than short gaps at the same levels of overall total missingness.

Short gap sensitivity analysis

Table 3.4 shows some results from this sensitivity analysis. Depending on the level of acceptable data loss, it is possible to establish a maximum gap length from this table of results. For example, if loss of under 2% travel distance is desired, gaps under 5 min in length may be acceptably handled by linear interpolation.

Some differences emerge between the three simulation studies. In Section 3.4.1, an increase in total missingness was associated with a decrease in RoG, whereas removal of 1-15 min per hour, equivalent to a range of q from $.01 - .25$, results in a small positive increase. This is attributable to the simulation design in which

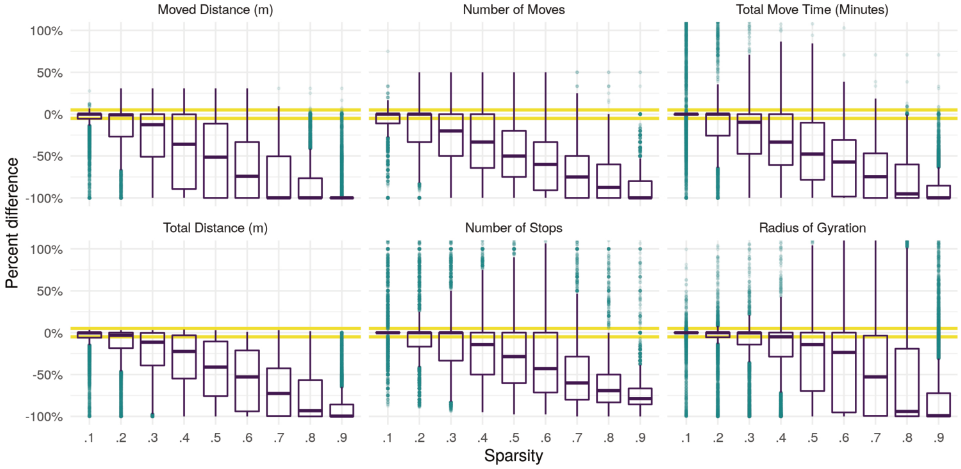


Figure 3.6 *MCAR long gap analysis*

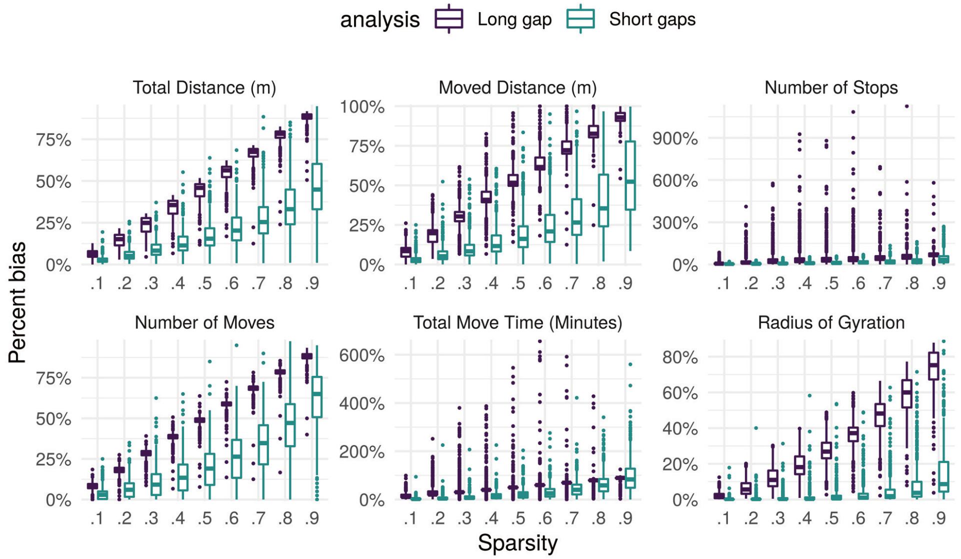


Figure 3.7 *Percent bias in calculated mobility metrics relative to sparsity*

Table 3.4 *Short gap sensitivity: Median (%) absolute differences*

Min Removed/Hr	Travel Distance (km)	RoG	Stops
1	0 (0%)	0.3 (0%)	0 (0%)
2	-0.1 (-0.4%)	0.5 (0%)	0 (0%)
3	-0.2 (-0.6%)	0.6 (0.1%)	0 (0%)
4	-0.3 (-0.9%)	0.8 (0.1%)	0 (0%)
5	-0.4 (-1.3%)	1.1 (0.1%)	0 (0%)
10	-1 (-4%)	0.8 (0.1%)	0 (0%)
15	-1.8 (-7.4%)	0.2 (0%)	0 (0%)
20	-2.9 (-10.8%)	-2.4 (-0.2%)	0 (0%)

minutes were removed from the beginning of each 60 min period, starting with the first location entry. As users frequently engaged with the app for the first time while at home, this results in a small upwards bias of RoG, since a slightly higher proportion of home locations were removed. Additionally, gaps occurring at stop-move boundaries inhibit accurate determination of movement initiation, leading to a larger proportion of time associated with movement behavior on average and consequently a larger time-weighted RoG.

Figure 3.8 shows the relationship between the percentage difference on calculated mobility measures between the complete data and the data with induced missingness. Some metrics, such as number of stops and moves, or the distance traveled, respond well through gap sizes of 10 min. Other metrics, such as Total Move Time, become quickly unreliable, even with very short gaps.

Importantly, while the point estimates may remain stable across all simulation study methodologies, the variance is considerable, with any one gap being more or less impactful depending on the sampled underlying behavior.

3.4.2. Covariates

Finally, we sought to establish a set of covariates that could impact the relationship between q and bias remaining after interpolation. Differential response across covariates could allow for extending the boundaries of what we consider the maximum acceptable gap time. On the other hand, similar response profiles can assure that no additional bias is introduced when filling short gaps through linear interpolation.

Mobility characteristics

The metrics of interest had a small impact on the percent bias introduced through interpolation. Figure 3.9 shows the mean bias of total distance, number of moves and RoG following interpolation across gaps at varying levels of sparsity under the long gap simulation condition. All three metrics demonstrate slightly lower bias

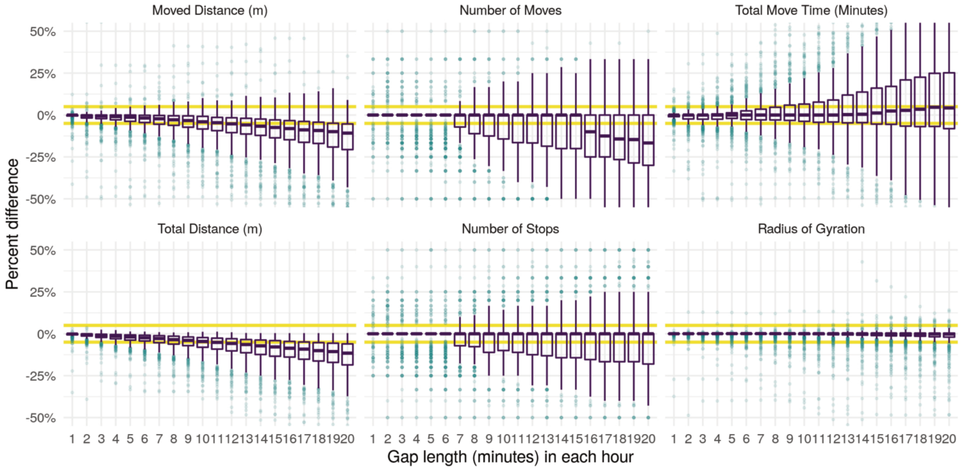


Figure 3.8 *Sensitivity simulation for small gaps*

when calculated in data sets where the true distance and number of moves was lower. We find the inverse relationship with true RoG in total distance and RoG estimation on the interpolated data sets. Across all three metrics, the absolute percent bias at $q < .1$ is low across varied true trip characteristics.

Personal characteristics

The Dutch population register contains basic information on all individuals living in the Netherlands through their registration with the municipality. This information was linked to the users who had participated in our study. Percent bias in travel metrics established in the long gap study were compared across several individual covariates. A full set of results are made available in Appendix B, Section B.2. The data establish minor relationships between covariates age, education and urbanicity and induced bias, but the overall effect sizes are small. Age, education and urbanicity are associated with differential mobility characteristics which demonstrate a stronger relationship with with percent bias in the travel metrics, potentially driving this relationship.

Time

Investigation of time as a metric was considered important, as both Android and iOS operating systems employ mechanisms for reducing device activity during times of lesser activity levels, contributing to long gaps during nights that are unlikely to con- tain travel behavior. The results from the Long Gap simulation study discussed in 3.4.1 provided a way to investigate the relationship between relative error and the time of day during which the gap occurs.

As shown in Figure 3.10, hours between 21:00 and 04:00 produce unbiased estimates of total distance, number of moves and RoG even when gap length exceeds

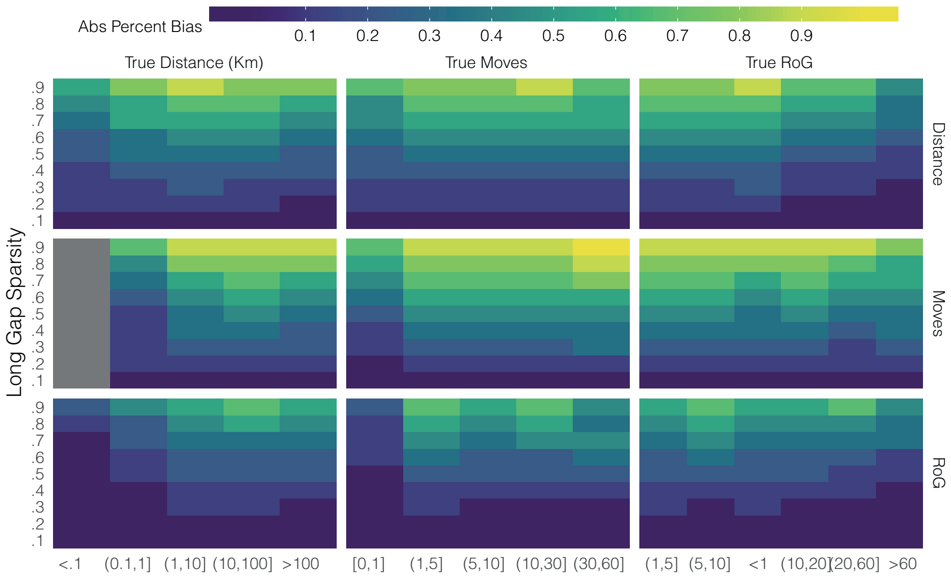


Figure 3.9 *Bias with respect to underlying mobility characteristics*

5 h. The low variance during these time periods suggests that overnight missingness may be appropriately resolved through unsophisticated methods that rely upon infrequent nighttime travel behavior.

Interestingly, there appear to be pockets during daytime hours that can bias results up to 25% with relatively short periods of missingness. These correspond with time periods in which people are more likely to engage in travel behavior, such as commuter traffic, with a distinct morning phase and diffuse evening phase. Missing data around noon is more likely to bias recorded move events than distance traveled or RoG reflecting mobility patterns shorter both in duration and distance that occur during this time.

3.5. Conclusion

Understanding the composition of the missing data is integral to making the correct decisions about its content. The composition can involve the length of the component gaps, the overall sparsity of the data, or the time at which the gaps begin or end. Working smartly within these boundaries, researchers can extend the use of data that might otherwise be excluded from analysis for being incomplete. In situations where the gaps are short – between five and ten min – even if they occur frequently, very little can perturb aggregate measures such as distance or RoG. This makes linear interpolation an acceptable solution for gaps caused by interrupted line-of-sight instances. More intensive methods of addressing these short gaps, such as map-matching or imputation, are unlikely to offer large gains in these situations and

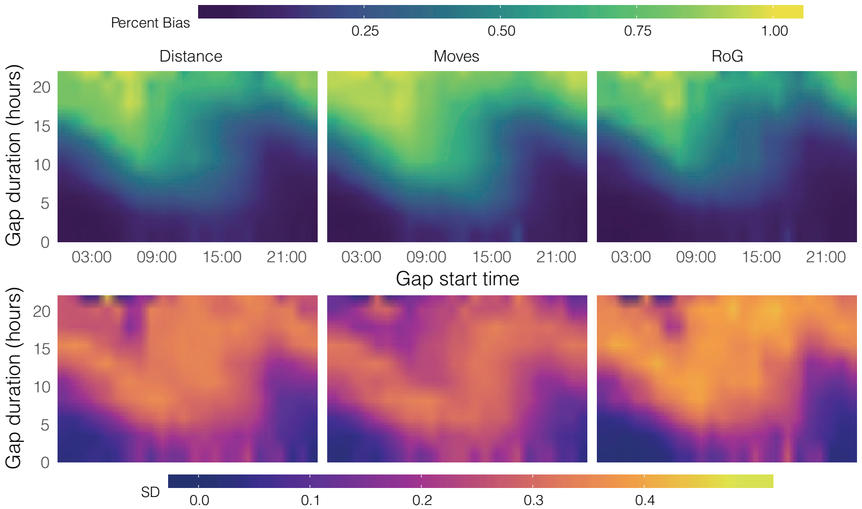


Figure 3.10 *Bias with respect to gap duration and time of gap*

the added complexity may hinder efforts to address the larger gaps.

We are limited to the metrics considered within this paper in our discussion of the consequences. One metric of great importance in mobility research yet not included in these analyses is the employed mode of transportation during move events. It may be possible that interpolation even across very short gaps has a negative impact on the prediction accuracy of mode of transportation. Additionally, although the metrics investigated within these simulations are shared across many fields collecting data on movement behavior, it is unlikely that our results will extend beyond data that is collected within the context of individual mobility.

It is clear that linear interpolation is a poor fit for addressing most long gaps. Radius of gyration and number of stops may see minimal impact if only portions of a trip are lost, as may be the case when a phone's battery dies en route and is then charged at the destination. However, the same situation would almost certainly result in a downward biasing of distance metrics using the same method. The uncertain impact on any individual's travel behavior necessitates the incorporation of data beyond the spatiotemporal aspects of the gap itself.

So what are the implications for analysis of time-location data in a travel survey? Our study shows that if the gaps don't extend beyond 10 min in length, and if they are relatively infrequent, say below 15%, biases in main travel statistics can be acceptable. Or seen from the other side, when setting requirements on the coverage of confidence intervals for travel statistics, our study gives insights into what gap characteristics may be problematic.

Methods of addressing these long gaps are documented in the literature. Imputation with a user's longitudinal data or with densely overlapping spatial data from other users are both promising methods using larger and longer data sources to account for the information lost. Although often users may contribute sufficient data to contextualize their own gaps, the longer the gap, the less data is otherwise available for this con text. And while users in densely populated urban areas may contain sufficient overlap in their trajectories to aid in completing other users' data, this is far from applying to all such collection opportunities, where it may be of interest to distribute a limited budget such that a wider range of geographical areas may be covered. External data sources providing information on the individual such as frequented addresses, or integrated surveys on car- and bike-ownership are used to provide context for estimation. Land-use characteristics and transportation infrastructure data feature in other long gap methodologies. Additionally, smart phones sensors like accelerometers and gyroscopes can be used to identify movement and travel activity and may additionally be of use in both contextualizing and handling missing data.

A clear path emerges for future research. Researchers need a comprehensive plan for addressing short and long gaps in the most contextually appropriate manner. This paper proposes that linear interpolation is an acceptable manner for short gaps, and defines short gaps as less than 10 min in length for the study of human mobility. Existing methods for addressing long gaps universally incorporate larger data and varied information sources. For researchers, however, the decision on which methods are appropriate for their particular solution remains opaque. The integration of multiple approaches using all available resources is a direct if not uncomplicated next step. As GPS technology improves, researchers can expect some challenges of location data to fall away. As the number of positioning satellites increases, and technology for triangulation improves, the accuracy of individual locations will certainly improve to create trajectories with less noise. New generations of satellite systems and receiving chips purport sensitivities that can travel through walls, reducing some line-of-sight based causes of missing data. But many causes of missing data are likely to remain for years. Battery capacity has improved year-on-year, but battery life has not as the demand for what a mobile phone must do has grown concurrently. Android versions are becoming more rather than less likely to kill an app, and iOS devices remain similarly opaque on when they are allowed to perform operations in the background. While undoubtedly a positive trend for users, a growing focus on user privacy may indirectly impact researchers as users are not aware of the switch to opt-in versus opt-out location options on their device.

Missing data is unlikely to be a solved problem for researchers in the near future androbust methods of addressing the missing data are integral to unlocking the potential of this new technology.

4

Dynamic Time Warping-based imputation of long gaps in human mobility trajectories

This chapter is currently available as a preprint McCool, D., Lugtig, P., & Schouten, B. (2024). Dynamic Time Warping-based imputation of long gaps in human mobility trajectories. In arXiv [stat.ME]. arXiv. http://arxiv.org/abs/2410.16096 (McCool et al., 2024).

Abstract

Individual mobility trajectories are difficult to measure and often incur long periods of missingness. Aggregation of this mobility data without accounting for the missingness leads to erroneous results, underestimating travel behavior. This paper proposes DTWBMI *as a method of filling long gaps in human mobility trajectories in order to use the available data to the fullest extent. This method reduces spatiotemporal trajectories to time series of particular travel behavior, then selects candidates for multiple imputation on the basis of the dynamic time warping distance between the potential donor series and the series preceding and following the gap in the recipient series and finally imputes values multiple times. A simulation study designed to establish optimal parameters for* DTWBMI *provides two versions of the method. These two methods are applied to a real-world dataset of individual mobility trajectories with simulated missingness and compared against other methods of handling missingness.* LI *outperforms* DTWBMI *and other methods when gaps are short and data are limited.* DTWBMI *outperforms other methods when gaps become longer and when more data are available.*

4.1. Introduction

The mobile device as a new source of data for researchers has promised returns across a variety of fields including human geography, transportation and mobile health (Rout et al., 2021). These gains have yet to materialize due to a lack of generic resources available to handle the frequent data quality issues, especially with respect to missing data (Bähr et al., 2022; Beukenhorst et al., 2022). Studies using standalone GNSS tracking devices have demonstrated the usefulness of trace data in collecting activity data, in predicting wandering episodes in adults with dementia and in investigating travel behavior (Furletti et al., 2013; Wojtusiak & Mogharab Nia, 2021; Zheng et al., 2008).

Researchers presupposed a natural progression from these standalone devices to the use of respondents' own devices as an economic and flexible alternative. This is within the technical capacity of the smartphone, but the reality of this data collection methodology is that it is highly prone to missingness (Harding et al., 2021; Yoo et al., 2020). In fact the same feature that makes it compelling for researchers – collection of large amounts of data over time – becomes a significant drawback when the data are incomplete. The temporal nature of these data make straightforward imputation of the exact missing locations an intractable problem. In this paper, we propose and evaluate an alternative method for imputation of longer sections of missing location data in travel surveys. DTWBMI uses aggregate statistics derived from continuously measured locations to impute travel metrics during the gap, allowing full use of the information that precedes and follows the gap.

Data can be missing for many reasons, some related to the physical surroundings and thus common to all GNSS measurements, and others related to the device, the user, or the interaction between the two (Bähr et al., 2022; Ranasinghe & Kray, 2018). The first category contains such mechanisms as line-of-sight problems, cold starts and urban canyons (Karaim et al., 2018). These tend to produce missingness that is consistent, occurring either over a limited geographic range or within a relatively brief period of time. On the other hand, the second category varies significantly between persons and their devices (Beukenhorst et al., 2021). Storing the location data requires that an app must be running on the smartphone, the smartphone must be on, and it must have the necessary permissions to record the data (Keusch, Wenz & Conrad, 2022). Both Android and iOS operating systems, in an effort to increase battery life on their devices, will often shut down apps indiscriminately, including those applications recording data for research purposes. This is a primary cause of missing data collected from smartphones (González-Pérez et al., 2022). In addition to this largest hurdle, apps must also contend with the user closing the app, turning off the phone or disabling the location services on the device. This can occur unintentionally, but may occasionally be due to privacy concerns (Kreuter et al., 2020) or concerns about battery life. Lastly, the battery may simply become drained, ceasing all location collection until the device is turned on again.

Developing strategies to handle the missing data is as necessary as it is challenging

(Moffat et al., 2007). Studies have demonstrated that the patterns of missing data are more often informative than not (Bähr et al., 2022; Mennis et al., 2018). Consequently, deletion of those persons or days with missing data can produce biased results in subsequent analyses (Hawthorne & Elliott, 2005; Honaker & King, 2010). Neither is it appropriate to interpolate across gaps of any appreciable size, leading to a considerable rate of underestimation as gaps lengthen to include whole trips (McCool et al., 2022; Phan et al., 2018) and up to a 10-fold increase in error variance computed on the metrics of interest (Barnett & Onnela, 2020). It is possible to aggregate the data to days or weeks and then to apply more traditional forms of missing data handling. This is the method most commonly employed by accelerometer studies (Stephens et al., 2018). However, this loss of granularity restricts the level of analysis to the aggregate level. More sophisticated methodologies are necessary to simultaneously maintain the useful characteristics of the data as well as allow for unbiased handling of missingness (Onnela, 2021).

These sophisticated methodologies do exist in the literature, but so far none have demonstrated the capacity to solve long gaps in the data without problems. For example, k-Nearest Neighbor (kNN) methodologies are often employed to impute gaps in this way, but tend to insufficiently consider the time aspect of univariate time series (Sun et al., 2017). Time-Delayed Deep Neural Networks have recently sown some promise for the imputation of long gaps, but currently consider only a single gap in data supported by related variables (Park et al., 2022).

Efforts to reconstruct full sets of potential paths with adequate coverage have been considered recently which may prove interesting, but may also lead to huge sets of possibilities when considering long gaps in human mobility (Parrella et al., 2021). Random walks may be used to generate sequences of behavior that may be used for gap-filling across long gaps, but current methodologies are still in development (Dekker et al., 2022).

Geographic or time constraints can help to identify some trajectories that are more likely, and to narrow the number of possible tracks. However, this requires additional data that limits their applicability in the general situation. For example, we may make use of the repetitive nature of human behavior and the fact that two weeks tend to contain a person’s primary activity spaces (Stanley et al., 2018; Zhu et al., 2022) if we extend data collection over a long enough temporal period to achieve coverage over the missing data periods (Chen et al., 2019; Dhont et al., 2021). Or, if a geographical constraint can be applied, routes that make use of public transportation or common road structures may be shared with a high frequency. Users may then have overlapping traces allowed for individuals to complete each others’ records with a high degree of precision.

External data has also been shown to provide additional benefits in the handling of missing data. Map-matching may provide a method for realistically imputing travel patterns provided that the gap is short enough that it can account for both the start and stop locations (Jagadeesh & Srikanthan, 2017; Knapen et al., 2018; Tanaka et al., 2021) or if the missed locations can be provided after the fact by users or by data donation (Boeschoten et al., 2020; Hollingshead, Quan-Haase et al., 2021; Keusch &

Conrad, 2022; Silber et al., 2021). It is possible that data on mode of transportation is useful in filling long gaps where the travel behavior depends strongly on the mode, but asking for this information increases the burden on respondents and is prone to error. Importantly, each of these methodologies requires specific additional data or constraints that are unlikely to be compatible with the low density and short observation periods common for travel studies.

Often there may be no helpful relationships within the data to allow for the prediction of missing data beyond what patterns exist within and between individuals in the recorded data. Dynamic Time Warping is a method used to find similar areas within two time series, which we propose as a selection mechanism for multiple imputation candidates on the basis of the location data itself. This paper 1) presents DTWBMI, a new generically applicable methodology for use in this scenario, and 2) evaluates its performance against other applicable methodologies.

The next section will describe and outline the integral elements of this methodology, including extraction of metrics of interest as time series, the calculation of Dynamic Time Warping (DTW) distances between these time series, followed by multiple imputation using this distance measure to select candidates for donation of their time series, and a description of Dynamic Time Warping-Based Imputation (DTWBI). We describe the history of this methodology and outline our adaptations of the method for use in the imputation of geolocations over time. Section 4.3 describes data preparation and a simulation study in which the methodologies were evaluated. Section 4.4 evaluates the relative efficacy of DTWBMI as compared with Dynamic Time Warping Based [Single] Imputation (DTWBI), interpolation, aggregation, Time Window imputation and mean imputation. Finally in Section 4.5, we discuss relevant issues regarding the procedures presented in this study.

4.2. Background

DTWBI was developed for more general use in the imputation of time series data. Because the method makes use of patterns within the temporal characteristics of the data, it is useful to evaluate its potential as a mechanism for correcting for long gaps in trajectory data. In this section, we describe the building blocks of this process.

4.2.1. Key metrics in travel data

Raw GPS data is comprised of location coordinates that can be used to segment the day into stops and trips, and to calculate various metrics of interest such as distance traveled. In comparison, travel diaries lack this granularity of location information, usually containing only start and end locations for trips. Travel diary data is therefore limited to count, time, distance or travel mode statistics. Aggregation often occurs at day- or week-level in these cases, and general statistics estimated on the basis of the sample data.

Location data collected via a smartphone under ideal settings generate traces representing the raw data of a person's location at a given point in time. When the goal

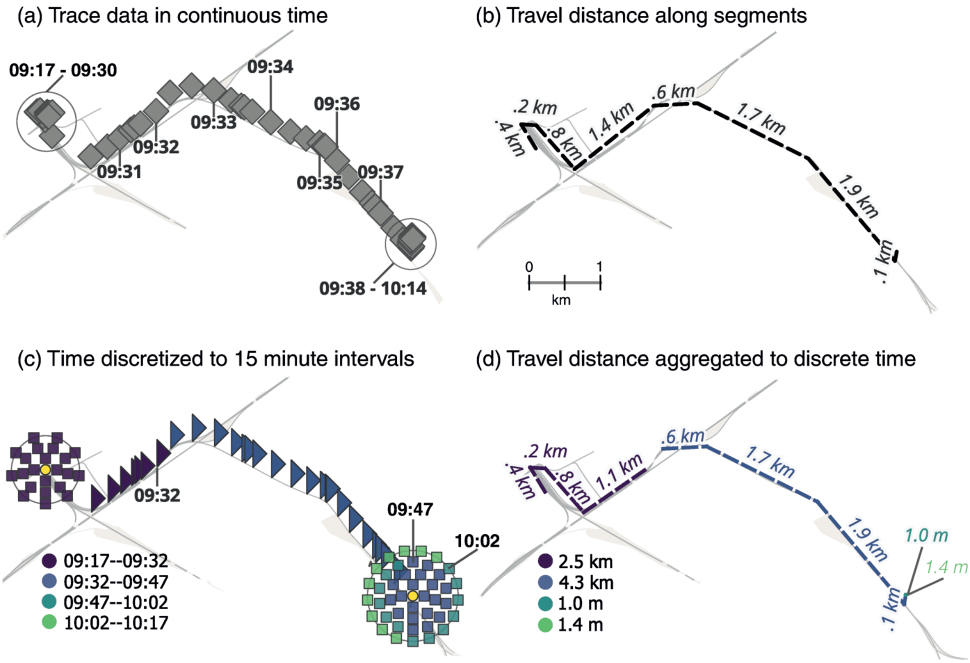


Figure 4.1 *Aggregation of travel distances into 15-minute intervals. (a) Timestampped geolocations with stay points (gray circles) and connecting track. (b) Segment-wise travel distance estimation. (c) Geolocations categorized into 15-minute segments, with stay points expanded for clarity; cutoffs are marked which occur during the track (09:32) and stay point (09:47, 10:02). (d) Segments recalculated within each time interval to sum travel distances, forming a discrete time series appropriate for our analyses.*

is to yield similar metrics, researchers must process these raw data by calculating the distance, defining stops and trips and establishing a mode of transportation.

Segmenting the trajectories into stops and trips without user input requires selection of an algorithm. In this study, we use an algorithm which defines stops on the basis of a radius and time parameter as described in Montoliu et al. (2013). In this study, we apply Top-Down Time Ratio (TDTR) segmentation on the raw data, summing the distances of the segments (McCool et al., 2022; Meratnia & de By, 2004). These two steps allow for the calculation of count, time and distance statistics at the desired level of aggregation.

4.2.2. Time Series

It is possible to generalize individual trace data by considering it as a time series of metrics of interest common between individuals. This removes the specific geographic information associated with each location. A person's travel behavior may be described as a time series representing distances traveled over a discrete period or the increase over time of the total covered area. Figure 4.1 shows an example in which trace data are discretized to 15 minute intervals and distance is aggregated within these periods to produce a four-element time series of distances. Such metrics can be calculated across all users to describe aspects of travel behavior against which different persons can be compared to find commonalities. This process involves two important considerations, namely the selected metrics and the way in which time is discretized.

Which metrics are selected should reflect the final analyses that the researcher intends to employ. In this paper, we consider a selection of key metrics typical to travel diaries, including Travel Distance (TD), RoG and Total Stops (TS) as discussed in Section 4.2.1. This method is extensible to a wide variety of metrics, including categorical measures such as transportation mode.

A secondary consideration involves the discretization of continuous time. Metrics may be aggregated across intervals reflecting increments of a given temporal resolution. As the temporal resolution changes, the time series will lose detail which will decrease the potential specificity in the matching process. The temporal resolution should be of sufficient length to ensure that there are data within most aggregated blocks containing complete data.

Using the notation from (Phan, Poisson Caillault, Lefebvre & Bigand, 2020), we describe a univariate time series x as a series of N measurements or observations indexed by discrete time t, as shown in Equation 4.1:

$$x = \{x_t \mid t = 1, 2, \dots, N\} \tag{4.1}$$

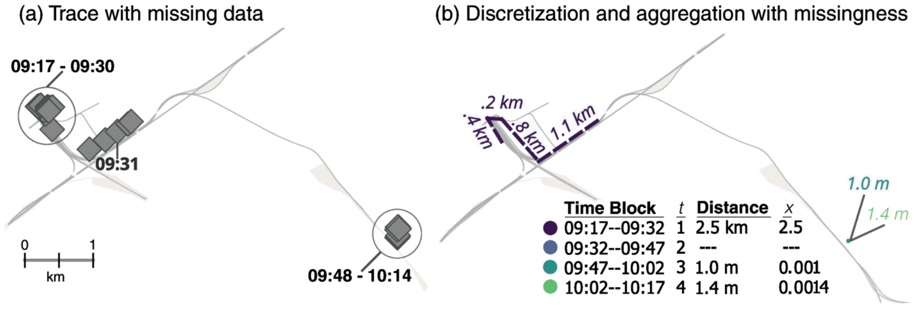
Using the example provided in Figure 4.1, we would reduce this trace to a time series x comprised of $N = 4$ elements representing kilometers traveled during each time period t: $\{2.5, 4.3, .001, .0014\}$.

In instances where x contains missing measurements, some time points t will lack observations. We introduce a binary response variable r_t, indicating the presence (1) or absence (0) of data at time t. Missing data can manifest as isolated instances or span multiple time points. For a given time index t, if $r_t = 0$ and $r_{t-1} = 1$, then t denotes the start of a gap. Let s be the smallest index greater than t such that $r_s = 1$, marking the gap's end. The length of the gap T beginning at time t is then given by Equation 4.2:

$$T^t = s - t \tag{4.2}$$

Figure 4.2 demonstrates the conversion of the original trace to a time series x with missing data occurring during time period $t = 2$. In this case, the time series x_t would be represented by $\{2.5, -, 0.001, 0.0014\}$ and the response series r_t by $\{1, 0, 1, 1\}$. The length of the gap T is calculated as $3 - 2 = 1$.

We extend both the time series x and the response variable r to multiple persons indexed by k, resulting in x^k and r^k respectively. We can further generalize to a matrix X^k representing multiple variables extracted as time series from an individual's data,


(a) Trace with missing data
09:17 - 09:30
09:31
09:48 - 10:14
0
1
km
(b) Discretization and aggregation with missingness
.2 km
.4 km
.8 km
1.1 km
1.0 m
1.4 m

	Time Block	t	Distance	x
●	09:17--09:32	1	2.5 km	2.5
●	09:32--09:47	2	---	---
●	09:47--10:02	3	1.0 m	0.001
●	10:02--10:17	4	1.4 m	0.0014

Figure 4.2 *Gaps between sucessive geolocations are considered to represent missing data if they exceed the length of the chosen discretization interval. (a) No locations are recorded after 09:31 and before 09:48, representing a gap of 16 minutes. (b) Time $t = 2$ therefore covers an unknown distance, and x_2 is missing.*

reflecting various aggregated metrics of interest, analogous to a design matrix. As each series within a person retains the same missingness pattern, the response vector r^k and associated gap lengths T maintain the dimensionality of any contained x.

4.2.3. Dynamic Time Warping

In DTW, we calculate a measure of similarity between two time series by finding a path of alignment that minimizes the sum of the Euclidean distances between aligned element pairs, as described in Equation 4.1 (Sakoe & Chiba, 1978). We distinguish between these series by denoting the query series[1] as q_t of length N and the reference series[2] as ρ_j of length M. We call the ordered selected pairs of points the warping path. Figure 4.3 compares a matching schema utilizing direct Euclidean distance between ordered elements to a matching schema under DTW.

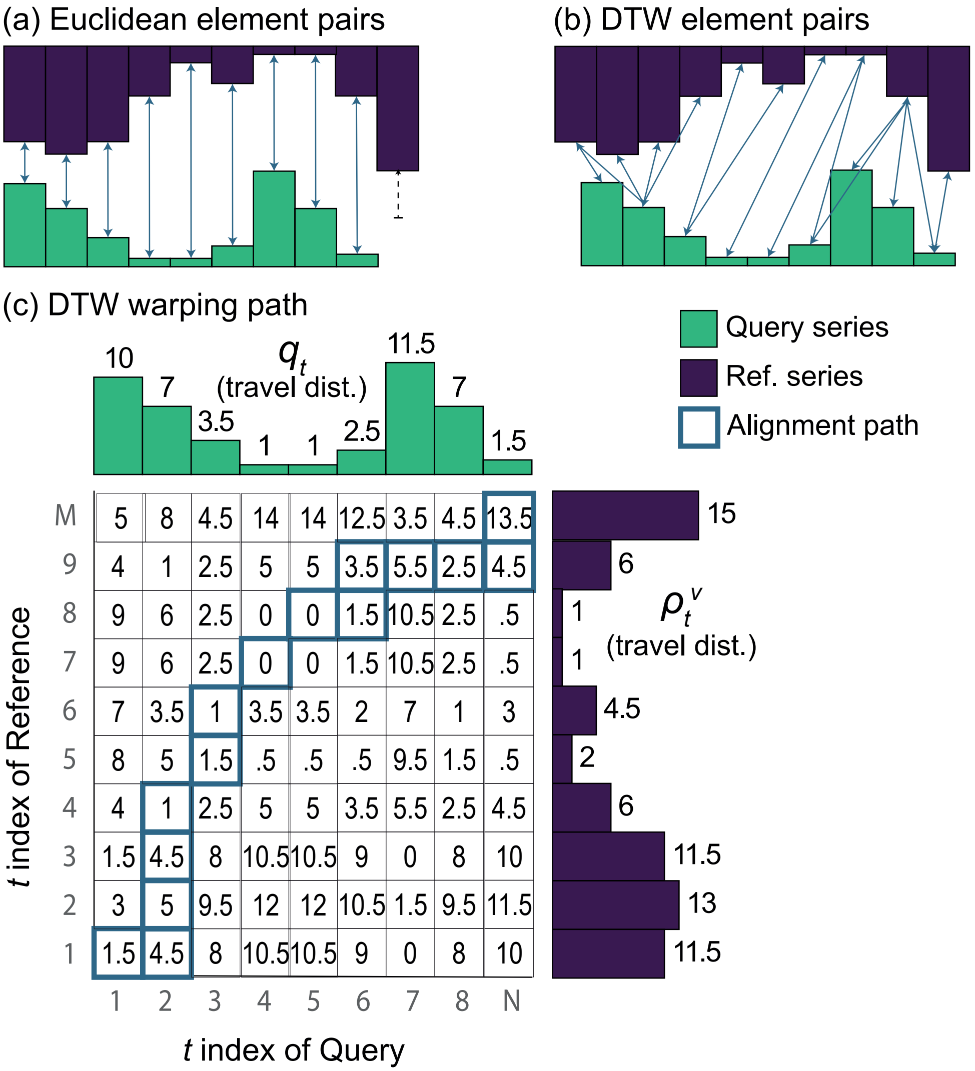
If $M = N$, calculating the Euclidean distance as shown in Figure 4.3(a) can be done in the usual way, as shown in Equation 4.3:

$$\text{dist}(q, \rho) = ||q - \rho|| = \sqrt{(q_1 - \rho_1)^2 + (q_2 - \rho_2)^2 + \cdots + (q_N - \rho_N)^2} \tag{4.3}$$

This requires both that q and ρ are the same length and also restricts the alignment to a one-to-one match with each q_t matching to a ρ_j where $t = j$. With DTW, we calculate a path of best fit that minimizes the distance between the series by allowing a disjoint matching of elements. A pairwise difference is calculated between all indices t from 1 to N and all j from 1 to M, resulting in a matrix D where each element D_{tj} is the difference between q_t and ρ_j, as shown in Table 4.1.

[1] Rabiner and Juang (1993) call this the test pattern $\mathcal{T}$.

[2] Rabiner and Juang (1993) use r to denote the singular instance, but this conflicts with our notation for the binary response variable.



(a) Euclidean element pairs

(b) DTW element pairs

(c) DTW warping path

M	5	8	4.5	14	14	12.5	3.5	4.5	13.5
9	4	1	2.5	5	5	3.5	5.5	2.5	4.5
8	9	6	2.5	0	0	1.5	10.5	2.5	.5
7	9	6	2.5	0	0	1.5	10.5	2.5	.5
6	7	3.5	1	3.5	3.5	2	7	1	3
5	8	5	1.5	.5	.5	.5	9.5	1.5	.5
4	4	1	2.5	5	5	3.5	5.5	2.5	4.5
3	1.5	4.5	8	10.5	10.5	9	0	8	10
2	3	5	9.5	12	12	10.5	1.5	9.5	11.5
1	1.5	4.5	8	10.5	10.5	9	0	8	10
	1	2	3	4	5	6	7	8	N

Figure 4.3 *Alignment of two time series representing kilometers traveled over a discretized interval of time. (a) Euclidean alignment, where each element in the query time series matches to the corresponding ordered element in the reference time series $\mathcal{R}$. (b) DTW alignment where the path of best fit is found where elements may match multiple times. (c) Cost matrix used to determine the path of best fit. The difference between each pair of elements is calculated and a path is drawn that results in the smallest total sum of differences, according to the standard DTW implementation in Equation 4.4.*

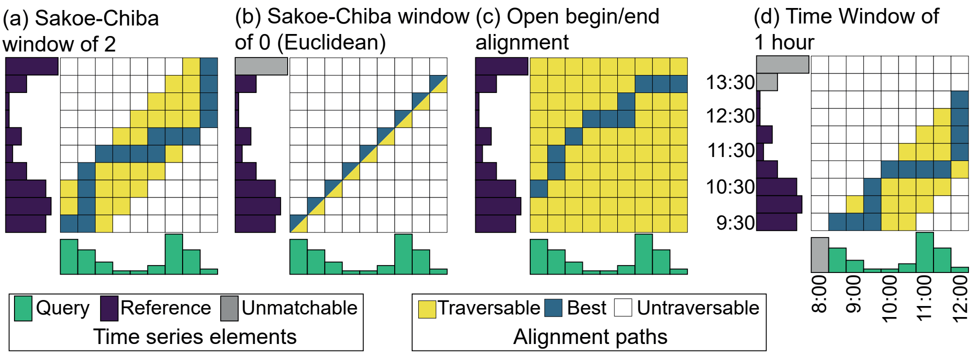


Figure 4.4 *Different alignment restrictions within DTW. (a) Sakoe-Chiba window of 2, which allows alignment of elements at most two distant in order, (b) Sakoe-Chiba window of 0, where only one path exists in a one-to-one correspondence between query and reference, (c) easing of the beginning and ending restriction where first and last elements of the query need not align to the first and last elements of the reference vector, (d) is an example of the use of time windows reflecting true time, where the possible path is restricted to a one-hour window at each element of the query series.*

Table 4.1 *Distance matrix D representing the cost of all possible routes*

	t = 1	t = 2	$\cdots$	t = N
j = 1	$dist(q_1, \rho_1)$	$dist(q_2, \rho_1)$	$\cdots$	$dist(q_N, \rho_1)$
j = 2	$dist(q_1, \rho_2)$	$dist(q_2, \rho_2)$	$\cdots$	$dist(q_N, \rho_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
j = M	$dist(q_1, \rho_M)$	$dist(q_2, \rho_M)$	$\cdots$	$dist(q_N, \rho_M)$

A path is selected through this matrix that provides the minimized cost through all traversable paths, as shown in Equation 4.4, where $D(t, j)$ gives the minimum distance cost up to point t in the query series and point j in the reference series. A path may consist of diagonal, vertical, or horizontal movements, reflected by $D(t-1, j-1)$, $D(t-1, j)$, and $D(t, j-1)$ respectively. Figure 4.3(c) demonstrates an example calculation of the path of best fit, given the cost matrix D.

$$D(t, j) = \text{dist}(t, j) + \min\{D(t-1, j-1), D(t-1, j), D(t, j-1)\} \tag{4.4}$$

In standard DTW, the first and last elements of the query and reference series must align with each other, meaning that the path must begin with $D(1, 1)$ and end with $D(M, N)$, regardless of cost. Additionally, a path may not allow backtracking (i.e., $D(t+1, \ldots)$ or $D(\ldots, j+1)$ are not allowed), and each element must be matched at least once (e.g., $D(t-2, \ldots)$ or $D(\ldots, j-3)$ to skip over one element in q or one element in ρ respectively). In addition to these constraints, it may also be desirable to restrict the path in other ways, such as requiring matched elements to be within a certain ordered distance, or window, of each other.

Table 4.2 *$\mathcal{R}$ with $V = 3$ representing 3 distinct sets for aligning on aggregated distance measurement across discrete time. ρ^1 is length $M = 5$, and ρ^3 is of length $M = 3$.*

v	j = 1	j = 2	j = 3	j = 4	j = 5
1	2 km	5 km	0 km	2 km	5 km
2	0 km	0 km	0 km	1 km	
3	1 km	0 km	2 km		

Figure 4.4 illustrates warping paths under different constraints. Figure 4.4(a) adds the constraint of the Sakoe-Chiba window of size 2. Differences are only calculated between those elements falling within 2 elements of the query series. In comparison, a Sakoe-Chiba window of size 0, shown in Figure 4.4(b), allows only one path, matching each element in the query to its ordered pair in the reference series. Imposing this restriction requires that the query and reference series are the same length, due to the restriction that the first and last elements of the query and reference series must be aligned. Easing this restriction is possible as well, as demonstrated in Figure 4.4(c), with Open Begin/End DTW or subsequence alignment (Tormene et al., 2009). All elements of the query vector must align to at least one element of the reference vector, but the reverse need not be true. When combined with a Sakoe-Chiba window of zero, this is known as the best fitting subsequence search (Sakoe & Chiba, 1978).

Figure 4.4(d) illustrates our new method, which restricts the available warping paths on the basis of a continuous-time windowing parameter. Only elements which are within the given time window are traversable by the algorithm.

Our objective is to determine the similarity between q and ρ to evaluate the fitness of ρ to serve as a donor during a gap T^t in q^k. To this end, we define a collection of V reference time series $\mathcal{R}$, as shown in Equation 4.5:

$$\mathcal{R}^V = \{\rho^1, \rho^2, \dots, \rho^V\} \tag{4.5}$$

where each ρ^v is defined as in Equation 4.6:

$$\rho^v = \{\rho^v_j \mid j = 1, 2, \dots, M_v\} \tag{4.6}$$

Table 4.2 provides an example of a univariate $\mathcal{R}$ containing aggregated distances in kilometers for $V = 3$ reference sets of varying lengths. If the only constraint applied is a Sakoe-Chiba window of 0, requiring a one-to-one match, both ρ^1 and ρ^2 can serve as reference series to the example query from Figure 4.2 because $M_v \geq N$, but ρ^3 cannot because it contains only 3 elements.

Given that DTW can only compute distances for complete observations, we split q into two segments: $q_{pre} = \{q_i \mid i = 1, 2, \dots, t{-}1\}$ and $q_{post} = \{q_i \mid i = s, s{+}1, \dots, N\}$, where t and s mark the start and end of the gap respectively. For any given q^k,

Table 4.3 *All possible fits of $\mathcal{R}$ to q under a Sakoe-Chiba window of 0. Respecting the length of q_{pre}, T, and q_{post}, ρ^1 has two possible alignments with its five elements, and rho^2 has only one possible alignment as it contains the same number of elements as q. Column T contains future potential donor values from each viable position in the ρ^v.*

	q_{pre}	T	q_{post}	
q	2.5	-	.001	.0014
$\rho^1_{j\in\{1,2,3,4\}}$	2.0	5.0	0.0	2.0
$\rho^1_{j\in\{2,3,4,5\}}$	5.0	0.0	2.0	5.0
$\rho^2_{j\in\{1,2,3,4\}}$	0.0	0.0	0.0	1.0

the associated $\mathcal{R}$ may contain ρ from different persons ($v \neq k$), or from the same person ($v = k$) but at times not in q ($j \notin i$).

For each ρ^v in $\mathcal{R}$, we calculate the distance between each $q_i \in \{q_{pre}, q_{post}\}$ and ρ_j, find the subset of $j \in M$ that minimizes the total sum of distances, and sum these distances to establish the overall dissimilarity, as shown in Equation 4.7:

$$d(q, \rho^v) = \sum_{i \in \{q_{pre}, q_{post}\}} \min_{j \in M_v} d(q_i, \rho_j) \tag{4.7}$$

From our earlier example query where $q = \{2.5, -, .001, .0014\}$, we may create $q_{pre} = \{2.5\}$ and $q_{post} = \{.001, .0014\}$. Table 4.2 provides the example reference series $\rho^1 = \{2, 5, 0, 2, 5\}$. We similarly split ρ^1 at each permissible spot that will allow for a comparison against both q_{pre} and q_{post}, separated by a gap of length $T = 1$, generating the following two comparison reference series with the gap from q occurring either at timepoint $j = 2$ or $j = 3$, producing respectively $\{2, (5), 0, 2\}$ and $\{5, (0), 2, 5\}$, in order to fit the pattern of missingness of q (later allowing the distances 5 and 0 respectively to be used to fill the gap). Table 4.3 shows the comparison of both possible alignment moments of ρ^1 with this q.

The example alignment from Table 4.3 of q to $\rho^1_{j\in\{1,2,3,4\}}$ would produce a total dissimilarity of $|2.5-2|+|.001-0|+|.0014-2| = 2.4996$. Meanwhile, the alignment of q to $\rho^1_{j\in\{2,3,4,5\}}$ would produce a total dissimilarity of $|2.5-5| + |.001-2| + |.0014-5| = 9.4976$, and to $\rho^2_{j\in\{1,2,3,4\}}$, a dissimilarity of $|2.5-0| + |.001-0| + |.0014-1| = 3.4996$. For $\mathcal{R}$, V total sets of indices and minimum dissimilarities are calculated, representing the best fit for each ρ^v. Because $\rho^1_{j\in\{1,2,3,4\}}$ is a better fit than $\rho^1_{j\in\{2,3,4,5\}}$, the alignment position and total dissimilarity of 2.4996 would be returned for ρ^1. Because only one possible alignment exists for ρ^2, alignment position $\{1, 2, 3, 4\}$ and dissimilarity 3.4996 will be returned for ρ^2.

We may constrain $\mathcal{R}$ by applying a set of restrictions, which we define as the function ϕ, as shown in Equation 4.8. This may limit persons (v) or time (j), or restrict the set of allowable minimization paths, as in Equation 4.7.

$$d(x, r^j) = \min(d_\phi(x, r^j)) \tag{4.8}$$

One logical limitation is implementing a time difference threshold between the query and reference sets. The maximum allowed difference in time between points in the query and points in the reference set we refer to as the window parameter. This limits the possible paths of alignment to be temporally similar so that morning travel behavior is imputed in the morning, midday travel behavior imputed for the midday, etc. Under this limitation, ϕ imposes the restriction $|t_i - t_j| \leq window$. This type of temporal windowing can also be applied to apply to days of the week or to allow for weekend/weekday restrictions. Reducing the number of computed similarities also has the side effect of decreasing the total computational time.

It is unclear which limitations perform best in this situation. In practice, one needs to make a choice for the parameters in ϕ, and it is not obvious how this should be done, nor have earlier studies provided much guidance. For this reason, we implemented a simulation study in the context of a week-long travel diary study in which GPS locations were measured over a period of a week to illustrate how different choices lead to different matching sets, and affect the statistics of interest. Appendix C.1 describes this simulation study.

4.2.4. Multiple Imputation

It is quite likely that the DTW procedure as explained in the previous section yields several possible matches. This is desirable, as it allows for the selection of a number of different matches instead of a single "best-fitting" example. We select m matches to create m different data sets against which we will calculate final statistics in a process called Multiple Imputation (Rubin, 2004). The basic premise behind multiple imputation is that the creation of multiple complete datasets from a single incomplete dataset allows us to obtain standard errors that are appropriately large (van Buuren, 2018). This is valuable because it permits calculation of uncertainty.

Multiple imputation can be used to describe a range of methods that can be used to fill in missingness. A subset of these methods, among which are Hot Deck imputation and Predictive Mean Matching, describe techniques that impute missing values with observed values (Little, 1988). This provides more plausible values as they are drawn from true occurrences, with the added advantage of requiring little in the way of model specification. Implementation of this method requires generation of a particular number of candidates on some basis of similarity. This similarity between the query and the reference sets that are possible donors is used to generate a probability of selection for each reference set (Siddique & Belin, 2008). This can be tuned by adjusting a parameter in the calculation to increase or decrease the probability of selection for similar donors. The multiple imputation procedure can be chained such that data that are incomplete at the start of the imputation procedure may themselves be used to impute subsequent data (van Buuren & Oudshoorn, 2000). This improves the quality of the imputations and in the case of missing time series data, allows for all available data to be used when determining the set of closest matches.

4.2.5. Dynamic Time Warping-Based Imputation

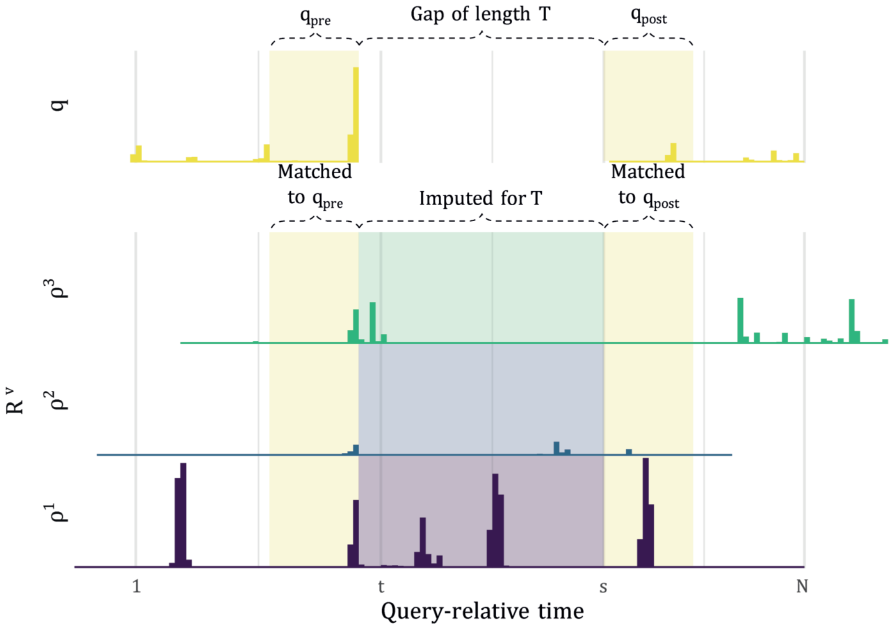


Figure 4.5 *An example of Dynamic Time Warping-Based Imputation. The match buffer for the query q is selected surrounding the gap, creating q_{pre} and q_{post}. Artificial gaps of length T are created in each reference series ρ, and buffers equal in length to q_{pre} and q_{post} (highlighted in yellow) are selected for matching against the query. Selected candidates donate the data between the buffer elements (highlighted in green, blue and purple respectively). The differing lengths of the series illustrate DTW in practice and its capacity to start at different points.*

The implementation of DTW as a selection mechanism for imputation has been previously described (Phan, Poisson Caillault & Bigand, 2020; Phan, Poisson Caillault, Lefebvre & Bigand, 2020; Phan et al., 2018). DTW may be used to find the best-fitting subsequence within candidate series based on matches against the observed data immediately preceding and following a gap in a target series in DTWBI. Variations on this concept have been implemented as a selection mechanism for gap-filling in various disciplines (Hung et al., 2022; Kostadinova et al., 2012; Zhang & Thorburn, 2021). This paper differs somewhat to previous implementations of DTWBI and extends the single imputation to multiple imputation.

First, the restrictions on the start and stop of query alignment are relaxed as in Open Begin/End DTW. Second, a strict one-to-one concordance is imposed between query and reference time series, which enforces the identifiability of the matching sequence and makes more intuitive sense with discretized time series. We also introduce Time Window, which is similar to the Sakoe-Chiba window, but restricts the query alignment to reference time points occurring within a specified time difference. This is done in order to align travel behaviors only with other travel behaviors occurring at the same time of day (although the days may differ). Lastly, we replace the gap in the target time series with a single time series element representing the interpolated measure, then excise and interpolate along each potential candidate at each potential gap moment, allowing us to make full use of all data. Figure 4.5 provides an example this procedure.

DTWBI has many parameters that describe its implementation. The efficacy of DTWBI is expected to depend on the selection of these parameters. In these analyses, we vary four parameters: two impacting the mechanisms of the dynamic time warping alignment, and two impacting imputation procedure. The parameters match buffer and time window reflect integral considerations for DTW alignment, while the parameters candidate specificity and number of imputations impact the subsequent imputation procedure.

The match buffer parameter describes the number of elements considered before and after the gap. A buffer of zero elements is equivalent to selecting candidates independent of DTW similarity. A larger buffer prefers longer patterns such as a person's work commute while a shorter buffer prefers shorter patterns, such as a person's moving preceding a gap but not afterwards.

The time window parameter describes the maximum time difference allowable between matchable elements in the target and query time series. For example, a window of two hours would allow a query element of 8AM to match against target elements occurring between 6AM and 10AM. The tighter the time window, the more candidate matches are restricted to conform to time-based routines. The wider the time window, the more candidate matches are allowed to favor similar behavior without requiring the time element. A time window of 12 hours would allow for unrestricted pattern matching in the time series (12 hours before, and 12 hours afterward).

DTWBMI is the extension of DTWBI to the multiple imputation framework. Whereas in DTWBI, only the best fitting candidate is selected for imputation, with DTWBMI we select a number of different candidates. This adds new parameters that can impact the imputation – the number of imputations and the candidate specificity. A higher number of imputations means that we will average results out across a number of worse- and better-fitting candidates from whose data the imputed data will be drawn. This has the advantage of allowing us to quantify the uncertainty of each estimate. Candidate specificity is a parameter that describes the relationship between the selection probability of an imputation candidate and how close a match the donor is. A higher candidate specificity leads to a higher preference for similarity, versus a low candidate specificity which gives us a high tolerance for misspecification. If the candidate specificity is too high, we may select the same donor across all imputations, reducing to the DTWBI case of a single best donor. If the candidate specificity is too low, we may select donors without regard for similarity at all, reducing to the Time Window imputation case.

These four parameters produce a number of different combinations whose effectiveness is expected to depend not only on the interactions with the other parameters, but also on differences in the data. Of primary interest is selecting parameters sets that are appropriate to extent of the information available within the data. As the amount of data increases, either through increasing the total number of persons or the length of data collection, a longer match buffer will be beneficial in increasing the chance that identical routes will be identified during the alignment procedure. Similarly, a higher candidate specificity would be beneficial in a high-information setting, increasing the frequency of selection of very-similar imputation candidates. The expected impact of the parameters time window and number of imputations is less clear. A set of parameters more appropriate for a low-information data set in which there is little overlapping data either within or between individuals may reflect instantaneous travel patterns, such as travel immediately preceding or following a gap, or may have a shorter time window in order to benefit from the temporal aspects of travel behavior. A low-information parameter set may thus prefer a shorter time window, shorter match buffer, reduced candidate specificity, and an increased number of imputation streams.

We expect improved performance of the high-information variant relative to the low-information variant to be related to the type of travel behavior generated by a person, but restricted by the amount of available data for the person. Persons with very consistent travel patterns across days would be expected to benefit from a high-information variant as they will be more likely to match against their own near-identical travel behavior than against that of others. Persons with a varying travel pattern across days, but whose travel behavior is unlikely to deviate from typical travel patterns within the sample, are unlikely to benefit from the high-information variant until the study periods are long enough to encapsulate most of the deviation. A third type of person, for whom travel patterns vary over the days, but whose activities deviate from the travel patterns of others within the sample (e.g. a large number of stops/tracks, long travels), may find that the low-information variant performs acceptably in shorter gap, but poorly in longer gaps, and that a high-information variant would have poorer performance until a sufficient number of similar travel activities were observed for the same or other individuals.

We performed a simulation study to establish these two sets of parameters, leading to two methods called DTWBMI-HI and DTWBMI-LO, representing high- and low-information availability respectively. Full results are available in Appendix C.1.

4.3. Analysis strategy

This section describes the selection of a suitable data set for the simulation study, criteria for comparison, and the other methods against which this new method is to be compared.

4.3.1. Example data set

As a motivating example, we consider the data collected from a 2018 field test of the Statistics Netherlands travel app. This field test concerned 1902 sample persons aged 16 and older. ODiN is an online-only study of individual mobility in the Dutch population (Centraal Bureau voor de Statistiek (CBS) & (rws-Wvl), 2020).

Both groups of respondents were contacted via post with a request to download the application onto their personal mobile devices, register using the enclosed personal username and password, and record seven days of movement behavior. Full details on app methodology and data structure have been outlined in (McCool et al., 2021). The app captured a participant's location once per second while the person was determined to be in motion, and once per minute while the person was determined to be stationary. This determination was based upon an algorithm that assessed whether or not the displacement between recorded intervals exceeded thresholds indicating movement behavior. Collectively, a total of 2087 person days were recorded amongst 576 participants.

We describe the missing data in terms of covered time and with respect to the number and length of gaps. A natural gap occurs between each two successive recorded locations within a single person's trajectory. If this gap is very small, on the order of seconds, very little information on a person's continuous trajectory is lost. As the time elapsed between locations increases, the amount of potential information lost increases. Deciding on a maximum allowable gap length is somewhat arbitrary, but should depend on the smallest movement behavior of interest to the researcher. The missing data patterns in this data set were determined on the basis of a maximum gap time of six minutes. The contiguous length of time elapsing for a person without exceeding the maximum gap time between any two successive locations we refer to as the covered time. Covered time can be expressed as a percentage of a discrete length of time, such as an hour or a day, as a measure of data completeness. We can then speak of the hourly, daily coverage. In this data set, the average length of covered time in a day was 8.2 hours, corresponding to a mean coverage of 34.2%. The mean number of gaps per calendar day across all participants was 4.0, and the mean gap length was 2.23 hours in length.

Following from the initial missing data analyses, we identified two specific aspects of interest across which we expect the effectiveness of the gap-filling methods to differ: participation length, and temporal variation. First, participants recorded data on a variable number of days, ranging from one to 43, creating a natural variation in the total information available per individual. Figure 4.6 (a) shows the distribution of persons in the data set by the number of days on which they submitted at least some data, grouped by their average hourly coverage. Some respondents maintain high levels of daily coverage, represented by the brighter yellow color, over a long period of time. As the amount of data from a person's own travel behavior increases, the additional days become available to form reference sets for imputing potentially missing data. Secondly, McCool et al. (2022) noted previously that nighttime hours have an increased incidence of missingness, due in part to operating system behaviors that restrict processing when the device is not in use. Figure 4.6 (b) illustrates the average coverage across time and day of the week. Coverage is lowest during night time hours and highest during commuting times through the week. Patterns of coverage are more dispersed within the weekends. Because most travel behavior occurs during the daytime hours, and much of the missing data occurs at night, imputation of missing data occurring during the evening is unlikely to outperform simpler methods.

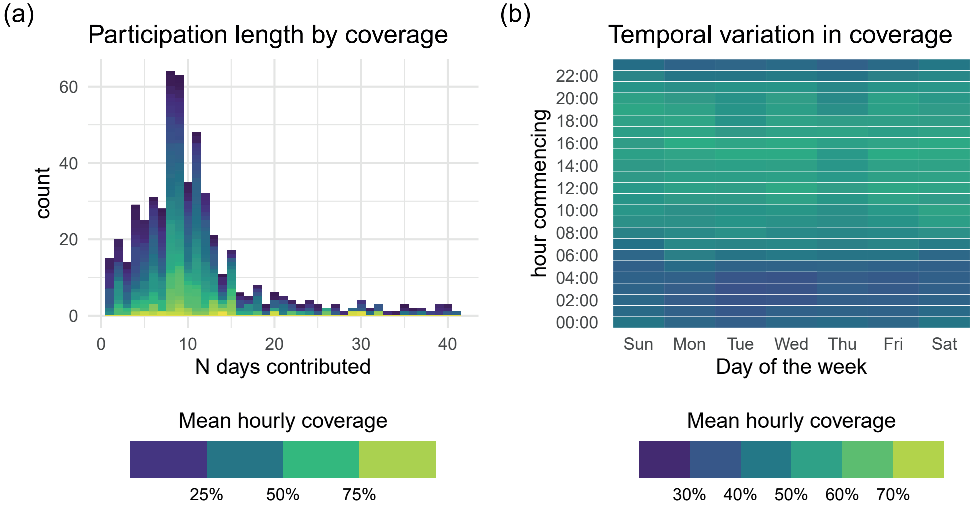


Figure 4.6 *Increased length of participation may increase the value of high-information gap-filling mechanisms by providing more reference sets of a person's own behavior. Missingness that occurs in the evening has a low likelihood of containing travel behavior, which may reduce the impact of high-information gap filling mechanisms.*

4.3.2. Comparison

From the data set described in Section 4.3.1, we selected all contiguous periods of at least 24 hours and with no gaps exceeding 6 minutes in duration. Individual trajectories with implausible speeds between successive locations were manually inspected to identify which locations were more likely to be incorrect. These were then flagged for removal, and the data were again filtered to remove any contiguous sets with gaps exceeding 6 minutes. Following this, 274 contiguous sets across 143 respondents remained. Figure 4.7 shows the spatial density of all recorded trajectories relative to the density of the set of trajectories covering at least 24 hours.

Because the length of the gap plays an integral role in the accuracy with which the data can be imputed, we created five different missingness scenarios, representing gap lengths of one hour, three hours, six hours, 10 hours and 12 hours. Within each scenario, 100 of the 274 sets were selected at random for introduction of missingness. The missingness was introduced into each of the selected sets as a single contiguous period. Each scenario randomized which sets contained missing data, as well as the start time of the missing data, but all methods compared within each scenario were applied to the same data.

In total, we compare six different methods on their capacity to address missing data induced at various levels. These methods are: 1) LI, 2) Mean Imputation (MI), 3) Time Window Imputation (TWI), 4) Dynamic Time Warping-Based [Single] Imputation (DTWBI), 5) Dynamic Time Warping-Based Multiple Imputation under High Information (DTWBMI-HI), and 6) Dynamic Time Warping-Based Multiple Im-

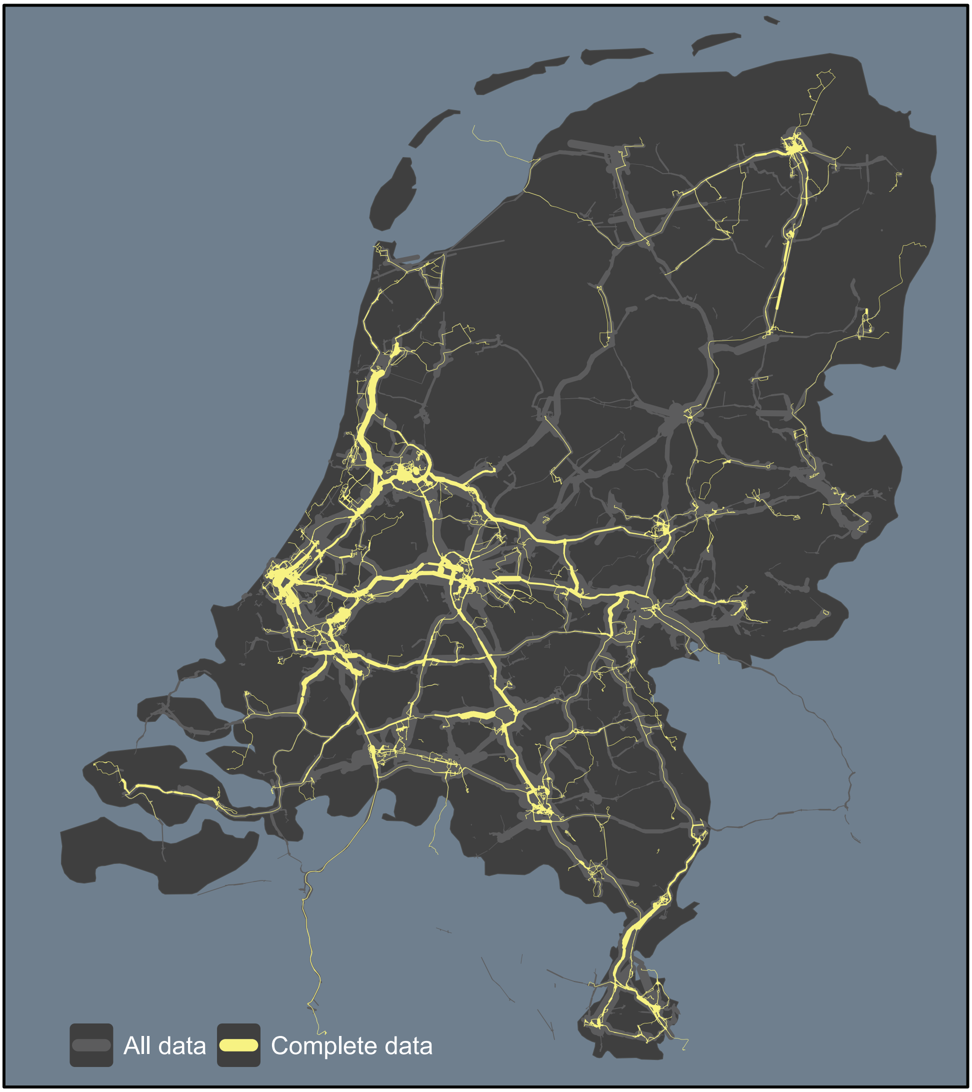


Figure 4.7 *Spatial density of trajectories recorded in the app by participants. Light grey areas show segments represented in the full set of data, which includes both data sets with gaps as well as data sets covering less than 24 hours in total. Yellow areas show the set of data used in the simulation study, which required at least 24 hours of contiguous data with no gaps exceeding 6 minutes.*

Note: Densities are calculated on the basis of the number of intersections with segments from other users, Non-intersecting segments were removed to protect privacy, but were used in the simulation study.

putation under Low Information (DTWBMI-LO). All methods were implemented in the R language and are available at https://anonymous.4open.science/r/DTWBI_Analyses-0178.

LI calculates the distance between the last point before the gap and the first point following the gap using the Haversine formula. This distance is divided equally across the N discrete missingness periods. The gap is then filled with a time series of identical travel behavior. This is done in order to preserve the time series format.

MI filled each gap with the mean value per hour of the applicable statistic, calculated on the basis of the non-missing portion. This personal mean is used to fill the gap, creating a time series of the travel behavior equal to the length of the gap.

TWI imputes missing data comparable time window selected from candidates with complete data. Candidates' travel behavior similarity is not otherwise assessed. Both the one-hour and three-hour values were considered for the time window parameter with similar performance. We compare one-hour TWI because of its slightly better aggregate performance. Ten imputations were performed because of the theoretical value of increasing sufficient variability.

DTWBI selects the best-fitting candidate on the basis of DTW with a match buffer of eight hours and a time window of one hour. This parameter set was selected on a theoretical basis for a desirable performance assuming one near-identical travel pattern was available as an imputation candidate.

DTWBMI-HI selects the three best-fitting candidates based on DTW alignment, utilizing an extended match buffer of eight hours, and a high candidate specificity to target finding a few very close matches – such as a person's own activity trajectory when available. The parameters for match buffer and candidate specificity were theoretically posited to enhance matching precision, selecting for trajectories that had close alignment over a longer interval. The simulation study described in Appendix C demonstrated the viability of the theoretical combination, and suggested a 12-hour time window and three imputation candidates as complimentary on the basis of providing the lowest absolute bias as well as an acceptable profile across other performance measures. Both of these align with theory, as we are likely to have fewer very close matches and will thus prefer fewer candidates, and the lack of time restriction may allow for matching morning commutes against evening commutes, or for otherwise identical trajectories that are made at different points in time.

DTWBMI-LO selects the 10 best fitting candidates on the basis of DTW alignment, using a match buffer of one hour, a medium candidate specificity, and a time window of three hours. In the case of DTWBMI-LO, the simulation study detailed in Appendix A was used as the primary basis for parameter selection. This combination provided the overall best profile across the selected measures of performance. Subsequent investigation of the matches produced by this method suggested a theoretical mechanism of action relying on a general relationship between travel behavior immediately preceding/following a gap, and the travel behavior contained within the gap.

Results are split according to gap length, and also with respect to the two characteristics identified in 4.3.1: number of available sets and night-only missingness. This may allow for fine-tuning method selection.

4.3.3. Performance criteria

We selected the following two key travel metrics for evaluation of performance: total distance and number of stops. We compare the parameters on the basis of Root Mean Square Error (RMSE) and mean absolute bias (Bias), as well as on a set of metrics developed to assess the accuracy and directional bias of the imputed travel distance and number of periods spent moving. RMSE and Bias both assess the accuracy of the underlying imputed metric in absolute terms. Because different parameter sets generated either a significant upward bias or downward bias on total distance, we compared under- and overestimation separately.

Travel Period overestimation (TP ↑) reflects the percentage of 15-minute travel periods imputed that did not exist in the true data set. Conversely, Travel Period underestimation (TP ↓) reflects the percentage of the true number of periods that were spent in movement that were not reflected in the imputation. Travel Period Accuracy (TP Acc.) reflects the percentage agreement with the total count of moving/stationary periods between the true data and the imputed data.

Distance overestimation (Dist ↑) reflects only the upward bias of the imputed values relative to the true distance. Distance underestimation (Dist ↓) similarly reflects only the downward bias of the imputed values relative to the true distance. Both are expressed in kilometers.

4.4. Results

4.4.1. Travel distance

Table 4.4 shows results for travel distance imputation following imputation with the six tested methods: LI, mean imputation, TWI, DTWBI, DTWBMI-HI, and DTWBMI-LO, averaged across all five missingness scenarios. The best average performance in terms of absolute bias (Abs Bias) is DTWBMI-LO, with a mean bias of 0.6 Km. The median bias (Med Bias) for all DTW methods is less than 100 meters, and 300 meters for the LI method. We can break out absolute bias into Dist ↑ or Dist ↓ which allows investigation of systematic directional biases in the gap-filling procedure. LI is incapable of overestimating the travel distance – instead all bias reflects an underestimation of travel distance of 5.8Km on average. All DTW-based methods systematically underestimate travel distance. DTWBMI-LO outperforms the other methods; despite a slight underestimation of travel distance, both overestimation and underestimation are small relative to other methods, and the difference between the average overestimation and underestimation is small.

The travel period metrics offer insight into the shape of the gap-filling method. DTWBI provides the highest accuracy with respect to reproducing the correct num-

Table 4.4 *Comparison of methods for imputing travel distance*

	Abs Bias[1]	Med Bias[1]	Dist ↑[1]	Dist ↓[1]	RMSE	TP Acc[2]	TP ↑[2]	TP ↓[2]
LI	5.9	−0.3	0.0	5.9	1.05	91.9	3.7	36.8
MI	1.9	5.7	7.8	5.8	1.38	93.7	0.0	42.4
TWI	1.1	2.0	8.4	7.3	1.97	91.8	23.5	23.3
DTWBI	1.8	0.0	2.7	4.5	1.40	95.5	7.4	20.8
DTWBMI-HI	0.7	0.0	4.8	5.5	1.49	94.2	13.4	21.3
DTWBMI-LO	0.6	0.0	3.2	3.8	1.44	95.3	12.2	16.1

Note. [1] Km, [2] %

ber of 15-minute periods in which a person was traveling. DTWBMI-HI and DTWBMI-LO also offer approximately 95% accuracy. TWI and LI are less accurate than the other methods. TP ↑ and TP ↓ allow for investigation of systematic biases in travel period estimation. DTWBMI-LO is the only method for which underestimation and overestimation percentages are similar and also low. Mean imputation has an overall TP ↑ of zero because the mean value imputed for the missing values does not exceed the movement threshold. Linear interpolation, although it is not capable of exceeding the total traveled distance, can assign a distance exceeding the movement threshold to a period which originally contained no movement. Here, DTWBI demonstrates a problem with systematic underestimation of travel periods.

Previous studies have demonstrated that the length of the gap is an important consideration in deciding on a method. Table 4.5 shows method performance metrics across the five simulations with differing gap lengths. With a gap one hour in length, DTWBI, DTWBMI-LO and mean imputation provided the least bias on average. Mean imputation offers quite a high median bias at 1.9 Km on average, with other methods all less than 100 meters. The over- and underestimation was relatively low across all methods due to the small likelihood of one hour intervals containing travel over long distances. The performance of DTWBI and DTWBMI-LO are very similar, performing best on these metrics, containing over- and underestimation in a similar ratio and to a small degree. All travel period metrics also favor DTWBMI-LO, although linear interpolation also offers acceptable performance.

As the gap length increases to 3 hours, the performance of DTWBI and DTWBMI-LO remain sufficient. Although both offer favorable profiles with low bias, there is little evidence for large systematic bias in distance estimation. LI and DTWBMI-HI offer acceptable performance with regard to overall bias metrics, but performs less well in the assessment of the imputed shape and accuracy of the travel periods. Methods TWI and mean imputation do not offer favorable profiles in the 3-hour gap length condition. In the 6-hour gap condition, DTWBMI-LO demonstrates the best performance. While TWI has the lowest absolute bias, it offers worse performance across the other metrics. DTWBMI-LO has a more favorable shape profile, with both 14% over and underestimation of the travel periods. The other methods offer relatively worse performance on average. Within the 10-hour gap simulation, DTWBMI-LO offers the best performance across all methods. The absolute bias is 300 meters on average, although the bias split by over- and underestimation is 6.8 Km and 6.5 Km respectively. Here, DTWBMI-HI also offers an acceptable performance with a small bias of 1.1 Km, and a favorable shape profile. In the 12-hour gap simulation, single imputation by DTWBI offers the best profile on average, with a bias of only 100 meters, although there is a meaningful bias towards travel period underestimation. DTWBMI-LO performs well here as well, with an absolute bias of only 200 meters, roughly equitable travel period over- and underestimation, but a median bias of 1.7 Km. Overall, increased gap length tends towards increased absolute bias across most methods, and an increase in both overestimation and underestimation of the total distance and the number of travel periods. The median bias remains low for all DTW-based methods, even as gap length increases to 12 hours. Both DTWBMI-LO and DTWBI are able to provide much better approximations of travel behavior than LI at gap lengths of 10 and 12 hours.

Table 4.6 shows performance of the imputation methods across time-based conditions in the 3-hour gap condition. Because people are less likely to travel during the night hours, different methodology or parameter sets may be preferable when imputing missing data occurring only at night. Records in which the missingness was induced only in the hours between 22:00 and 05:00 were identified and marked as "Night Only," to be compared to records also containing missingness occurring during the daytime hours. Missingness occurring only at night was imputed with less absolute bias in the DTWBMI-LO, DTWBI, and LI conditions. Systematic over- and underestimation was lower across all methods except for mean imputation when the missingness occurred at night. DTWBMI-LO and DTWBI offer the best daytime performance, with comparable absolute bias and median bias. DTWBMI-HI demonstrates less systematic bias towards underestimation of travel periods than DTWBI.

As the data contained multiple different sets of data for some respondents, it was possible to investigate the amount to which access to someone's own complete travel behavior from different time points could be beneficial in the imputation procedure. In Table 4.7, the methods are compared against the quantity of own data available as different data sets for serving as imputation candidates. For persons where only a single set was available and whom therefore could not serve as their own imputation candidate, DTWBMI-LO was the preferred method on the basis of all measures. When 2-3 sets were available, DTWBMI-HI and DTWBMI-LO provided similar performance with respect to the mean bias and over- and under-estimation. However, the travel period metrics demonstrate preference for the DTWBMI-LO method. In the 4+ sets condition, MI had the lowest absolute bias despite otherwise poor performance, while DTWBMI-LO had a preferable profile for reduction of systematic bias both for travel periods and distance. Notably, the performance of all DTW methods worsens as the number of sets increases, running counter to expectations.

To determine whether this result was due to inherent differences between groups with varying amounts of extra data, for example due to a relationship between

Table 4.5 *Comparison of methods for imputing travel distance by gap length*

		Abs Bias[1]	Med Bias[1]	Dist ↑[1]	Dist ↓[1]	RMSE	TP Acc[2]	TP ↑[2]	TP ↓[2]
1-hour gap	LI	0.8	0.0	0.0	0.8	0.79	93.0	5.5	5.0
	MI	0.9	1.9	1.5	2.5	1.37	93.0	0.0	16.0
	TWI	1.4	0.2	1.2	2.6	1.47	89.3	10.7	13.4
	DTWBI	0.5	0.0	0.3	0.9	0.96	95.0	3.2	8.8
	DTWBMI-HI	1.4	0.0	0.1	1.6	0.88	94.1	3.1	10.3
	DTWBMI-LO	0.7	0.0	0.2	0.9	0.75	95.7	3.2	5.5
3-hour gap	LI	2.4	−0.1	0.0	2.4	0.76	92.0	2.5	28.0
	MI	1.2	5.7	4.2	3.0	1.08	93.0	0.0	32.0
	TWI	1.9	1.6	5.4	3.5	1.69	90.4	18.7	20.8
	DTWBI	1.0	0.0	1.0	2.0	1.06	94.5	6.2	16.7
	DTWBMI-HI	1.9	0.0	1.1	3.0	1.05	93.4	8.2	19.7
	DTWBMI-LO	0.9	0.0	0.8	1.7	1.04	94.5	7.7	13.9
6-hour gap	LI	5.4	−0.2	0.0	5.4	0.99	92.9	3.3	37.0
	MI	1.4	11.5	7.9	6.4	1.31	94.5	0.0	42.0
	TWI	0.2	3.3	7.6	7.8	1.89	93.0	24.6	22.7
	DTWBI	3.4	0.0	0.9	4.3	1.37	96.5	7.4	17.6
	DTWBMI-HI	3.4	0.1	2.9	6.3	1.44	94.8	11.8	20.1
	DTWBMI-LO	1.9	0.1	2.3	4.1	1.42	95.6	13.1	14.8
10-hour gap	LI	11.7	−5.1	0.0	11.7	1.76	87.5	6.2	58.0
	MI	2.9	8.9	9.0	11.9	1.94	92.8	0.0	65.0
	TWI	4.0	2.7	11.0	15.1	2.67	92.8	25.4	34.2
	DTWBI	4.0	−0.2	4.5	8.6	2.21	95.5	10.8	25.7
	DTWBMI-HI	1.1	0.0	8.5	9.6	2.48	94.4	17.9	27.9
	DTWBMI-LO	0.3	0.4	6.8	6.5	2.54	94.7	18.2	22.5
12-hour gap	LI	9.4	−1.9	0.0	9.4	0.94	94.4	0.9	56.0
	MI	10.9	21.2	16.3	5.3	1.18	95.2	0.0	57.0
	TWI	9.3	13.0	16.9	7.6	2.13	93.8	37.9	25.3
	DTWBI	0.1	−0.4	7.0	6.9	1.41	95.9	9.4	35.2
	DTWBMI-HI	4.5	2.4	11.3	6.8	1.63	94.3	26.0	28.5
	DTWBMI-LO	0.2	1.7	6.1	5.9	1.44	96.0	18.9	23.9

Note. [1] Km, [2] %

Table 4.6 *Method comparison across night only vs day, 3-hour gap condition*

		Abs	Med	Dist ↑[1]	Dist ↓[1]	RMSE	TP	TP ↑[2]	TP ↓[2]
		Bias[1]	Bias[1]				Acc[2]		
Day missing	LI	3.2	−0.4	0.0	3.2	0.99	89.5	3.3	36.8
	MI	0.3	5.0	3.7	4.0	1.27	90.8	0.0	42.1
	TWI	2.3	3.4	6.9	4.6	2.17	87.5	23.1	27.4
	DTWBI	1.3	0.0	1.3	2.6	1.39	92.8	8.2	22.0
	DTWBMI-HI	2.7	0.0	1.3	4.0	1.34	91.5	9.5	25.9
	DTWBMI-LO	1.2	0.0	1.0	2.2	1.37	92.8	9.8	18.3
Night only	LI	0.0	0.0	0.0	0.0	0.01	100.0	0.0	0.0
	MI	5.8	5.8	5.8	0.0	0.48	100.0	0.0	0.0
	TWI	0.6	0.1	0.6	0.0	0.14	99.4	4.6	0.0
	DTWBI	0.0	0.0	0.0	0.0	0.01	100.0	0.0	0.0
	DTWBMI-HI	0.5	0.0	0.5	0.0	0.11	99.4	4.2	0.0
	DTWBMI-LO	0.0	0.0	0.1	0.0	0.02	99.8	0.8	0.0

Note. [1] Km, [2] %

Table 4.7 *Method comparison across number of reference sets, imputing travel distance, 3 hour gap condition*

		Abs	Med	Dist ↑[1]	Dist ↓[1]	RMSE	TP	TP ↑[2]	TP ↓[2]
		Bias[1]	Bias[1]				Acc[2]		
No extra data	LI	4.8	−0.6	0.0	4.8	1.01	91.9	4.3	38.8
	MI	3.7	5.6	7.7	4.0	1.32	94.0	0.0	45.3
	TWI	3.7	2.3	9.6	5.9	2.07	91.4	25.2	24.9
	DTWBI	0.5	−0.1	3.7	4.1	1.41	94.7	9.2	22.1
	DTWBMI-HI	0.7	0.0	5.0	4.3	1.50	94.1	17.1	22.7
	DTWBMI-LO	0.2	0.0	3.4	3.1	1.45	95.0	14.6	17.1
2-3 sets	LI	6.8	−0.4	0.0	6.8	1.09	92.0	3.0	40.6
	MI	1.6	5.7	7.6	6.0	1.42	93.0	0.0	45.5
	TWI	1.2	2.4	8.6	7.4	2.01	91.6	22.9	24.5
	DTWBI	1.4	0.0	3.2	4.6	1.45	95.5	7.5	21.9
	DTWBMI-HI	0.4	0.0	4.8	5.2	1.51	94.0	12.0	21.7
	DTWBMI-LO	0.4	0.0	3.6	4.0	1.44	95.3	11.9	16.1
4+ sets	LI	5.6	−0.1	0.0	5.6	1.02	91.9	4.2	28.5
	MI	0.8	4.0	8.1	7.3	1.36	94.5	0.0	34.3
	TWI	1.6	1.5	7.0	8.6	1.80	92.8	22.6	19.6
	DTWBI	3.7	0.0	1.0	4.7	1.32	96.2	5.5	17.6
	DTWBMI-HI	2.4	0.0	4.7	7.1	1.48	94.6	11.9	19.1
	DTWBMI-LO	1.7	0.0	2.5	4.2	1.43	95.6	10.3	15.2

Note. [1] Km, [2] %

Table 4.8 *Method comparison across all cases: Number of trips*

	Abs Bias	Med Bias	RMSE	TP ↑	TP ↓	TP Acc.
LI	0.02	0.00	0.30	23.8%	10.0%	74.9%
TWI	0.04	0.00	0.27	26.1%	20.0%	91.2%
DTWBI	0.03	0.00	0.21	14.0%	15.0%	94.4%
DTWBMI-HI	0.02	0.00	0.21	12.7%	19.4%	94.3%
DTWBMI-LO	0.02	0.00	0.21	14.4%	14.4%	94.7%

more travel behavior leading to longer tracking times, an additional simulation study was conducted using only those persons with at least four sets. The details are described in Appendix C.2. In this restricted simulation study, performance does indeed improve with additional own data sets, but the preference for the high-information DTWBMI-HI is not demonstrated.

4.4.2. Number of trips

DTWB(M)I methods are not restricted to the imputation of travel distance. All travel metrics that occur as a function of time may benefit from imputation as time series. Table 4.8 presents a comparison of the same methods applied to the number of trips occurring within a gap. While all methods present little bias, the methods DTWBMI-LO and DTWBI offer an improved TP Acc. and less systematic overestimation of the number of periods containing movement.

Table 4.9 shows the results of the simulations stratified by the length of missingness. While LI performs with more bias as the length of the missingness increases, the DTWBMI-LO and DTWBI methods deliver relatively consistent performance with regards to average absolute bias, and offer better relative performance in TP Acc., TP ↑, and TP ↓. At a gap length of 12 hours, all DTW-based methods maintain a TP Acc. of approximately 95%, while LI offers only a 63% TP Acc.

4.5. Conclusion

In this research, we introduced a unique approach to imputing travel behavior characteristics in human trajectory data, which we call Dynamic Time Warping Based Multiple Imputation (DTWBMI). We tested the performance of this methodology on a real-life dataset and conducted a simulation study to gauge the impact of model parameters. The outcomes convincingly indicated that the DTWBMI technique superseded other gap-filling strategies, such as mean imputation, Time Window Imputation, and linear interpolation, specifically for long gaps.

We demonstrate that different methods for gap-filling may provide better or worse results on the basis of the nature of the missingness, e.g. depending on gap length or time or day, or the nature of the data itself, e.g. depending on the number of own reference sets. Linear interpolation is an appropriate method for small gaps, and when missingness occurs at night, because the chance of travel is generally low in

Table 4.9 *Method comparison across varying gap lengths, imputing number of trips*

		Abs Bias	Trips	Med Bias	RMSE	TP ↑	TP ↓	TP Acc.
1-hour gap	LI	0.000	1.03	0.000	0.10	6.5%	1.0%	93.2%
	TWI	0.002	1.03	0.000	0.21	15.3%	11.1%	87.2%
	DTWBI	0.000	1.03	0.000	0.13	4.5%	9.8%	92.5%
	DTWBMI-HI	0.000	1.03	0.000	0.09	0.9%	8.4%	95.3%
	DTWBMI-LO	0.003	1.03	0.000	0.11	4.7%	4.5%	95.1%
3-hour gap	LI	0.000	0.98	0.000	0.17	14.2%	3.0%	85.2%
	TWI	0.025	0.98	0.000	0.21	23.5%	12.5%	91.5%
	DTWBI	0.000	0.98	0.000	0.14	9.2%	8.7%	95.7%
	DTWBMI-HI	0.007	0.98	0.000	0.13	6.7%	11.5%	95.5%
	DTWBMI-LO	0.012	0.98	0.000	0.13	7.7%	8.4%	96.2%
6-hour gap	LI	0.000	0.95	0.000	0.32	26.0%	7.0%	72.8%
	TWI	0.030	0.95	0.000	0.28	25.7%	21.9%	91.6%
	DTWBI	0.010	0.95	0.000	0.23	12.0%	15.8%	94.2%
	DTWBMI-HI	0.013	0.95	0.000	0.24	15.9%	20.0%	92.9%
	DTWBMI-LO	0.015	0.95	0.000	0.24	14.3%	14.6%	94.1%
10-hour gap	LI	0.040	0.98	0.000	0.46	37.2%	16.0%	60.5%
	TWI	0.029	0.98	0.000	0.32	33.4%	26.7%	92.5%
	DTWBI	0.030	0.98	0.000	0.29	25.4%	17.6%	94.4%
	DTWBMI-HI	0.003	0.98	0.000	0.28	20.2%	25.8%	93.5%
	DTWBMI-LO	0.007	0.98	0.000	0.29	24.6%	19.3%	93.7%
12-hour gap	LI	0.060	0.98	0.000	0.44	35.1%	23.0%	62.6%
	TWI	0.027	0.98	0.000	0.32	32.6%	27.6%	93.2%
	DTWBI	0.030	0.98	0.000	0.27	18.7%	23.0%	95.4%
	DTWBMI-HI	0.003	0.98	0.000	0.28	20.0%	31.5%	94.2%
	DTWBMI-LO	0.000	0.98	0.000	0.28	20.7%	25.2%	94.5%

these situations. On the other hand, in situations where it is necessary to fill a long gap, imputation methods will likely provide superior results because of their capacity to appropriately consider travel behavior variance. When a sufficient quantity of a person's own data is available to use as donor candidates, methods selecting fewer candidates with more closely aligned travel behavior should be preferred for their capacity to reduce the variance of the estimate.

DTWBMI-HI, with an 8-hour matching buffer and high candidate specificity was expected to be able to match on the basis of longer travel behavior patterns, such as commuting behavior, and was expected to have an advantage over DTWBMI-LO in scenarios in which travel behavior was more predictable, such as in the Daytime Missing or 4+ reference set conditions. In fact, DTWBMI-HI underperformed relative to DTWBMI-LO in almost all scenarios, including when imputation was restricted to persons with additional own data sets as in Appendix C.2. DTWBMI-LO is defined by a short 1-hour matching buffer, medium candidate specificity, and unrestricted time window, and thus matched imputation candidates on the travel behavior immediately preceding and following the gap. This proved to be a good fit across many scenarios, including as the length of the gap increased to 10 and even 12 hours.

Comparisons between the performance of DTWBMI-LO and DTWBI across scenarios contextualize the benefit of single versus multiple imputation. While DTWBI sometimes had the lowest absolute bias, the method tended toward a larger systematic underestimation. DTWBMI-LO performed more consistently across scenarios, despite its systematic overestimation of distance and travel periods. The source of this systematic bias is unclear, but may be due to differences in imputation candidates available under the two parameter sets. The longer match buffer and higher candidate specificity of DTWBMI-HI may more often be restricted to only the longer sets, where there is a slight negative correlation with distance traveled per period. DTWBMI-LO imputes on average an equal number of travel period overestimations and underestimations, indicating that the distance per period is too high.

We recognize certain constraints associated with the DTWBMI method, including the necessity for at least one complete dataset and the requirement for extended observation periods. Consequently, we recommend future research endeavors focus on investigating the limits of feasibility of the DTWBMI method for imputing travel behavior, and in broadening its potential applications.

In terms of future directions, we suggest examination of optimal parameters in datasets encompassing a larger population, prolonged participation periods, or data exhibiting Missing Not At Random characteristics. This will allow for establishing the generalizability and robustness of the DTWBMI method, and may additionally shed light on the appropriateness of varying parameter sets for differing datasets. Given that travel mode is a fundamental component of travel diary studies, it would also be worth exploring the possibility of imputing travel mode information concurrently with travel behavior characteristics via the DTWBMI method.

Notwithstanding these limitations, we posit that the DTWBMI method holds pragmatic implications for augmenting the precision of app-based travel diaries, and addressing its pervasive missing data problems, particularly when external data resources are limited. We urge researchers and practitioners to recognize this methodology as a potential solution for filling long gaps in human mobility data and to consider its capacity for expansion in various contexts.

5

Imputing missing data in app-based travel diaries

Danielle McCool, Barry Schouten, and Peter Lugtig

Abstract

Missing data in smartphone-based app studies of human mobility is the primary obstacle in their large-scale introduction as replacements for traditional diary-based surveys. The data are prone to periods of missingness that can range from minutes to days in length. Researchers seeking to answer questions on the basis of the collected data must account for this missingness, as it is likely to be informative, especially since data are more likely to be missing during periods of low travel. In this article, we investigate the impact of different methods for handling missing data arising from the 2018 Dutch Travel Survey in terms of the impact on aggregate travel behavior statistics. We propose and illustrate a method that recovers the missing travel behavior by imputing at the granular level using Dynamic Time Warping-based Multiple Imputation, followed by Multivariate Imputation by Chained Equations to address problems of representativity.

5.1. Introduction

Data on human mobility are integral to national transportation models guiding policy and practice (Axhausen, 2012; Haggar & Cooper, 2023; Harrison et al., 2020). Consequently, surveys on the topic are administered at a relatively high frequency, and with considerable care towards the accurate generation of the aggregate statistics (RVU Sverige, 2023), at high cost and effort (Morency et al., 2024). These data are consumed not only by the National Statistical Institutes (NSIs) or transportation agencies that generate them, but also by NGOs, research institutions, and companies. The consequences of inaccurate, biased, or imprecise estimates of mobility behavior can therefore be far-reaching.

Smart surveys have become increasingly important tools for information gathering when traditional methods encounter problems such as high respondent burden, questions for which respondents don't have answers, or topics that aren't well-suited for a straightforward question-answer approach (Schouten et al., 2025). Data on human mobility, traditionally gathered in the form of the Travel Diary Study seem to be a good fit for the smart survey methodology, and, threatened by increasing rates of non-response and declines in response quality (Bayart & Simas, 2024), many NSIs have initiated the process of developing smart surveys in the form of travel apps to replace or augment the TDS (Morency et al., 2024). Although the primary goal is generally easier, cheaper, and more accurate capture of the same data formats, the fact that the data are generally richer and cover more days has been noted as a nice added benefit (Allahviranloo & Recker, 2015; Harrison et al., 2020).

Considering the potential of passive data collection to combat the perennial limitations of the traditional TDS, such as underreporting (Stopher & Greaves, 2010; Witlox, 2007) and departure/arrival time rounding (Sfeir et al., 2024), coupled with the increasing prevalence of sensor-enabled smart devices, why do so many organizations that adopt these applications opt out again so quickly (Greaves et al., 2023; Harrison et al., 2020; Morency et al., 2024)? Greaves et al. (2023) suggest a combination of inertia, disinterest from participants, and technical challenges, Morency et al. (2024) identify difficulties with managing the data format differences, while Harrison et al. (2020) proposes the limited standards available for methodology and applications as a primary reason. We believe that the larger issue underscoring each of these difficulties is the massive impact of missing data arising from apps that passively collect travel behavior data.

Missingness, sparsity, or low observation frequency is a consistent problem with travel apps (Harding et al., 2021; Thomas et al., 2018).[1] Studies frequently demonstrate missing data ratios of 50% (Halabi et al., 2024). One week-long study reported that only 2 persons out of the 821 who downloaded the app provided 7 days of data, despite the relatively liberal initial requirement of at least one data point per hour for 16 hours (Lan & Helbich, 2023). Complicating matters further, this

[1] This problem exists in non-app-based Travel Diary Studies as well, where it is less identifiable (Sammer et al., 2018).

missingness is not equally distributed: temporal spikes, user characteristics, device aspects, geographical considerations and even travel behaviors themselves are correlated with the probability of providing a location. For example, users traveling by car may be less likely to have missing data during their journeys, and users traveling by underground train more likely. This missingness can result in inaccurate features (Currey & Torous, 2023) and biases in inferred mobility patterns (Uğurel et al., 2024).

A recent review on best practices for GPS data usage found that a majority of researchers making use of GPS data in their protocols do not report on missingness at all (Pearson et al., 2024). It is tempting to mistake the absence of data for evidence of absence and assume that no data implies no movement. This is, in fact, sometimes the case in accordance with device behavior (Meseck et al., 2016), but both user feedback and studies using parallel devices (Greaves et al., 2023; Montini et al., 2015; Thierry et al., 2024) that devices frequently fail to record travel behavior.

Current methods for addressing the missingness fall broadly under two categories: trajectory reconstruction, and behavior imputation. With trajectory reconstruction, a gap in a respondent's trajectory is replaced with plausible spatiotemporal elements to impute at the same level of granularity as the data collection itself. This includes both simple techniques, such as linear interpolation or exponential smoothing (Huo et al., 2010), and more sophisticated techniques such as map matching (Knapen et al., 2018) or the Gaussian Process Multi-Task reconstruction method proposed by Uğurel et al. (2024). Simple methods are generally acceptable for short gap reconstruction (Huo et al., 2010), but perform less well as the gap length increases (McCool et al., 2022) as they cannot account for the complexity of human travel behavior. Map matching can be used to extend the acceptable length of a short gap, providing more realistic alternatives than simple interpolation, but can't be used to impute whole trips that are missing. More complex methods can impute both trajectories and whole trips, solving the problem of long gaps. This can be done on the basis of similarities in and/or between users' trajectories (Ren et al., 2021; Thierry et al., 2024; Uğurel et al., 2024). Critically, these methods require longitudinal data spanning months or years, making them unsuitable for travel apps that typically collect data over days to weeks.

This leaves travel behavior imputation as the only remaining option for addressing long gaps. These methods also vary in complexity, though they share the key element of aggregation, which is the process by which the granular recorded trajectory becomes the travel statistic. In its simplest form, researchers may aggregate to day level statistics by summing across the collected data without respect to missingness. This is the same as imputing no travel behavior across any gaps. To avoid the downward biasing of travel statistics, aggregation must occur only across complete data, but obtaining this from most participants is unrealistic, and restricting analysis to participants with complete data may induce selectivity (Currey & Torous, 2023; Keusch, Bähr et al., 2022). In this paper, we present an improvement over this process by restricting aggregation to intervals that can be judged to be complete, after which travel behavior is multiply imputed.

This chapter presents a comprehensive methodological framework that addresses missing data in app-based travel diaries through a three-step sequential strategy. Rather than applying a single imputation approach across all types of missingness, we develop targeted methods for distinct gap categories: (1) interpolation for short gaps under 30 minutes, where spatial proximity permits trajectory reconstruction; (2) Dynamic Time Warping-Based Multiple Imputation (DTWBMI) for medium gaps of 30 minutes to 12 hours, where temporal patterns provide information about likely travel behaviors; and (3) Multiple Imputation by Chained Equations (MICE) for day-level missingness, where demographic and contextual variables must compensate for absent trajectory data. This sequential approach recognizes that different missing data mechanisms require fundamentally different inferential strategies, with each step designed to maximize information preservation while progressively addressing sources of bias.

The framework is demonstrated using data from the 2018 Statistics Netherlands Travel App field test, in which a probability-based sample of 1,902 individuals aged 16 and older were invited to record a week of mobility patterns. Among the 674 participants who provided data, only 5 recorded seven complete days without gaps, despite accumulating 2,087 person-days of observation – a stark illustration of the missing data challenge. By systematically addressing different types of missingness, our approach transforms these fragmented observations into a complete dataset suitable for population-level inference.

This study contributes to the emerging literature on smart survey methodology in three key dimensions. First, it provides empirical evidence of how different missing data treatments affect substantive travel behavior estimates, demonstrating biases of up to 25% in distance traveled when missingness is ignored. Second, it develops a generalizable framework that can be adapted across different data collection contexts, as the sequential strategy depends only on timestamped coordinates rather than platform-specific features. Third, it bridges the methodological gap between sophisticated trajectory reconstruction techniques designed for long-term tracking and practical requirements of short-duration travel surveys. The remainder of this chapter presents the theoretical foundations, implementation details, and empirical results of this integrated approach to missing data in app-based mobility measurement.

5.2. Types and causes of missing data in travel apps

The methodological framework depends on addressing missingness in a progressive set of steps. We first establish a taxonomy of patterns of missingness in app-based travel data. This classification scheme, grounded in empirical observation where possible, provides the analytical framework for the sequential imputation stragey.

5.2.1. Operational definitions and threshold selection

The establishment of operational thresholds for gap classification represents a critical methodological decision point that influences subsequent analytical procedures. In choosing the thresholds, we seek to balance theoretical considerations of human mobility patterns with the practical constraints of data availability.

Gap detection threshold (5 minutes) A challenge inherent in establishing boundaries for missingness is establishing a threshold at which the interval between measurements – natural artifacts of discrete sampling of a continuous time process – are large enough to constitute missing data. Although location provisioning services nominally operate at requested intervals, practical variability in signal acquisition, delays in processing, and power management protocols introduce stochastic elements that complicate gap identification. The selection of a 5-minute threshold is based on both empirical and theoretical considerations:

- Behavioral completeness - five minutes represents the approximate duration required to cross a meaningful activity location (e.g. residential building, retail establishment)
- Technical stability - This interval exceeds many typical small delays in signal acquisition and processing
- Analytical considerations - The threshold is fine-grained enough to allow for trip detection

Short gap threshold (30 minutes Conceptually, short gaps can be distinguished from the gap detection threshold by permitting the possibility of missing travel behavior, and from long gaps by whether we can preserve spatial trajectory information through deterministic methods. The 30-minute threshold represents an empirically validated boundary considering:

- Distance and trips are preserved - Empirical analyses suggest <5% total distance loss when interpolating gaps under this threshold, and a low likelihood of complete trip concealment (McCool et al., 2022)
- Analytical considerations - Linear interpolation across short gaps provides a larger, more comprehensive, and more representative set of complete data for the subsequent steps.

Long gap threshold (12 hours The transition from short gaps to long gaps occurs where behavioral patterns rather than exact trajectories become the primary inferential resource, and the transition from long gaps to day-level missingness similarly represents a switch to the use of demographics as a proxy for these behaviors. The 12-hour threshold is set with consideration of the following aspects:

- Circadian alignment - The threshold captures the split between active and rest periods

- Sufficient measured behavior - Gaps exceeding 12 hours will provide a decreasing amount of contextual information
- Individual variance - Variations in daily behavior, including sleep and commute schedules, may still be inferrable when the gaps don't exceed 12 hours.

These operational definitions represent theoretically grounded boundaries that align missingness with appropriate methodological responses. The progressive increase in threshold duration comes alongside escalating complexity of inferential challenges.

5.2.2. Short gaps (<30 minutes)

Although short gaps are classified on the basis of duration rather than cause, they often share similar causes which result in a disruption in the signal. One common cause of short gaps is the ”cold start” phenomenon, shown in Figure 5.1 a), where there is a delay in geolocation recording as the device switches from WiFi-based triangulation to GNSS-based triangulation (McCool et al., 2022). This delay can obscure the beginning of a trajectory and affect the accuracy of departure times, and can be much longer in built up environments. Another frequent cause of short gaps is signal loss due to occlusion(Stopher & Greaves, 2010). Figure 5.1 b) illustrates a signal gap resulting from a short tunnel, but large buildings, trees, or even atmospheric conditions can also disrupt the signal.

Because the maximal threshold for identifying a short gap is based on their low probability of concealed trips, these gaps can be corrected by deterministic interpolation across the gap. This preserves the trajectory geometry while accepting minor distance underestimation.

5.2.3. Long gaps (30 minutes - 12 hours)

The causes of long gaps are more varied than short gaps. One of the most frequent cause of long gaps is inherent in the design of modern smartphone operating systems, which restrict an app's functionality during certain periods (Bähr et al., 2022). This often occurs during the nighttime hours, while the device is charging, or more generally while the device is not in motion. During these periods, there is a very low likelihood of travel behavior. If this were the only cause of missingness, imputing these gaps as stationary behavior would have very little impact on the subsequent estimation of travel statistics.

Frustratingly, these gaps are generally indistinguishable from causes that are due to a recording failure within the app, which have a very high likelihood of containing travel behavior. Diagnosing the reason for recording failure is difficult without respondent feedback, but may stem from any of the following:

1. The device kills the app or restricts its processing behavior for battery optimization
2. The respondent closes the app or turns off the device
3. An empty battery leading to the device shutting off

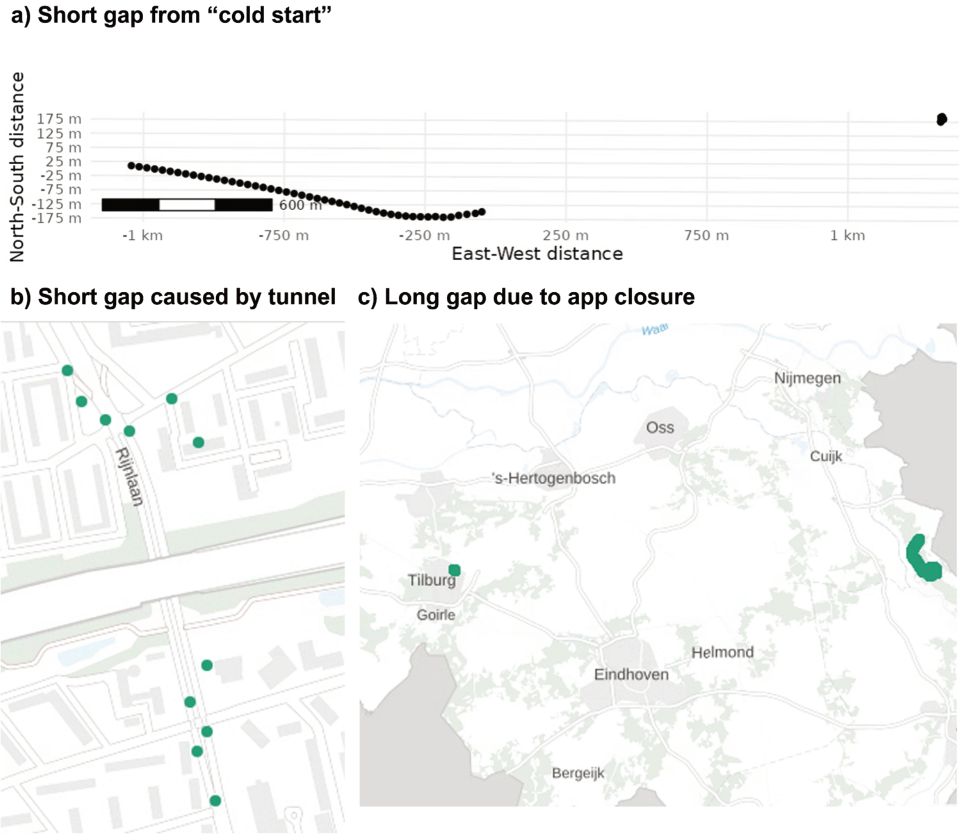


Figure 5.1 *Gaps in recorded geolocations can be due to many underlying reasons, many of which lead to identical outcomes. a) "Cold-start" gap resulting from a delayed location as a sufficient number of GNSS satellites are found for triangulation. This results in gaps that obscure the beginning of a trajectory. b) A very short gap caused by tunnel occlusion. Interpolating across these gaps loses little data as tunnels often take the shortest path. c) Long gap due to app closure. The large distance covered makes the gap evident, but a gap of equivalent duration without a spatial component may obscure whole trips.*

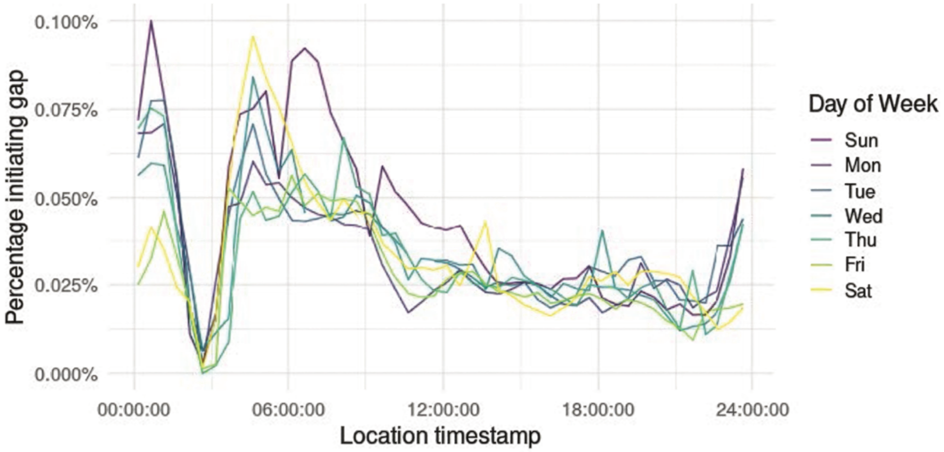


Figure 5.2 *Temporal proportional distribution of time invervals commencing periods of missingness. The distributions suggest both patterns that are consistent across days, showing increased missingness in evenings, as well as different across days, such as the larger and time-shifted spikes on Sundays.*

4. Disabled or restricted location permissions
5. Hardware- or location-based GNSS signal failure
6. Technical deficiencies in the travel app itself

We would like to distinguish between gaps of the first type, which imply a stationary device, and the second type, which do not. Incorrectly filling a type one gap with travel-containing behavior or a type two gap with stationary behavior would lead to incorrect estimates. If a temporal gap occurs with a spatial gap, as in 5.1 c), missing travel behavior is evident. However the reverse need not be true, as long temporal gaps with no spatial gap may conceal entire round trips.

Without additional data, we are limited to the relationships that can be found within timestamped geographic coordinates. The periodicity of human mobility provides one potential inroad. Most people (and thus also their devices) follow a 24-hour sleep-wake cycle which lends itself to a long period of being stationary abutted on either end with periods of travel – usually commuting behavior for work days. Using the temporal patterns in the data to inform the model is therefore crucial for addressing the gap with the Dynamic Time Warping approach described in Section 5.4.2.

Figure 5.2 shows the temporal proportional distribution of time intervals that start periods of missingness. Data are more likely to be missing in the evening, both as a consequence of user behavior (for example, some users report turning their phones off before going to sleep (Keusch, Wenz & Conrad, 2022)) and as a consequence of planned device down-time as described above. The distributions also differ per day, with Sunday showing both larger spikes as well as more pronounced temporal clustering. This temporal clustering aligns with documented patterns of reduced Sunday observation frequency, suggesting systematic behavioral variations in device interaction patterns (Uğurel et al., 2024).

The temporal patterns surrounding these gaps provide sufficient information for probabilistic reconstruction through Dynamic Time Warping, which leverages behavioral regularities without requiring extensive longitudinal data.

5.2.4. Day-level missingness (>12 hours)

Where long gaps end, day-level missingness begins, but this encompasses many different situations including non-response, break-off, device incompatibility, and app-stoppage. With day-level missingness, we lack sufficient data to permit gap-level imputation and must instead impute behaviors at a higher level of aggregation. In the case of non-response or break-off this is self-evident: there is no gap to impute. On the other hand, when does a long gap become day-level missingness?

Conceptually, we are interested at the cutoff point where the recorded location data will improve rather than bias our estimate of what occurred during the gap. The choice of 12 hours as an upper boundary is inspired by the 24-hour sleep-wake cycle and by earlier experiments that demonstrated acceptable properties with this cutoff (McCool et al., 2025). Although gaps over 12 hours may be estimated as long gaps, we expect to find that the benefit of estimating at the gap level decreases between 12 and 24 hours.

Imputing travel statistics on the basis of the recorded data would incorporate only the day of the week, ignoring any selectivity in the non-response. For example, retired persons may be both less likely to respond to an app-based survey (Stone et al., 2023) and may also travel less (Susilo et al., 2019). Imputing average travel behavior under selective non-response could lead to bias. We can improve our estimates by incorporating demographic variables that are related both to the response and the different travel behaviors, ensuring that the missing data in the analysis can be plausibly assumed to depend on observed data (Little et al., 2024; Rubin, 2004). The absence of trajectory information and associated shift to demographic and contextual covariates makes Multiple Imputation by Chained Equations (van Buuren, 2018) the appropriate methodological choice.

5.2.5. Relationship to methodological framework

This classification of missing data patterns directly guides our sequential imputation framework, with each type of data gap tackled using specific methodological approaches.

As gap duration increases, the inferential challenge transforms fundamentally. Short gaps require only geometric reconstruction; long gaps demand behavioral modeling; day-level gaps necessitate population-level inference. This progression motivates our sequential imputation strategy, where each method is optimized for its specific inferential context.

Table 5.1 *Gap classification and characteristics*

Gap Type	Primary Challenge	Methodological Solution	Key Assumption
Short	Spatial path	Interpolation	Proximate endpoints
Long	Behavioral	DTWBMI	Temporal regularity
Day-level	Selection bias	MICE	MAR given covariates

5.3. Methodological Framework

5.3.1. Data Source and Preparation

We use data gathered from the 2018 field test of the Statistics Netherlands travel application as a motivating example. The field test, based on methodology from an online TDS named ODiN, involved a sample of 1902 individuals selected at random from the Dutch population register, restricted to those aged 16 and over. Participants were invited via postal mail to download the application, register with a personal username and password, and document a week's worth of mobility patterns. The application requested the user's location every second during movement and every minute during stationary periods, although data were often received at a considerably lower frequency. Usable location data were gathered across 576 participants, amounting to a total of 2087 person-days. Full details on app methodology and data structure are available in McCool et al. (2021), and the open source app has been made available at https://gitlab.com/tabi/archive/tabi-app.

Raw location data underwent systematic quality control procedures. All entries with device-reported accuracy exceeding 200 meters were removed to maintain spatial precision. Each trajectory underwent visual assessment fo implausible transitions, defined as movements requiring speeds $\geq$ 200kph between contiguous locations. This manual inspection procedure identified approximately 20,000 locations for removal from 17.3 million total observations. Complete methodological details are available in Appendix C.

5.3.2. Conceptual framework for sequential imputation

The varied nature of the missingness patterns present in the data requires an analytical approach that benefits from differential treatment. Building on the gap classification proposed in Section 5.2, we develop a sequential imputation framework that aligns methodological strategies with the inferential requirements of each gap category. Currently, no generally accepted method exists for addressing missingness in geolocation data for human trajectories, with researcher decisions often constrained by methodological familiarity and data scale limitations.

The 3-step procedure seeks to address the tension between data granularity and inferential validity. Where data are present, we would like to avoid discarding available information by premature aggregation, but we also should be hesitant to overextend the capabilities of granularity-preserving methods. The framework operates on three core principles:

Principle 1: Hierarchical information preservation Each processing stage maximizes the retention of available information before proceeding to more abstract imputation methods. This hierarchy moves from reconstruction at the spatiotemporal level (interpolation) through behavioral modeling (DTWBMI), to demographic inference (MICE).

Principle 2: Progressive bias mitigation The sequential structure enables targeted correction of specific sources of bias at each stage. This prevents the accumulation of systematic errors that could result from a single-stage imputation.

Principle 3: Methodological alignment Each gap category requires fundamentall y different inferential approaches, necessitating method selection on the basis of available information density.

Currently, no generally accepted method exists for addressing missingness in geolocation data for human trajectories. The decisions made by individual researchers often depend on their familiarity with general methodology for treatment of missing data and their set of options is often limited by the scale of the captured data. The 3-Step processing architecture provides a framework for filling gaps sequentially from smallest to largest.

5.3.3. Implementation Overview

Figure 5.3 illustrates the operational framework that transforms fragmented mobility traces into complete datasets through systematic application of targeted imputation strategies.

Short gap resolution (<30 minutes) Short gaps, representing 75% of all gaps exceeding five minutes, arise primarily from brief signal disruptions. These are addressed deterministially through linear interpolation, which preserves the trajectory geometry. This step reduces the bias associated with the selectivity of restricting further analyses to fully complete data.

Long gap imputation (30 minutes-12 hours) Long gaps require probabilistic reconstruction based on temporal behavior patterns. The removal of spatial components necessitates aggregation to behavioral time series, enabling pattern matching through Dynamic Time Warping. Section 5.4 presents comparative analyses of alternative approaches, including restriction to complete weeks only or complete days only as methodological baselines. This step tries to resolve the bias arising from the relationship between the underlying travel behavior and the missingness itself.

Day-level reconstruction (>12 hours) Day-level missingness encompasses both unit non-response and extensive within-person gaps. The absence of trajectory information mandates reliance on demographic covariates from the Statistics

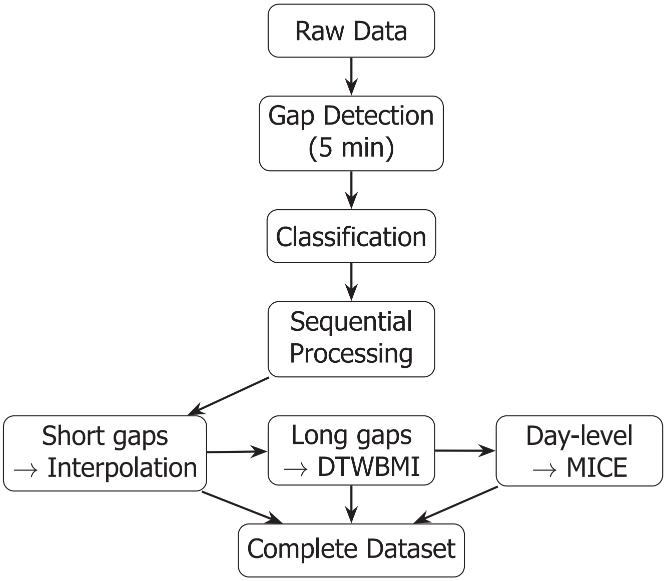


Figure 5.3 *Sequential imputation framework*

Netherlands population register. Variables used from the population register were selected based on their relationship with both travel behavior, as indicated in previous literature, as well as their relationship with the propensity to be missing. This step provides a correction for response propensity bias. The final layer results in five complete data sets suitable for analysis, after which the results can be combined using Rubin's rules (Rubin, 1976).

5.4. Implementation of Gap-Filling Methods

5.4.1. Step 1: Interpolation for short gaps

For each person, the set of locations was sorted based on time. Gaps between 5 minutes and 30 minutes where the distance was greater than 200 meters were replaced with a set of records at one-minute intervals. The difference in latitude and longitude between start and end position was calculated and incremented along the records.

This was a necessary step to facilitate the use of trajectories containing short gaps as donor trajectories during DTWBMI.

5.4.2. Step 2: Comparative approaches for long gaps

Each of the following proposed methods slots into the long gap filling layer and provides a data set that may be used for analysis in the final stage. If the missingness was uninformative, unrelated either to observed or unobserved characteristics, final estimates for the travel statistics would not be significantly different across models, and we would note only larger confidence intervals consequent to the loss of power. On the other hand, if the models produce large deviations, this would be indicative of informative missingness that must be addressed in order to reduce the bias of the missingness.

Using only complete weeks would discard data from all participants who did not register 7 days of complete data. This is likely to introduce bias into our data, as the largest factor in providing fully complete data is the device on which the data are collected. In restricting the analyses to include only those respondents who provide a full week of data, the selection is dependent on cell phone characteristics rather than the second largest contributor to missing data: travel behavior.

This has the benefit of removing the biasing impact of missingness related to travel behavior. The data may still have patterns of missingness associated with individual characteristics contributing to unit response propensities or particular cellphone ownership – aspects that may also be related to travel behavior.

Restriction to the set of complete weeks only is suspected to be relatively unbiased for the calculation of point estimates for travel behavior under particular assumptions, but in many studies, this restricts the sample to such a degree that calculation of population level statistics across even high-level registry characteristics such as age or population density, may be unstable or even impossible.

Using only complete days would increase the amount of data available for estimating travel behavior. With this method, we exclude participant days that are incomplete. This results in an unbalanced dataset and may be seen as a rough analogue to pairwise deletion. This notes two distinct advantages: an increase in the total *N*, and the simplification of the data curation process through the elimination of the selection of only a single day of the week for an individual [2]

It would also be possible to consider this method at a different level of aggregation, for example at the hour-level. As we are seeking to provide outcome statistics at the day-level, the day is a convenient unit of measurement for judging completeness.

The disadvantage to analysis of all complete days is in the potential to bias subsequent analyses. The data that are collected are likely to be missing in ways related to the travel behavior within this subset of participants. On a day-level basis, more travel behavior implies a higher likelihood of the behavior being recorded. This is evidenced by the fact that respondents who record only one day of travel tend to record a high distance traveled in comparison to people who record many days. In addition, this is in line with the expected behavior of device operating systems. This would violate the assumption that the missingness is unrelated to our statistics of interest and is expected to bias these estimates upwards.

[2] For example, a participant may contribute three complete Mondays to the estimation, but have no data for Thursday.

Dynamic Time Warping-Based Multiple Imputation as described by McCool et al. (2024) can be used to impute aggregate travel characteristics that preserves the maximum amount of data. DTWBMI looks at the pattern of movement behavior over time, aggregating it into time series with the goal of finding similar patterns.

DTW has been successfully used in other research to find similarities between full spatial trajectories (Thierry et al., 2024), and this capacity has been used by Zhang et al. (2024) as a basis for imputation not of the trajectories themselves, but of missing personal characteristics. While this study makes use of DTW in a different way in order to find donor candidates for aggregate series of travel behaviors, the same properties that allow the algorithm to perform well at subsequence matching in these other cases are expected to be a good fit for the imputation mechanism necessary here.

One particularly different element between this and the other long-gap filling methods in the literature is the stripping away of the spatial element of the trajectories, leaving behind only an aggregate travel behavior. While this is necessary due to the lack of sufficient overlap in the data, the properties are unclear. Chakrabarti et al. (2023) suggests using a very similar procedure for wearables[3] in which data surrounding the gap are aggregated to improve the imputation procedure. One previous study has shown this to be a successful procedure in the case of synthetically generated missingness in which the probability for being missing was completely random and unrelated to any data characteristic. Given that the true missingness in the data is certainly related to temporal and personal characteristics, or even directly to the travel behavior we're trying to measure, it may be difficult to obtain accurate estimates.

A discussion of several of the finer points required to implement of DTWBMI within this study can be found in Appendix D.

5.4.3. Step 3: MICE for day-level imputation

While DTWBMI allows for imputation of incomplete days with at least some travel information, we cannot use it to impute the data of non-responders. For this, it is necessary to use a more conventional method of handling missing data on the aggregate daily measures. MICE is an R package that allows fully conditional specification of multiple variables with missing data, and in which each variable may be defined by its own imputation model (van Buuren & Groothuis-Oudshoorn, 2011). Using a set of relevant variables from the Statistics Netherlands population register, relevant mobility statistics for a full week were imputed for every non-responder, and for days of the week with insufficient data to support DTWBMI.

Using MICE following DTWBMI allows us to correctly treat data that are differentially missing due to personal characteristics or measured variables. This is a necessary step, as previous research has demonstrated relationships with age, gender, length

[3] Such as accelerometers or heart rate monitors that are worn on the person, generating similar time series of sensor data

of participation in the study, and working status. The imputation model we select must consider the variables that demonstrate both a relationship with the propensity to be missing, as well as with the travel behaviors themselves. This sample, pulled from the Dutch Personal Records Database, has relatively complete records on variables known to be related to travel behavior, including income, urbanicity, and other demographics.

The model introduced in the MICE step as part of the final layer must also correct for two more critical elements: the day of missingness, and the unbalanced dataset. One method of correcting for the unbalanced datasets is by using only whatever complete weeks are available following the DTWBMI step. While this is possible if the number of sets is quite large, it is unlikely to be feasible in our situation, as the number is expected to be too low, especially relative to the full data set of 13,314 days across 1902 persons. Therefore, the model must correct for this in another way. Here, we have opted to implement a multilevel model specified for systematically missing predictors, from the mdmice package in R. Because many people, including the non-responders and those with insufficient data for imputing any days, will have systematically missing travel behaviors, it will be necessary to impute these by drawing from the posterior distribution on the basis of the persons where the data was either fully observed, or sporadically observed following the first chain.

5.5. Results

Table 5.2 presents the mean estimates and confidence intervals for three selected travel metrics: distance traveled, number of trips, and the number of short trips (trips $\leq$ 2 km), segmented by Weekdays, Saturday, and Sunday. The methods presented reflect progressively more sophisticated approaches that is able to account for more sources of error. The progression from linear interpolation deletion to multiple imputation on the aggregate data set allows more respondents to contribute complete weeks at each stage in the process, as the missing data handling metric fills progressively larger gaps. Comparing the results for complete days only, in which the estimates are not generated on the basis of complete weeks, demonstrates the necessity of weighting to reduce the impact of overrepresentation of person-days with more travel, which are more likely to be recorded in the app than person-days with less travel. The interpolation process primarily adds the set of person days that have had a significant amount of movement during the day that was recorded by their device, generally interspersed with stops during which no locations were recorded. This is borne out in the increase in both trips and short trips. That we again see an increase in travel behavior relative to the estimates calculated on the basis of the complete weeks only group is not surprising, as this step has biased the set of complete data towards incorporating highly-mobile persons. Despite this fact, this step is important, as this provides an otherwise missing pattern of stop-and-go travel behavior into the data.

Estimates made on the basis of respondents with a full week following the DTWBMI

Table 5.2 *Mean estimates and confidence intervals for travel metrics – including distance traveled, number of trips, and number of short trips – by day of the week (Monday – Friday, Saturday, and Sunday.) Metrics are computed on the basis of complete weeks through the three-step sequential imputation process, with comparison methods and reference dataset presented for context.*

		Sequential Imputation Process									Comparison Methods								
		Step 1: INT (N^1=13)			Step 2: DTWBMI (N^1=86)			Step 3: MICE (N^1=1896)			CW (N^1=5)			CD^2			ODiN (N^3=5897)		
Metric		*M*	95% CI		*M*	95% CI		*M*	95% CI		*M*	95% BCa CI		*M*	95% BCa CI		*M*	95% CI	
			LL	*UL*		*LL*	*UL*		*LL*	*UL*		*LL*	*UL*		*LL*	*UL*		*LL*	*UL*
Distance (km)	M-F	59	46	73	47	41	53	53	46	60	45	33	56	49	41	58	42	39	45
	Sat	49	28	71	55	40	71	55	38	71	50	17	92	50	36	64	38	33	42
	Sun	37	17	60	34	22	46	33	26	40	39	2	63	24	13	35	34	27	42
Trips	M-F	5.8	4.8	6.8	4.7	4.4	5.0	4.8	4.5	5.0	5.0	3.8	6.4	5.6	5.0	6.4	3.9	3.7	4.0
	Sat	3.9	2.5	5.6	4.5	3.8	5.1	4.7	4.3	5.1	4.4	2.5	6.4	6.3	5.0	8.1	3.2	3.0	3.5
	Sun	3.5	1.9	5.2	3.4	2.7	4.1	3.5	3.3	3.7	2.4	0.6	5.1	2.9	2.1	3.8	2.2	2.1	2.4
Trips $\leq$ 2km	M-F	2.9	2.3	3.4	1.8	1.6	2.0	2.5	1.7	3.4	2.1	1.5	2.7	2.5	2.2	2.8	1.6	1.5	1.7
	Sat	2.6	1.7	3.5	1.6	1.2	2.1	2.7	1.6	3.8	2.1	1.0	3.2	2.9	2.2	3.5	1.3	1.1	1.5
	Sun	2.2	1.1	3.5	1.4	1.0	1.8	1.8	0.9	2.7	1.7	0.3	4.0	1.4	0.9	2.0	0.8	0.7	0.9

Note. BCa CI = Bias-Corrected and Accelerated Confidence Interval, bootstrapped with 2000 iterations; LL = lower limit; UL = upper limit; INT = interpolation; DTWBMI = Dynamic Time Warping-based multiple imputation; MICE = Multiple Imputation through Chained Equations; CW = Complete Weeks only analysis restricted to weeks with 7 weekdays of $\geq$95% temporal coverage; CD = Complete Days only analysis restricted to days with $\geq$95% temporal coverage; ODiN = Underway in the Netherlands.

[1] Number of respondents with complete seven-day weeks

[2] Using all complete days results in varying n per day: $n_{M-F} = 224$, $n_{Sat} = 51$, $n_{Sun} = 43$

[3] ODiN is a one-day travel diary, this N reflects the total number of respondents during the selected period; $n_{M-F} = 4225$, $n_{Sat} = 806$, $n_{Sun} = 866$. Metrics are weighted according to the ODiN personal weighting factor.

procedure have integrated the temporal relationship with travel behavior. This provides a mechanism to partially alleviate the biasing impact of heavy travel days. When gaps occurring during the night hours are filled, they are filled with behavior that is similar in pattern before and after the gap, and that occurs within one hour of the missingness. This naturally leads to evening and night gaps being appropriately filled with stationary behavior if the preceding and following pattern suggested a stop, or movement behavior if the preceding or following pattern was suggestive of movement. The imputation model here is implicit; we do not need to define an explicit model for this temporal distribution, and because we select from multiple donors across multiple independent imputation chains, we also introduce an amount of variance that corresponds to the distribution of the underlying behavior within the full dataset.

The final column presents comparable results from a reference dataset ODiN (CBS-CvB, 2018). Weighted estimates are provided, using the total weights provided by SN to reflect the total population of the Netherlands. ODiN distinguishes categories of certain travel types, including travel outside the Netherlands, travel for work purposes with or without a freight truck, and travel to serial locations. Achieving the same categorization with the Statistics Netherlands app data was not possible, so ODiN data was aggregated across all conditions to achieve a data format as similar as possible to the the app-based data. Trip legs were selected as the comparable element to number of trips, and trip legs under 2 km were considered short trips.

Methodologies that use both the short and long gap filling layers produce results that are closest to what we would expect relative to the traditional TDS. We expected that our final estimates would indeed be higher than those of the traditional TDS, as this has been a common finding of many other studies. Considering that addressing the underreporting in TDSs is one of the proposed benefits of using apps to measure travel behavior, this is quite logical. However, the difference for Saturday is remarkable. It is additionally relatively persistent across the different missing data treatments. While the wide confidence intervals of the Complete Weeks, Complete Days and Interpolation methods encompass the ODiN estimate, neither DTWBMI nor MICE do.

Regarding short trips, we do see the increase in the total number of short trips as expected. In the MICE estimate, we estimate on average one more short trip for weekdays, but much less for weekends. All methods estimate the mean travel behavior on Sunday to be less than the mean travel behavior on other days of the week, consistent with other findings (Astroza et al., 2018; Dahmen et al., 2024).

Table 5.3 presents the unweighted mean estimates and confidence intervals for three selected travel metrics: distance traveled, number of trips, and the number of short trips (trips $\leq$ 2 km), segmented by Weekdays, Saturday, and Sunday. These results are computed on the newly added set of all person-days at each progressive stage of missing data handling, representing the incremental data added at each step. The data are presented this way in order to more clearly demonstrate the underlying changes in distributions between models.

Table 5.3 *Mean estimates and confidence intervals for travel metrics – including distance traveled, number of trips, and number of short trips – by day of the week (Monday-Friday, Saturday, and Sunday.) Metrics are calculated on the newly added set of all person-days at each progressive stage of missing data handling and represent the incremental data added at each step.*

		INT vs CD				DTWBMI vs INT				MICE vs DTWBMI			
Metric		*n*	*M*	95% BCa CI		*n*	*M*	95% CI		*n*	*M*	95% CI	
				LL	*UL*			*LL*	*UL*			*LL*	*UL*
Distance (km)	M-F	205	55	45	66	943	56	51	60	8392	53	44	61
	Sat	39	56	35	83	196	66	54	78	1668	54	35	72
	Sun	45	35	23	48	198	36	28	45	1675	33	25	41
Trips	M-F	205	5.8	5.2	6.4	943	5.0	4.8	5.2	8392	4.8	4.5	5.0
	Sat	39	6.1	4.8	7.6	196	5.0	4.5	5.6	1668	4.7	4.2	5.1
	Sun	45	4.1	3.2	5.0	198	3.7	3.3	4.1	1675	3.5	3.3	3.7
Trips $\leq$ 2 km	M-F	205	3.0	2.6	3.4	943	1.9	1.8	2.0	8392	2.5	1.5	3.5
	Sat	39	2.7	2.0	3.6	196	2.0	1.6	2.3	1668	2.6	1.3	3.9
	Sun	45	2.0	1.6	2.5	198	1.6	1.3	1.8	1675	1.8	0.7	2.9

Note. Results are computed across the total number of person-days added at each step, allowing respondents to contribute between 1 and 40 days. BCa CI = Bias-Corrected and Accelerated Confidence Interval, bootstrapped with 2000 iterations; LL = lower limit; UL = upper limit; CD = analysis restricted to days with $\geq$95% temporal coverage; INT = interpolation; DTWBMI = Dynamic Time Warping-based multiple imputation; MICE = Multiple Imputation through Chained Equations.

While Table 5.2 demonstrate that the sequential steps result in shifts in the estimation of the travel parameters, Table 5.3 focuses explicitly on the characteristics of all days that become available at each progressive step and therefore represent an unbalanced set per person. Interpolation and DTWBMI provides a set of higher average distances than are imputed at the day level in Step 3. Trip frequency demonstrates monotonic reduction (5.8 → 5.0 → 4.8 trips), while short trips display a non-monotonic pattern (3.0 → 1.9 → 2.5 trips $\leq$2km). These compositional differences provide insight into the distinct missing data mechanisms addressed by each methodological layer—interpolation preferentially captures high-mobility behaviors with intermittent gaps, DTWBMI reconstructs moderate-activity patterns through temporal regularities, and MICE leverages demographic covariates to impute low-mobility days and unit non-response. The observed gradients in travel metrics across imputation stages corroborate the hypothesis that passive data collection exhibits systematic biases favoring high-activity recording, which requires the sequential approach in order to mitigate the selectivity from device-level features and population-level representativeness. The widening confidence intervals in later stages appropriately reflect the propagation of uncertainty.

5.6. Discussion

The problem of missing data in smartphone-captured geolocation data has remained a central issue plaguing researchers over the past decade, but the number of studies introducing passively-acquired mobility data into their protocols has only increased.

Many studies ignore the missing data, or exclude all persons or days that do not meet certain thresholds without considering the potential bias introduced by these naive methods. We have developed a hierarchical imputation pipeline that fills gaps successively by maintaining the highest available granularity with respect to the gap characteristics. To do this, we have leveraged three methods: interpolation for short gaps, Dynamic Time Warping-Based Multiple Imputation (DTWBMI) for long gaps, and multiple imputation for unit-nonresponse. Each of these steps solves a particular need in reducing the overall bias: through interpolation, the addition of more complete days partially addresses the biasing impact of the combination of device type and travel behavior, through DTWBMI, we capitalize on temporal correlations and reduce bias from systematic patterns in missing data segments, and through multiple imputation, we account for uncertainty in the completely missing units. This study found that different treatments of missing data can lead to substantially different estimates of travel behavior. While this is expected, it reaffirms the notion that researchers who are working with passively-collected geolocation data must explicitly consider the problem of missingness rather than dismissing it as a consequence of the data collection mechanisms.

More specifically, this study demonstrates the biasing impact of a missingness generation mechanism that is correlated with the travel behavior being measured. Days containing more travel behavior, here best seen in the travel distance and number of tracks, are more likely to be recorded. Users who travel more frequently will, as a consequence, generate more days, resulting in an unbalanced dataset. If these days are allowed to contribute to the estimates without consideration to the nested structure, as in the complete days only example, the travel estimates made on the basis of these data will be biased upwards. Of course to do this without addressing the missing data will often result in a sample too small to be useful in addition to the potentially bias induced by selective personal characteristics. Handling the problem of gap-filling with three different steps to account for different gap lengths resulted in a final data set that was much closer to the the ODiN reference data set in terms of distance traveled, and where it deviated, it did so in expected ways. Much like other researchers, we found a slightly-increased total travel distance, and an increase in the number of short trips (Storesund Hesjevoll et al., 2021).

One primary goal of this study was was to demonstrate the significant impact of the chosen method of missing data handling on travel behavior-related outcomes. Lacking a ground truth makes it difficult to state with certainty that the imputation process has been able to address the bias completely, or that the resulting estimates are definitive. What we have shown is that naive strategies for handling data result in very imprecise estimates that are also quite volatile. The final estimates resulting from the MICE imputation model are aligned with estimates from other studies correcting for missingness in passively collected data, which we believe is an indication that the hierarchical imputation pipeline proposed here is at the very least more correct than the alternatives. In this paper, the spatial aspect of a person's travel behavior is lost at the DTWBMI stage. With sufficient data covering the same geographic area, either because the geographic area chosen for the study is very small, leaving many people to traverse the same areas, or because the study follows the same person for a long time, the DTWBMI step can impute actual trajectories with relatively good success (Thierry et al., 2024). We believe that this would improve the imputations at the granular step. However, progressing on to the MICE step to impute non-responders with no preexisting geographic information would still remove the spatial aspect. On the other hand, there are certainly benefits to using multiple imputation as the final step of the imputation pipeline. The resulting nicely complete dataset which can allow for many different types of model-based inference, such as agent based models (Li et al., 2024), or more traditional mode choice models (Huang & Levinson, 2015).

This method has some limitations that should be addressed. First and foremost is the fact that the long-gap and day-level imputation stages remove the data from their spatial context. This greatly limits the utility of the procedure if precise route reconstruction is required. Although DTWBMI is theoretically capable of imputing geolocations, this would require dense spatial overlap of trajectories to produce meaningful results. Another concern is that the method currently relies on temporal dependence to address some of the challenges of missingness related to the underlying travel behavior. This presumes that people tend to have similar temporal patterns on the whole that can be predicted by a short buffer before and after. If some individuals have highly irregular temporal patterns, this may not be sufficient, and the imputation model may not capture these differences.

Although this paper uses specific implementations at each stage, namely interpolation for reconstructing trajectories for small gaps, for establishing trajectory/behavior similarity for large gaps, and multiple imputation for larger gaps, these implementation details can be substituted with alternatives better suited for the data set at hand. For example, map matching would almost certainly offer an improvement on interpolation, subsequently opening the field for more sophisticated imputation or trajectory reconstruction options at the large gap stage. Extensibility of the DTWBMI methodology is also possible, as it supports a wide variety of travel metrics, including activity and travel model, if the model is built at trip leve. Dynamic Time Warping is also extensible to multivariate time series data, which would allow for simultaneous estimation of additional correlated metrics such as radius of gyration, or the introduction of person-level or day-level characteristics into the donor selection model (Lion & Shahar, 2021; Tormene et al., 2009). The implementation details are secondary to the idea behind them: data should optimally be solved for at the lowest level of aggregation feasible in order to reduce the various forms of bias.

Part III

Concluding Remarks

Plain Language Summary

Smartphones seem like the perfect tool for measuring how people travel. People carry them around with them at all times, they have GPS, and they don't require respondents to remember whether they biked to work or took the bus last Thursday. The problem is that the data they collect is often very spotty. Rather than day-length logs, we tend to have intermittent coverage for people, and this can vary a lot both over people and over the time of day.

In the field test that forms the backbone of this thesis, the app recorded a mean of just 8.2 hours of location data per person per day—roughly a third of the complete 24-hour day we were looking for. Phones lose satellite signal, batteries die mid-commute, and operating systems quietly kill apps to save power. The result is data that's more likely to be missing than present over the course of a week.

This thesis was crafted to be a formula for addressing this reality. It starts by explaining how and why we want to collect the data, then asks when we should start to worry about the gaps (short ones are not very problematic and can be ignored, but long ones are more of an issue), and ends by reconstructing what might have happened during the gaps by using a person's own travel patterns as a template. The method - Dynamic Time Warping-Based Multiple Imputation - sounds complicated because the details are, in fact, a bit complex. However, the core insight is simple: people are creatures of habit, and their past behavior can fill in for their missing present.

The final takeaway is that we shouldn't see missing data in smartphone-based travel surveys as a fatal flaw, because the missingness comes part and parcel with the many benefits. Rather, we should see it as a methodological problem, and address it with an appropriate methodological solution.

6

Conclusion

6.1. Reframing the missingness discussion

In the introduction to this thesis, I made a pretty bold claim: that the most significant barrier to smartphone-based travel surveys becoming viable replacements for traditional diary methods isn't respondent willingness or institutional investment – it's missing data. Our work is framed within the broader pattern of technological innovation alongside promising technologies like ARPANET and Next Generation Sequencing, where the initial failure to meet expectations ultimately transformed their respective fields through methodological refinement.

When NASA first designed the Mercury spacecraft, engineers envisioned a fully automated system where astronauts would merely be passengers. In fact, they didn't even include windows or manual controls. But reality intervened - astronauts needed situational awareness and the ability to make adjustments when automated systems faltered. The resulting hybrid approach, combining automation with human oversight, became the standard for space travel (Swenson et al., 1966). Similarly, smart surveys sit at this same intersection - trying to balance automated data collection with necessary human input for context and interpretation. Like NASA, our initial attempts tend to err on the side of assuming that the automation will fully address the issues we're facing. We would do better to take it in stride that smart surveys still require human intervention, whether that takes the form of input from the user during the survey, or post-hoc researcher-based methodological intervention.

The reality is that traditional survey methodology—which may appear straightforward from the outside—is actually a complex system built up over decades of refinement. For smart surveys, we're replicating these systems in a figurative[1] circuit board where each component serves a specific purpose in transforming raw data

[1] occasionally literal

into meaningful insights. In this thesis, I've focused on just one critical connection in that circuit: how to address missing data in smartphone-based travel surveys through Dynamic Time Warping-Based Imputation. By isolating this component, we can see its impacts not only within the circuit but also on the overall process. When we first asked the questions ”How can we identify when things are going wrong?”, ”Where should we look for the causes?”, and ”Against what should we be measuring?”, it was not yet clear that these would represent very fundamental questions for the field that would extend beyond our specific application, but they have.

We could certainly have chosen a different wire to follow — perhaps an app-based Household Budget Survey or a smart Time Use Survey. During the Smart Survey Implementation (SSI) project, I observed others encountering the same data missingness challenges I had faced years earlier. They asked the same questions I asked when I started[2], and which this thesis addresses. While the specific answers provided here, such as ”a gap of 20 minutes is probably fine if you're measuring travel behavior at the day-level,” may not interest smart survey researchers in other domains, the methodological framework for deriving such answers certainly does. An abridged version of this thesis would essentially be a formula for dissecting a smart survey to identify a problem and systematically develop a solution.

We could also easily have picked another switch in the circuit board that was not DTWBMI. The years between 2020 and 2023 were spent buried in half a dozen open tabs at any given moment, most of them suggesting shiny new ways of addressing missingness in data that were near-identical to the data from the 2018 Travel Survey App, but always slightly different in some meaningful way: too long, too car-centric, too infrequent, or paradoxically too complete[3]. Eventually, it became clear that any methodological choice would be imperfect, and the best approach was simply to start somewhere. We could have opted for Gaussian Processes of various flavors (Ahmed et al., 2022; Uğurel et al., 2024), or diffusion models (Yang et al., 2024). While our specific results might have differed, the process of applying these methods—and the broader insights about handling missing data in smart surveys—would have remained remarkably similar.

6.2. Key findings and their implications

6.2.1. Foundations for understanding the landscape

Chapter 2 laid the groundwork for everything that followed by introducing the Statistics Netherlands travel app and systematically documenting its first real-world implementation. While it functions as a technical description of the app, and a methodological description of the field test that is the backbone of the thesis, it was also an exploration of what happens when theory meets practice in an Smartphone-Based

[2] Not to mention the same questions that they asked back in 1983 at the Innovations in Travel Surveys conference Ampt et al., 1985

[3] Most of these datasets were presumably of the ”too-long” genus, but truncated with listwise or pairwise deletion.

Travel Surveys (SBTS).

One of the most sobering findings was the extent of the missing data problem. The app recorded a mean of only 8.2 hours of location data per participant per day – roughly one-third of the total time we'd hoped for. Far from being a minor technical issue, it proved to be a fundamental limitation that threatened the viability of the smartphone as a data collection apparatus. Even more concerning was the discovery that data completeness varied systematically across demographic groups and device characteristics, raising serious questions about representativeness.

The chapter also revealed insights about participation: incentives increased initial uptake, but had minimal impact on continued engagement. Device characteristics such as OS and manufacturer were demonstrated to be key aspects in data completeness. As these can't be standardized across participants, this represents a new layer of heterogeneity that must be accounted for.

Chapter 2 also established a clear picture of the relationship between actively reported and passively collected data. In comparing the data from the app with that from the traditional ODiN TDS, we could see both the potential and the limitations of the smartphone-based approach. The app captured more trips overall, particularly the expected short trips, but showed concerning and unlikely discrepancies in total travel distance.

By thoroughly documenting both the successes and failures of this initial implementation, Chapter 2 did more than identify the missing data problem – it characterized it so that it could be addressed. The documentation of the real-world experience opened up the path for the subsequent methodological innovations.

6.2.2. Establishing practical thresholds

Chapter 3 provided a framework for deciding on thresholds for when missing data becomes problematic enough to require intervention. The magnitude of the missingness doesn't necessarily tell the whole story. We're much more interested in whether or not the gaps meaningfully impact our ability to measure travel behavior. By introducing metrics like sparsity and establishing maximum interpolable gap lengths, we've provided practical tools for researchers to assess their own datasets.

The simulations presented in this chapter revealed something quite important: the relationship between missing data and bias in mobility metrics is neither linear nor universal. A 30-minute gap at midnight typically contains no travel behavior and can be safely interpolated, but the same gap during morning rush hour might obscure an entire commute. This temporal dependency creates both opportunities and challenges for researchers.

Our most actionable finding—that gaps under 10 minutes can generally be addressed through simple linear interpolation while maintaining bias below 5%—offers a straightforward guideline for handling short periods of missingness. This threshold remains reasonably stable across different travel metrics, including total distance traveled, radius of gyration, and number of moves. However, the simulation studies also showed that as gaps extend beyond this threshold, bias increases non-linearly, particularly for measurements of travel distance.

Perhaps most significantly, we demonstrated that the conventional approach of linear interpolation across longer gaps substantially underestimates travel behavior. In simulations with 12 total hours of missing data, linear interpolation underestimated travel distance by nearly 10 kilometers when the missingness was contiguous, versus only 3.6 kilometers when the missingness was dispersed throughout the day. This finding makes a compelling case against the naïve handling of missingness that remains prevalent in the field.

By establishing these empirical thresholds, Chapter 3 allows researchers to make informed decisions about when simple solutions are sufficient and when more sophisticated approaches become necessary. This foundation proved essential for the development of the DTWBMI methodology in subsequent chapters, as it clearly delineated the problem space where advanced imputation methods could provide their greatest value.

6.2.3. A novel methodology for long gaps

Chapter 4 represents the most substantial methodological contribution to this thesis. In it, we introduce Dynamic Time Warping-Based Multiple Imputation (DTWBMI) to address long gaps in human mobility trajectories. The seed of this idea arose from the concept that, when you record someone's movements for a longer period of time, they tend to go many of the same places and take the same routes. This is most evident when it comes to the work-home commute. When we would plot participants' behaviors on a map and arrange them by days, oftentimes you could overlay the complete days over the days with missingness and complete the picture. If you use the movement behavior leading up to the gap, and occurring after the gap, you can use this to pattern match against the complete sets, and use the ones with good fit to impute the travel behavior during the gap. This works, but very few people in our data set had enough data to make the direct imputation of spatiotemporal data a viable approach. The question then became: how could we make use of this method in a more general sense?

The most striking finding from this chapter was how effectively the low information variant DTWBMI-LO performed across the various simulation scenarios. With a mean absolute bias of .6 kilometers across all scenarios, it provided remarkably accurate imputations even as gap lengths increased to 10 to 12 hours. This is in stark contrast to linear interpolation, which underestimated travel distance by nearly 12 kilometers in 10-hour gaps.

What makes the approach here valuable is that it was achieved without requiring external data sources, map-matching, or extra sensors. It draws solely on the patterns in the most basic data available: timestamped geolocations. This has the benefit of making it broadly applicable across diverse mobility studies, regardless of geographical context or study length.

Interestingly, our expectations about which variant would perform best didn't hold up. We had anticipated that the high information variant DTWBMI-HI, with its extended matching buffer and high candidate specificity, would do quite well when fed with the sort of data that inspired the concept: people with predictable travel patterns with lots of own data. Actually, both methods performed worse for these candidates, but the DTWBMI-LO variant still outperformed DTWBMI-HI. This suggests that, while mobility patterns are generally predictive of travel behavior, what's most important is what's happening immediately before a gap occurs, which is a finding with broader implications.

The simulation studies also revealed nuanced performance differences based on contextual factors: nighttime gaps were universally easy to solve, for example. For daytime gaps, particularly those spanning typical commuting hours, the advantages of DTWBMI became much more pronounced, reinforcing the temporal dependency patterns identified in Chapter 3.

Finally, Chapter 4 demonstrated that the multiple imputation approach can effectively capture the uncertainty inherent in this type of missing data. In generating multiple plausible variations rather than a single ”best guess” option, DTWBMI produces more honest estimates of the variance of the travel metrics. This would be difficult to achieve without the use of an empirical method.

By systematically comparing these methods across varying gap lengths, temporal patterns, and data availability scenarios, we've provided researchers with clear guidance on when sophisticated imputation approaches become necessary and which variant might best suit their particular data characteristics. This transforms what was previously a subjective judgment call into an evidence-based methodological decision.

6.2.4. Real-world application and integration

Chapter 5 takes the theoretical and simulation-based work from earlier chapters and tests it in the messy reality of actual travel survey data. While simulations are invaluable for establishing methodological principles, they can't fully capture the complex patterns of missingness that emerge in practice. This chapter bridges that gap, demonstrating how our hierarchical approach to addressing missing data performs when confronted with the 2018 Statistics Netherlands travel app data.

The results revealed something quite interesting about the relationship between missing data handling methods and the resulting travel metrics. When we applied our full hierarchical imputation pipeline—addressing short gaps with interpolation, long gaps with DTWBMI, and day-length gaps with multiple imputation—we obtained estimates for travel behavior that aligned reasonably well with traditional travel diary studies, but with some key differences that reflect the known underreporting issues in those traditional methods.

Perhaps most significantly, our findings demonstrated that the choice of missing data handling methodology substantially impacts the resulting estimates. Naïve approaches like listwise deletion produced wildly varying estimates with huge confidence intervals, due to the fact that it reduced our sample to 5 total participants with 7 complete days. The hierarchical smart-, long-, day-level approach produced much more stable estimates with narrower confidence intervals, particularly when concerning weekday travel behaviors where travel patterns are more consistent.

Ultimately, this chapter demonstrated the existence of a practical pathway forward for researchers who find themselves with a dataset that was unexpectedly sparse. This is a situation that, as we've established throughout this thesis, is virtually inevitable. Rather than discarding the partially complete datasets, or trying to make inferences based on 5 participants, researchers now have evidence-based guidance for implementing a comprehensive missing data strategy.

Our comparison with ODiN suggests that this methodology is heading in the right direction. Although the absolute values differed somewhat, the overall patterns of travel behavior described in both studies aligned well, with similar variations across the days of the week, and trip frequencies that were more or less comparable. Hopeful to our cause was the fact that the differences that we did observe – particularly with respect to the higher trip counts in the imputed dataset – aligns with previous research suggesting trip underreporting remains an issue with self-report TDSs.

This final chapter pulls together the methodological threads from the entire thesis, demonstrating that what began as a targeted solution to a specific technical problem (missing data in smartphone-based travel surveys) has broader implications for how we approach sensor-based data collection more generally. By systematically addressing the missingness challenge at multiple levels, we've provided not just a specific solution but a framework for thinking about data quality in passive measurement systems.

Most importantly, Chapter 5 transforms what originally seemed like an insurmountable data quality issue into a manageable methodological challenge. By breaking down the problem into distinct categories of missingness and applying appropriate techniques to each, we've shown that smartphone-based travel surveys can indeed produce reliable mobility metrics despite their inherent limitations.

6.3. Practical applications for researchers and institutions

This thesis is not solely for methodologists looking for solutions for their missing data (although it is, of course, *also* for them), but instead something with concrete, immediate, and practical applications for a number of different stakeholders.

For App-Based Travel Survey Designers

- Work intimately with your app development team and collaborate directly on the implementation of the algorithms as they are implemented in the app
- Design with awareness of the systematic patterns of missingness

- Involve the user without being a burden in order to get complementary data that can help to identify, understand, and correct for gaps
- Build quality assessment into the data collection process to identify problematic patterns early

For Data Analysts Working with Existing Datasets

- Deal with different gaps in different ways: short gaps have many good options, long gaps currently very few
- Use DTWBMI in situations with high levels of missing data where the spatial component isn't important
- Use multiple imputation on the data aggregated to day level for gaps that can't be reasonably imputed
- Consider the temporal and behavioral patterns of missingness when conducting analyses

For National Statistical Institutes and Transportation Agencies

- Recognize that smartphone-based approaches, despite their limitations, offer valuable complementary data
- Plan ahead for the high level of missingness when deciding on sample size
- Leave time for the methodological development to occur - this reflects the actual goal of first attempts with technological innovation
- Don't go in with the expectation that smartphone-based approaches will replace traditional approaches by the second or third iteration if longitudinal stability is paramount
- Work towards a set of standards and best practices for handling missing data in passive mobility studies that consider the continuous nature of sampling

6.4. The road ahead: recommendations for future research

Right now, many researchers have found themselves in the challenging position of having already fielded an Smartphone-Based Travel Surveys (SBTS) only to find that they are now faced with datasets plagued by missing data. They may be in a position where they are now returning hat-in-hand to the stakeholder to bring news of an experiment that felt like a bit of a let-down. If this is you, then I ask you to consider the Hubble Space Telescope.

When the Hubble was first launched in 1990 – to be specific, *after* the Hubble was first launched – it was launched with its primary mirror incorrectly ground. As a result, all the images that it sent back were blurry and pretty disappointing. The Hubble was funded in the 1970s, took decades to engineer, and two months after launch, it failed to do the thing everyone had been anticipating for decades: deliver a crystal-clear view of outer space. The solution? Send up glasses[4]. Determining the correct specifications for the glasses required the researchers to develop digital image reconstruction algorithms for working backwards from the blurry data, which ended up being particularly useful in medical imaging(Luke et al., 2002).

Although to the best of my knowledge, no one is clamoring to abandon smart surveys altogether, I do think many are disappointed with their results. In a way, then, this thesis provides a set of methodological ”glasses” that correct a design flaw in smartphones that we seem to be stuck with. This is a good first step, but we need more. More concretely, we need adjustments that are precision-targeted at a specific solution. The difficulty is that that requires a large degree of collaboration between a number of different players. We need people with technical knowledge about the intricacies of software design: they are the only ones who can do anything meaningful to reduce the missingness. We need people who are knowledgeable about the sensors we want to use: they have decades of methodological work that can be extraordinarily useful to use, but often even identifying these works requires substantial domain knowledge, and the reading still more. We need machine learning methods, because oftentimes the data don't easily predispose themselves to more traditional techniques. More than anything, we need people who are willing to keep pushing, wanting more from Smart Surveys than traditional surveys have been able to provide.

This raises another point of contention, of course. The careful reader will have noticed that this thesis has skirted around the very real issue of how any of these methods can be validated. It feels disingenuous to suggest in the Introduction that the smart survey improves upon the measurement of the non-smart survey, only to turn around in the Results and use it as a benchmarking tool. Although I think no one would object to its usage to ensure that both modes produce estimates that are in the same order of magnitude, even the process of doing this is a nod to the current ”ground truth” status of the original survey. This is perhaps most evident at the moment where you must do something as simple as split an otherwise continuous dataset along discrete days.

The capacity of existing travel surveys to serve this function is limited, and limiting. When we choose estimates such as travel distance or number of trips, it's because these are the things we could have estimated before. You can't reasonably calculate route choice behavior from a traditional TDS, or use it to estimate the length of time that someone biking waits at a stoplight, but you *can* do these things with the SBTS. How suitable is the data, plagued by all its missingness, for this function? We don't really know, because this *is* the innovation that we are looking for. We can play fast and loose by making comparisons between the SN Travel App and ODiN only when we decide that we will remove the app from its context and meet the survey where it stands.

While I think these two future directions are the most critical, it's worth providing

[4] Corrective Optics Space Telescope Axial Replacement (COSTAR)

an overview of some concrete suggestions:

1. Multivariate DTWBMI applications: This thesis has demonstrated its success with univariate data, but the extension to multivariate data is trivial, and would open the field for incorporating other sensor data, personal characteristics, or simultaneous imputation of multiple travel characteristics.
2. Introducing other short-gap filling techniques: While we judged linear interpolation to be ”good enough” for our purposes, techniques such as map-matching or OSM-based routing are likely to offer an improvement, especially when considering distance as a measure,
3. Direct route imputation: DTWBMI is very well suited for matching trajectories against each other, which could be accomplished in a fully-passive dataset where users had many more days of complete and partial coverage.

6.5. Conclusion

The Smart Survey journey is following the pattern common in all technological evolution – the things that appear conceptually straightforward reveal layers of complexity in their implementation. This thesis has tackled one significant aspect of that complexity: the missing data problem in smartphone-based travel diaries. In developing a comprehensive understanding of the nature of the missingness, and establishing methodologies to address gaps of varying lengths, we've taken an important step towards making the SBTS a viable tool for mobility research.

Just as the Hubble Space Telescope's initial blurry images were corrected with COSTAR, and just as ARPANET's early transmission failures led to the robust error-checking protocols that power today's internet, the methods developed here transform what initially seemed like a fatal flaw into a manageable aspect of the research process. Our DTWBMI methodology serves as the ”corrective optics” for smartphone-based travel surveys, allowing them to fulfill their original promise despite inherent limitations.

The methods developed here don't eliminate the challenge of missing data, and it's unlikely that any ever will. What they do is transform it from a barrier to a manageable aspect of the research process. As these techniques mature and combine with other methodological advances, smartphone-based travel surveys will likely follow the same trajectory as other technologies that began with limitations but ultimately revolutionized their fields.

Part IV

Appendicies

References

Adler, T., Rimmer, L., & Carpenter, D. (2002). Use of internet-based household travel diary survey instrument [doi: 10.3141/1804-18]. *Transportation research record*, *1804*, 134–143. https://doi.org/10.3141/1804-18

Ahmed, H. M., Abdulrazak, B., Blanchet, F. G., Aloulou, H., & Mokhtari, M. (2022). Long gaps missing IoT sensors time series data imputation: A bayesian gaussian approach. *IEEE Access*, *10*, 116107–116119. https://doi.org/10.1109/ACCESS.2022.3218785

Allahviranloo, M., & Recker, W. (2015). Mining activity pattern trajectories and allocating activities in the network. *Transportation*, *42*, 561–579. https://doi.org/10.1007/s11116-015-9602-5

Allström, A., Kristoffersson, I., & Susilo, Y. (2017). Smartphone based travel diary collection: Experiences from a field trial in stockholm. *Transportation Research Procedia*, *26*, 32–38. https://doi.org/10.1016/j.trpro.2017.07.006

Ampt, E. S., Richardson, A. J., & Brög, W. (1985, December). *New survey methods in transport: Proceedings of 2nd international conference, hungerford hill, australia, 12-16 september 1983*. VSP. https://play.google.com/store/books/details?id=nvYSL9JPB00C

Arentze, T., Bos, I., Molin, E., & Timmermans, H. (2005). INTERNET-BASED TRAVEL SURVEYS: SELECTED EVIDENCE ON RESPONSE RATES, SAMPLING BIAS AND RELIABILITY [doi: 10.1080/18128600508685648]. *Transportmetrica*, *1*, 193–207. https://doi.org/10.1080/18128600508685648

Assemi, B., Jafarzadeh, H., Mesbah, M., & Hickman, M. (2018). Participants' perceptions of smartphone travel surveys. *Transportation research. Part F, Traffic psychology and behaviour*, *54*, 338–348. https://doi.org/10.1016/j.trf.2018.02.005

Astroza, S., Bhat, P. C., Bhat, C. R., Pendyala, R. M., & Garikapati, V. M. (2018). Understanding activity engagement across weekdays and weekend days: A multivariate multiple discrete-continuous modeling approach. *Journal of choice modelling*, *28*, 56–70. https://doi.org/10.1016/j.jocm.2018.05.004

Axhausen, K. W. (2012). Transport modelling. In *Computer modelling for sustainable urban design* (pp. 149–174). Routledge. https://doi.org/10.4324/9781849775403-8/transport-modelling-kay-axhausen

Axhausen, K. W. (1995). Travel diaries: An annotated catalog. (2nd ed). https://rosap.ntl.bts.gov/view/dot/13806

Axhausen, K. W., Molloy, J., & Tchervenkov, C. (2020). Has switzerland recovered? https://doi.org/10.3929/ETHZ-B-000417445

Bähr, S., Haas, G.-C., Keusch, F., Kreuter, F., & Trappmann, M. (2022). Missing data and other measurement quality issues in mobile geolocation sensor data.

Social science computer review, *40*, 212–235. https://doi.org/10.1177/0894439320944118

Baratchi, M., Meratnia, N., Havinga, P. J. M., Skidmore, A. K., & Toxopeus, B. A. K. G. (2014). A hierarchical hidden semi-markov model for modeling mobility data. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. https://doi.org/10.1145/2632048.2636068

Barnett, I., & Onnela, J.-P. (2020). Inferring mobility measures from GPS traces with missing data. *Biostatistics*, *21*, e98–e112. https://doi.org/10.1093/biostatistics/kxy059

Batool, T., Neven, A., Smeets, C. J. P., Scherrenberg, M., Dendale, P., Vanrompay, Y., Adnan, M., Ross, V., Brijs, K., Wets, G., & Janssens, D. (2022). A randomised controlled trial to enhance travel-related physical activity: A pilot study in patients with coronary heart disease. *Journal of transport & health*, *25*, 101344. https://doi.org/10.1016/j.jth.2022.101344

Bayart, C., & Simas, M. (2024). Workshop synthesis: Mixed modes and devices - integrating technology into traditional national travel surveys. *Transportation research procedia*, *76*, 657–664. https://doi.org/10.1016/j.trpro.2023.12.088

Berger, M., & Platzer, M. (2015). Field evaluation of the smartphone-based travel behaviour data collection app "smartmo". *Transportation Research Procedia*, *11*, 263–279. https://doi.org/10.1016/j.trpro.2015.12.023

Beukenhorst, A. L., Druce, K. L., & De Cock, D. (2022). Smartphones for musculoskeletal research - hype or hope? lessons from a decennium of mHealth studies. *BMC musculoskeletal disorders*, *23*, 487. https://doi.org/10.1186/s12891-022-05420-8

Beukenhorst, A. L., Sergeant, J. C., Schultz, D. M., McBeth, J., Yimer, B. B., & Dixon, W. G. (2021). Understanding the predictors of missing location data to inform smartphone study design: Observational study. *JMIR mHealth and uHealth*, *9*, e28857. https://doi.org/10.2196/28857

Bierlaire, M., Chen, J., & Newman, J. (2013). A probabilistic map matching method for smartphone GPS data. *Transportation research. Part C, Emerging technologies*, *26*, 78–98. https://doi.org/10.1016/j.trc.2012.08.001

Bihrmann, K., & Ersbøll, A. K. (2015). Estimating range of influence in case of missing spatial data: A simulation study on binary data. *International journal of health geographics*, *14*, 1. https://doi.org/10.1186/1476-072X-14-1

Boeschoten, L., Ausloos, J., Moeller, J., Araujo, T., & Oberski, D. L. (2020). Digital trace data collection through data donation. *arXiv [cs.CY]*. http://arxiv.org/abs/2011.09851

Bricka, S., Zmud, J., Wolf, J., & Freedman, J. (2009). Household travel surveys with GPS: An experiment [doi: 10.3141/2105-07]. *Transportation research record*, *2105*, 51–56. https://doi.org/10.3141/2105-07

Carrion, C., Pereira, F., Ball, R., Zhao, F., Kim, Y., Nawarathne, K., Zheng, N., Zegras, C., & Ben-Akiva, M. (2014). *Evaluating FMS: A preliminary comparison with a traditional travel survey* (research rep.). trid.trb.org. https://trid.trb.org/view/1289999

CBS-CvB. (2018, December). *Onderweg in nederland (ODiN) onderzoeksbeschrijving 2018* (research rep.). Centraal Bureau voor de Statistiek. Retrieved April 28, 2025, from https://www.cbs.nl/-/media/_pdf/2020/03/onderweg-in-nederland-odin-2018.pdf

Cellina, F., Bucher, D., Mangili, F., Veiga Simão, J., Rudel, R., & Raubal, M. (2019). A large scale, app-based behaviour change experiment persuading sustainable mobility patterns: Methods, results and lessons learnt. *Sustainability: Science Practice and Policy*, *11*, 2674. https://doi.org/10.3390/su11092674

Centraal Bureau voor de Statistiek. (2019). *ICT-gebruik van huishoudens en personen (ICT). [dataset]*. Retrieved June 8, 2019, from https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/ict-gebruik-van-huishoudens-en-personen--ict--

Centraal Bureau voor de Statistiek. (2022, February 10). *Onderweg in nederland (ODiN) 2018-2020*. Retrieved February 9, 2023, from https://www.cbs.nl/nl-nl/longread/rapportages/2022/onderweg-in-nederland--odin---2018-2020

Centraal Bureau voor de Statistiek (CBS) & (rws-Wvl), R. (2020, August 20). *Onderzoek onderweg in nederland - ODiN 2019*. DANS Data Station Social Sciences; Humanities. https://doi.org/10.17026/DANS-XPV-MWPG

Chakrabarti, S., Biswas, N., Karnani, K., Padul, V., Jones, L. D., Kesari, S., & Ashili, S. (2023). Binned data provide better imputation of missing time series data from wearables. *Sensors (Basel, Switzerland)*, *23*, 1454. https://doi.org/10.3390/s23031454

Chambers, T., Pearson, A. L., Kawachi, I., Rzotkiewicz, Z., Stanley, J., Smith, M., Barr, M., Ni Mhurchu, C., & Signal, L. (2017). Kids in space: Measuring children's residential neighborhoods and other destinations using activity space GPS and wearable camera data. *Social science & medicine (1982)*, *193*, 41–50. https://doi.org/10.1016/j.socscimed.2017.09.046

Chen, C., Gong, H., Lawson, C., & Bialostozky, E. (2010). Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the new york city case study. *Transportation research. Part A, Policy and practice*, *44*, 830–840. https://doi.org/10.1016/j.tra.2010.08.004

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation research. Part C, Emerging technologies*, *68*, 285–299. https://doi.org/10.1016/j.trc.2016.04.005

Chen, G., Viana, A. C., Fiore, M., & Sarraute, C. (2019). Complete trajectory reconstruction from sparse mobile phone data. *EPJ Data Science*, *8*, 30. https://doi.org/10.1140/epjds/s13688-019-0206-8

Cich, G., Knapen, L., Bellemans, T., Janssens, D., & Wets, G. (2015). TRIP/STOP detection in GPS traces to feed prompted recall survey. *Procedia computer science*, *52*, 262–269. https://doi.org/10.1016/j.procs.2015.05.074

Clarke, M., Dix, M., & Jones, P. (1981). Error and uncertainty in travel surveys. *Transportation*, *10*, 105–126. https://doi.org/10.1007/BF00165261

Cottrill, C. D., Pereira, F. C., Zhao, F., Dias, I. F., Lim, H. B., Ben-Akiva, M. E., & Zegras, P. C. (2013). Future mobility survey: Experience in developing a smartphone-based travel survey in singapore [doi: 10.3141/2354-07]. *Transportation research record*, *2354*, 59–67. https://doi.org/10.3141/2354-07

Currey, D., & Torous, J. (2023). Increasing the value of digital phenotyping through reducing missingness: A retrospective review and analysis of prior studies. *BMJ mental health*, *26*, e300718. https://doi.org/10.1136/bmjment-2023-300718

Dabove, P., & Di Pietra, V. (2019). Towards high accuracy GNSS real-time positioning with smartphones. *Advances in space research: the official journal of the Committee on Space Research*, *63*, 94–102. https://doi.org/10.1016/j.asr.2018.08.025

Dahmen, V., Martinez, S. Á.-O., Loder, A., & Bogenberger, K. (2024). Making large-scale semi-passive gps travel diaries valuable: A quality enhancement method. *Transportation Research Board Annual Meeting*. https://mediatum.ub.tum.de/1719353

Dekker, L., Đelić, K., van Dijk, M., Holtrop, S., Keuper, N., van der Laan, L., van Leeuwen, T., Meijs, C., Schroten, H., & van Wijk, L. (2022). Interpolating location data with brownian motion. *arXiv [stat.AP]*. http://arxiv.org/abs/2207.01618

Dhont, M., Tsiporkova, E., & González-Deleito, N. (2021). Deriving spatio-temporal trajectory fingerprints from mobility data using non-negative matrix factorisation. *2021 International Conference on Data Mining Workshops (ICDMW)*, 750–759. https://doi.org/10.1109/ICDMW53433.2021.00098

Douglas, D. H., & Peucker, T. K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica The International Journal for Geographic Information and Geovisualization*, *10*, 112–122. https://doi.org/10.3138/fm57-6770-u75u-7727

Forrest, T. L., & Pearson, D. F. (2005). Comparison of trip determination methods in household travel surveys enhanced by a global positioning system [doi: 10.1177/0361198105191700108]. *Transportation research record*, *1917*, 63–71. https://doi.org/10.1177/0361198105191700108

Fritz, M., Keusch, F., Volk, J., Häufglöckner, L., Blanke, K., De Vitiis, C., D'Amen, B., De Fausti, F., Inglese, F., Lorè, B. M., Pappagallo, A., Piccolo, F., Terribili, M., Perez, M., Van Tienoven, T. P., Lusyne, P., McCool, D., Lugtig, P., Struminskaya, B., … Holmøy, A. (2025, April 25). *Deliverable 2.3 smart advanced stage* (research rep.). ESSnet. Brussels, Belgium, European Commission.

Furletti, B., Cintia, P., Renso, C., & Spinsanti, L. (2013). Inferring human activities from GPS tracks. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, (Article 5), 1–8. https://doi.org/10.1145/2505821.2505830

Gadziński, J. (2018). Perspectives of the use of smartphones in travel behaviour studies: Findings from a literature review and a pilot study. *Transportation research. Part C, Emerging technologies*, *88*, 74–86. https://doi.org/10.1016/j.trc.2018.01.011

Geurs, K. T., Thomas, T., Bijlsma, M., & Douhou, S. (2015). Automatic trip and mode detection with move smarter: First results from the dutch mobile mobility panel. *Transportation Research Procedia*, *11*, 247–262. https://doi.org/10.1016/j.trpro.2015.12.022

Gomes, A., & Korf, B. (2018, January). Genetic testing techniques. In *Pediatric cancer genetics* (pp. 47–64). Elsevier. https://doi.org/10.1016/b978-0-323-48555-5.00005-3

Gong, L., Morikawa, T., Yamamoto, T., & Sato, H. (2014). Deriving personal trip data from GPS data: A literature review on the existing methodologies. *Procedia - Social and Behavioral Sciences*, *138*, 557–565. https://doi.org/10.1016/j.sbspro.2014.07.239

González-Pérez, A., Matey-Sanz, M., Granell, C., & Casteleyn, S. (2022). Using mobile devices as scientific measurement instruments: Reliable android task scheduling. *Pervasive and mobile computing*, *81*, 101550. https://doi.org/10.1016/j.pmcj.2022.101550

Gootzen, Y., Klingwort, J., & Schouten, B. (2025). Data quality aspects for location-tracking in smart travel and mobility surveys. https://www.cbs.nl/-/media/_pdf/2025/24/cbs_discussion_paper___ava_data_quality.pdf

Greaves, S., Ellison, A., Ellison, R., Rance, D., Standen, C., Rissel, C., & Crane, M. (2015). A web-based diary and companion smartphone app for travel/activity surveys. *Transportation Research Procedia*, *11*, 297–310. https://doi.org/10.1016/j.trpro.2015.12.026

Greaves, S. P., Cobbold, A., Stanesby, O., Sharman, M., Jose, K., Evans, J., & Cleland, V. (2023). Who stays and who plays? participant retention and smartphone app usage in a longitudinal travel survey. Retrieved October 28, 2024, from https://ses.library.usyd.edu.au/handle/2123/32005

Haggar, P., & Cooper, C. (2023). TrafRed model specification. https://users.cs.cf.ac.uk/CooperCH/TrafRed/methodology/TrafRed%20model%20specification%201.0.pdf

Halabi, R., Selvarajan, R., Lin, Z., Herd, C., Li, X., Kabrit, J., Tummalacherla, M., Chaibub Neto, E., & Pratap, A. (2024). Comparative assessment of multimodal sensor data quality collected using android and iOS smartphones in real-world settings. *Sensors (Basel, Switzerland)*, *24*. https://doi.org/10.3390/s24196246

Harding, C., Faghih Imani, A., Srikukenthiran, S., Miller, E. J., et al. (2021). Are we there yet? assessing smartphone apps as full-fledged tools for activity-travel surveys. *Transportation*. https://link.springer.com/article/10.1007/s11116-020-10135-7

Harding, C. (2019). From smartphone apps to in-person data collection: Modern and cost-effective multimodal travel data collection for evidence-based planning. https://utoronto.scholaris.ca/bitstreams/53b08745-9d40-49ed-990d-43acc2de4d49/download

Harrison, G., Grant-Muller, S. M., & Hodgson, F. C. (2020). New and emerging data forms in transportation planning and policy: Opportunities and challenges for "track and trace" data. *Transportation research. Part C, Emerging technologies*, *117*, 102672. https://doi.org/10.1016/j.trc.2020.102672

Hawthorne, G., & Elliott, P. (2005). Imputing cross-sectional missing data: Comparison of common techniques. *The Australian and New Zealand journal of psychiatry*, *39*, 583–590. https://doi.org/10.1080/j.1440-1614.2005.01630.x

Hecker, D., Stange, H., Korner, C., & May, M. (2010). Sample bias due to missing data in mobility surveys. *2010 IEEE International Conference on Data Mining Workshops*. https://doi.org/10.1109/icdmw.2010.162

Hollingshead, W., Quan-Haase, A., et al. (2021). Ethics and privacy in computational social science: A call for pedagogy. *of Computational Social …* https://doi.org/10.4324/9781003024583-13/ethics-privacy-computational-social-science-william-hollingshead-anabel-quan-haase-wenhong-chen

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American journal of political science*, *54*, 561–581. https://doi.org/10.1111/j.1540-5907.2010.00447.x

Hong, L., Zheng, Y., Yung, D., Shang, J., & Zou, L. (2015). Detecting urban black holes based on human mobility data. *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. https://doi.org/10.1145/2820783.2820811

Huang, A., & Levinson, D. (2015). Axis of travel: Modeling non-work destination choice with GPS data. *Transportation Research Part C: Emerging Technologies*, *58*, 208–223. https://doi.org/10.1016/j.trc.2015.03.022

Huang, J., Mao, B., Bai, Y., Zhang, T., & Miao, C. (2020). An integrated fuzzy C-means method for missing data imputation using taxi GPS data. *Sensors (Basel, Switzerland)*, *20*, 1992. https://doi.org/10.3390/s20071992

Humphreys, T. E. (2018). Centimeter positioning with a smartphone-quality GNSS antenna. https://doi.org/10.15781/T2HT2GV0R

Hung, L.-C., Hu, Y.-H., Tsai, C.-F., & Huang, M.-W. (2022). A dynamic time warping approach for handling class imbalanced medical datasets with missing values: A case study of protein localization site prediction. *Expert systems with applications*, *192*, 116437. https://doi.org/10.1016/j.eswa.2021.116437

Huo, J., Cox, C. D., Seaver, W. L., Robinson, R. B., & Jiang, Y. (2010). Application of two-directional time series models to replace missing data. *Journal of Environmental Engineering*, *136*, 435–443. https://doi.org/10.1061/(ASCE)EE.1943-7870.0000171

Hwang, S., VanDeMark, C., Dhatt, N., Yalla, S. V., & Crews, R. T. (2018). Segmenting human trajectory data by movement states while addressing signal loss and signal noise. *Geographical Information Systems*, *32*, 1391–1412. https://doi.org/10.1080/13658816.2018.1423685

Jagadeesh, G. R., & Srikanthan, T. (2017). Online map-matching of noisy and sparse location data with hidden markov and route choice models. *IEEE Transactions on Intelligent Transportation Systems*, *18*, 2423–2434. https://doi.org/10.1109/TITS.2017.2647967

Karaim, M., Elsheikh, M., Noureldin, A., & Rustamov, R. B. (2018). GNSS error sources. *Multifunctional Operation and Application of GPS*, 69–85. https://books.google.nl/books?hl=en&lr=&id=knqQDwAAQBAJ&oi=fnd&pg

=PA69&dq=sources+of+GNSS+error&ots=N5piLqtD0x&sig=SuB5V_eN9m3eD-TufTWylTPyl0k

Kelly, P., Krenn, P., Titze, S., Stopher, P., & Foster, C. (2013). Quantifying the difference between self-reported and global positioning systems-measured journey durations: A systematic review [doi: 10.1080/01441647.2013.815288]. *Transport Reviews*, *33*, 443–459. https://doi.org/10.1080/01441647.2013.815288

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., Trappmann, M., & Eckman, S. (2022). Non-participation in smartphone data collection using research apps. *Journal of the Royal Statistical Society. Series A*, *185*, S225–S245. https://doi.org/10.1111/rssa.12827

Keusch, F., & Conrad, F. G. (2022). Using smartphones to capture and combine self-reports and passively measured behavior in social research. *Journal of Survey Statistics and*. https://academic.oup.com/jssam/article-abstract/10/4/863/6375741

Keusch, F., Wenz, A., & Conrad, F. (2022). Do you have your smartphone with you? behavioral barriers for measuring everyday activities with smartphone sensors. *Computers in human behavior*, *127*. https://doi.org/10.1016/j.chb.2021.107054

Kiillaars, L., Schouten, J. G., & Mussman, O. (2019). Stop and go detection in GPS position data.

Knapen, L., Bellemans, T., Janssens, D., & Wets, G. (2018). Likelihood-based offline map matching of GPS recordings using global trace information. *Transportation Research Part C: Emerging Technologies*, *93*, 13–35. https://doi.org/10.1016/j.trc.2018.05.014

Körner, C. (2012). *Modeling visit potential of geographic locations based on mobility data*. Universitäts- und Landesbibliothek Bonn. https://hdl.handle.net/20.500.11811/5293

Kostadinova, E., Boeva, V., Boneva, L., & Tsiporkova, E. (2012). An integrative DTW-based imputation method for gene expression time series data. *2012 6th IEEE International Conference Intelligent Systems*, 258–263. https://doi.org/10.1109/IS.2012.6335145

Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S., & Trappmann, M. (2020). Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Social science computer review*, *38*, 533–549. https://doi.org/10.1177/0894439318816389

Lam, S. S.-S. (1974). *Packet switching in a multi-access broadcast channel with application to satellite communication in a computer network* [Doctoral dissertation, University of California, Los Angeles]. https://search.proquest.com/openview/96e299835561e709682f13e182b161f1/1?pq-origsite=gscholar&cbl=18750&diss=y

Lan, Y., & Helbich, M. (2023). Short-term exposure sequences and anxiety symptoms: A time series clustering of smartphone-based mobility trajectories. *International journal of health geographics*, *22*, 27. https://doi.org/10.1186/s12942-023-00348-1

Langley, R. B. (2015, October 8). *Innovation: Faster, higher, stronger*. Retrieved April 26, 2025, from https://www.gpsworld.com/innovation-faster-higher-stronger/

Lee, W.-C., & Krumm, J. (2011). Trajectory preprocessing. In *Computing with spatial trajectories* (pp. 3–33). Springer New York. https://doi.org/10.1007/978-1-4614-1629-6_1

Li, B., Cai, Z., Kang, M., Su, S., Zhang, S., Jiang, L., & Ge, Y. (2021). A trajectory restoration algorithm for low-sampling-rate floating car data and complex urban road networks. *Geographical Information Systems*, *35*, 717–740. https://doi.org/10.1080/13658816.2020.1825721

Li, J., Rombaut, E., & Vanhaverbeke, L. (2024). Agent-based digital traffic model generation for regions facing data scarcity using aggregated cellphone data: A case study for brussels. *International journal of digital earth*, *17*. https://doi.org/10.1080/17538947.2024.2407046

Lion, M., & Shahar, Y. (2021). Implementation and evaluation of a multivariate abstraction-based, interval-based dynamic time-warping method as a similarity measure for longitudinal medical records. *Journal of biomedical informatics*, *123*, 103919. https://doi.org/10.1016/j.jbi.2021.103919

Little, R. J., Carpenter, J. R., & Lee, K. J. (2024). A comparison of three popular methods for handling missing data: Complete-case analysis, inverse probability weighting, and multiple imputation. *Sociological methods & research*, *53*(3), 1105–1135. https://doi.org/10.1177/00491241221113873

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of business & economic statistics: a publication of the American Statistical Association*, *6*, 287–296. https://doi.org/10.1080/07350015.1988.10509663

Liu, G., & Onnela, J.-P. (2021). Bidirectional imputation of spatial GPS trajectories with missingness using sparse online gaussian process. *Journal of the American Medical Informatics Association: JAMIA*, *28*, 1777–1784. https://doi.org/10.1093/jamia/ocab069

Liu, J., Hu, Y., Zhang, D., & Liu, H. (2017). Performance assessment of GNSS measurements from android platform. *2017 6th International Conference on Computer Science and Network Technology (ICCSNT)*, 472–476. https://doi.org/10.1109/ICCSNT.2017.8343742

Luke, D. R., Burke, J. V., & Lyon, R. G. (2002). Optical wavefront reconstruction: Theory and numerical methods. *SIAM Review*, *44*, 169–224. https://doi.org/10.1137/S003614450139075

Lunardelli, I., van den Heuvel, J., Schouten, B., D'Amen, B., Loré, B., Nuccitella, A., Perez, M., & Zgonec, M. (2024). *How do respondents think about surveys with smart features?* (Tech. rep.). Statistics Netherlands.

Lynch, J., Dumont, J., Greene, E., & Ehrlich, J. (2019). Use of a smartphone GPS application for recurrent travel behavior data collection. *Transportation research record*, *2673*, 89–98. https://doi.org/10.1177/0361198119848708

Marra, A. D., Becker, H., Axhausen, K. W., & Corman, F. (2019). Developing a passive GPS tracking system to study long-term travel behavior. *Transportation Research Part C: Emerging Technologies*, *104*, 348–368. https://doi.org/10.1016/j.trc.2019.05.006

McCool, D., Lugtig, P., Mussmann, O., & Schouten, B. (2021). An app-assisted travel survey in official statistics: Possibilities and challenges. *Journal of official statistics*, *37*(1), 149–170. https://doi.org/10.2478/jos-2021-0007

McCool, D., Lugtig, P., & Schouten, B. (2022). Maximum interpolable gap length in missing smartphone-based GPS mobility data. *Transportation*. https://doi.org/10.1007/s11116-022-10328-2

McCool, D., Lugtig, P., & Schouten, B. (2024). Dynamic time warping-based imputation of long gaps in human mobility trajectories. *arXiv [stat.ME]*. https://doi.org/10.48550/arXiv.2410.16096

McCool, D., Struminskaya, B., & Lugtig, P. (2025). Integrating smart surveys with traditional methods: Challenges, opportunities, and methodological innovations. https://doi.org/10.48550/arXiv.2510.07521

Menard, T., Miller, J., Nowak, M., & Norris, D. (2011). Comparing the GPS capabilities of the samsung galaxy S, motorola droid X, and the apple iPhone for vehicle tracking using FreeSim_Mobile. *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 985–990. https://doi.org/10.1109/ITSC.2011.6083141

Mennis, J., Mason, M., Coffman, D. L., & Henry, K. (2018). Geographic imputation of missing activity space data from ecological momentary assessment (EMA) GPS positions. *International journal of environmental research and public health*, *15*. https://doi.org/10.3390/ijerph15122740

Meratnia, N., & By, R. A. D. (2003). A new perspective on trajectory compression techniques. https://research.utwente.nl/en/publications/a-new-perspective-on-trajectory-compression-techniques

Meratnia, N., & de By, R. A. (2004). Spatiotemporal compression techniques for moving point objects. *Advances in Database Technology - EDBT 2004*, 765–782. https://doi.org/10.1007/978-3-540-24741-8_44

Meseck, K., Jankowska, M. M., Schipperijn, J., Natarajan, L., Godbole, S., Carlson, J., Takemoto, M., Crist, K., & Kerr, J. (2016). Is missing geographic positioning system data in accelerometry studies a problem, and is imputation the solution? *Geospatial health*, *11*, 403. https://doi.org/10.4081/gh.2016.403

Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr, A. G., Beckstein, C., Braswell, B. H., Churkina, G., Desai, A. R., Falge, E., Gove, J. H., Heimann, M., Hui, D., Jarvis, A. J., Kattge, J., Noormets, A., & Stauch, V. J. (2007). Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agricultural and Forest Meteorology*, *147*, 209–232. https://doi.org/10.1016/j.agrformet.2007.08.011

Molloy, J., Castro Fernández, A., Götschi, T., Schoeman, B., Tchervenkov, C., Tomic, U., Hintermann, B., & Axhausen, K. W. (2020, August). A national-scale mobility pricing experiment using GPS tracking and online surveys in switzerland: Response rates and survey method results. https://doi.org/10.3929/ETHZ-B-000441958

Molnar, M. Z. (2012). *Error correction in next generation DNA sequencing data* [Doctoral dissertation, The University of Western Ontario]. The University of Western Ontario.

Montini, L., Prost, S., Schrammel, J., Rieser-Schüssler, N., & Axhausen, K. W. (2015). Comparison of travel diaries generated from smartphone data and dedicated GPS devices. *Transportation Research Procedia*, *11*, 227–241. https://doi.org/10.1016/j.trpro.2015.12.020

Montoliu, R., Blom, J., & Gatica-Perez, D. (2013). Discovering places of interest in everyday life from smartphone data. *Multimedia tools and applications*, *62*, 179–207. https://doi.org/10.1007/s11042-011-0982-z

Morency, C., Trépanier, M., Harding, C., Verreault, H., & Bourbonnais, P.-L. (2024). Scenarios for a national household travel survey in the province of quebec. *Transportation research procedia*, *76*, 1–12. https://doi.org/10.1016/j.trpro.2023.12.033

Nawaz, A., Huang, Z., Wang, S., Akbar, A., AlSalman, H., & Gumaei, A. (2020). GPS trajectory completion using end-to-end bidirectional convolutional recurrent encoder-decoder architecture with attention mechanism. *Sensors (Basel, Switzerland)*, *20*. https://doi.org/10.3390/s20185143

Ogle, J., Guensler, R., & Elango, V. (2005). Georgia's commute atlanta value pricing program: Recruitment methods and travel diary response rates. *Transportation research record*, *1931*, 28–37. https://doi.org/10.1177/0361198105193100104

Onnela, J.-P. (2021). Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology*, *46*, 45–54. https://doi.org/10.1038/s41386-020-0771-3

Palmer, M. C. (2008). Calculation of distance traveled by fishing vessels using GPS positional data: A theoretical evaluation of the sources of error. *Fisheries research*, *89*, 57–64. https://doi.org/10.1016/j.fishres.2007.09.001

Park, J., Muller, J., Arora, B., Faybishenko, B., Pastorello, G., Varadharajan, C., Sahu, R., & Agarwal, D. (2022). Long-term missing value imputation for time series data using deep neural networks. *arXiv [cs.LG]*. http://arxiv.org/abs/2202.12441

Park, M. H., Kim, H. C., Lee, S. J., & Bae, K. S. (2014). Performance evaluation of android location service at the urban canyon. *16th International Conference on Advanced Communication Technology*, 662–665. https://doi.org/10.1109/ICACT.2014.6779045

Parrella, M. L., Albano, G., Perna, C., & La Rocca, M. (2021). Bootstrap joint prediction regions for sequences of missing values in spatio-temporal datasets. *Computational statistics*, *36*, 2917–2938. https://doi.org/10.1007/s00180-021-01099-y

Patterson, Z., Fitzsimmons, K., Jackson, S., & Mukai, T. (2019). Itinerum: The open smartphone travel survey platform. *SoftwareX*, *10*, 100230. https://doi.org/10.1016/j.softx.2019.04.002

Pearson, A. L., Tribby, C., Brown, C. D., Yang, J.-A., Pfeiffer, K., & Jankowska, M. M. (2024). Systematic review of best practices for GPS data usage, processing, and linkage in health, exposure science and environmental context research. *BMJ open*, *14*, e077036. https://doi.org/10.1136/bmjopen-2023-077036

Pejović, V. (2025). Embracing shifting trends and reviving smartphone sensing. *IEEE pervasive computing*. https://doi.org/10.1109/mprv.2025.3542291

Petter, O., Hirsch, M., Mushtaq, E., Hevesi, P., & Lukowicz, P. (2019). Crowdsensing under recent mobile platform background service restrictions: A practical approach. *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 793–797. https://dl.acm.org/doi/abs/10.1145/3341162.3344867

Phan, T.-T.-H., Bigand, A., & Caillault, É. P. (2018). A new fuzzy logic-based similarity measure applied to large gap imputation for uncorrelated multivariate time series. *Applied computational intelligence and soft computing*, *2018*, 1–15. https://doi.org/10.1155/2018/9095683

Phan, T.-T.-H., Poisson Caillault, É., & Bigand, A. (2020). eDTWBI: Effective imputation method for univariate time series. *Advanced Computational Methods for Knowledge Engineering*, 121–132. https://doi.org/10.1007/978-3-030-38364-0_11

Phan, T.-T.-H., Poisson Caillault, É., Lefebvre, A., & Bigand, A. (2020). Dynamic time warping-based imputation for univariate time series data. *Pattern recognition letters*, *139*, 139–147. https://doi.org/10.1016/j.patrec.2017.08.019

Prelipcean, A. C., Gidofalvi, G., & Susilo, Y. O. (2015). Comparative framework for activity-travel diary collection systems. *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. https://doi.org/10.1109/mtits.2015.7223264

Prelipcean, A. C., Gidofalvi, G., & Susilo, Y. O. (2016). Measures of transport mode segmentation of trajectories. *International journal of geographical information science: IJGIS*, *30*, 1763–1784. https://doi.org/10.1080/13658816.2015.1137297

Prelipcean, A. C., Gidófalvi, G., & Susilo, Y. O. (2018). MEILI: A travel diary collection, annotation and automation system. *Computers, environment and urban systems*, *70*, 24–34. https://doi.org/10.1016/j.compenvurbsys.2018.01.011

Pronello, C., & Kumawat, P. (2021). Smartphone applications developed to collect mobility data: A review and SWOT analysis. In *Advances in intelligent systems and computing* (pp. 449–467). Springer International Publishing. https://doi.org/10.1007/978-3-030-55187-2_35

Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Prentice-Hall, Inc. https://dl.acm.org/doi/abs/10.5555/153687

Ramer, U. (1972). An iterative procedure for the polygonal approximation of plane curves. *Computer graphics and image processing*, *1*, 244–256. https://doi.org/10.1016/s0146-664x(72)80017-0

Ranacher, P., Brunauer, R., Trutschnig, W., Van der Spek, S., & Reich, S. (2016). Why GPS makes distances bigger than they are. *Geographical Information Systems*, *30*, 316–333. https://doi.org/10.1080/13658816.2015.1086924

Ranasinghe, C., & Kray, C. (2018). Location information quality: A review. *Sensors*, *18*. https://doi.org/10.3390/s18113999

Ren, C., Tang, L., Long, J., Kan, Z., & Yang, X. (2021). Modelling place visit probability sequences during trajectory data gaps based on movement history. *ISPRS International Journal of Geo-Information*, *10*, 456. https://doi.org/10.3390/ijgi10070456

Richardson, A. J., Ampt, E. S., & Meyburg, A. H. (1995). *Survey methods for transport planning*. Eucalyptus Press Melbourne. http://www.academia.edu/download/37634873/Survey_Methods_For_Transport_Planning.pdf

Robusto, C. C. (1957). The cosine-haversine formula. *The American mathematical monthly: the official journal of the Mathematical Association of America*, *64*, 38. https://doi.org/10.2307/2309088

Roddis, S., Winter, S., Zhao, F., & Kutadinata, R. (2019). Respondent preferences in travel survey design: An initial comparison of narrative, structured and technology-based travel survey instruments. *Travel Behaviour and Society*, *16*, 1–12. https://doi.org/10.1016/j.tbs.2019.03.003

Rout, A., Nitoslawski, S., Ladle, A., & Galpern, P. (2021). Using smartphone-GPS data to understand pedestrian-scale behavior in urban settings: A review of themes and approaches. *Computers, environment and urban systems*, *90*, 101705. https://doi.org/10.1016/j.compenvurbsys.2021.101705

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (2004, June 9). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons. https://play.google.com/store/books/details?id=bQBtw6rx_mUC

RVU Sverige. (2023). *The swedish national travel survey* (research rep.). RVU Sverige. https://www.trafa.se/globalassets/statistik/resvanor/2023/kvalitetsdeklaration-resvanor-i-sverige-2023.pdf

Safi, H., Assemi, B., Mesbah, M., & Ferreira, L. (2016). Trip detection with smartphone-assisted collection of travel data. *Transportation research record*, *2594*, 18–26. https://doi.org/10.3141/2594-03

Safi, H., Assemi, B., Mesbah, M., & Ferreira, L. (2017). An empirical comparison of four technology-mediated travel survey methods. *Journal of Traffic and Transportation Engineering (English Edition)*, *4*, 80–87. https://doi.org/10.1016/j.jtte.2015.12.003

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, *26*, 43–49. https://doi.org/10.1109/TASSP.1978.1163055

Sammer, G., Gruber, C., Roeschel, G., Tomschy, R., & Herry, M. (2018). The dilemma of systematic underreporting of travel behavior when conducting travel diary surveys–a meta-analysis and methodological considerations to solve the problem. *Transportation research procedia*, *32*, 649–658. https://www.sciencedirect.com/science/article/pii/S2352146518301571

Schouten, B., Lugtig, P., & Luiten, A. (2025). Can smart surveys have a positive business case? an evaluation based on three case studies. *Journal of official statistics*, *41*(2), 547–568. https://doi.org/10.1177/0282423x251321634

Schuessler, N., & Axhausen, K. (2009). Map-matching of GPS traces on high-resolution navigation networks using the multiple hypothesis technique (MHT). *Arbeits-*

berichte Verkehrs-und Raumplanung, *568*, 1–22. https://www.research-collection.ethz.ch/handle/20.500.11850/19956

Servizi, V., Pereira, F. C., Anderson, M. K., & Nielsen, O. A. (2021). Transport behavior-mining from smartphones: A review. *European transport research review*, *13*. https://doi.org/10.1186/s12544-021-00516-z

Sfeir, G., Rodrigues, F., Abou-Zeid, M., & Pereira, F. C. (2024). Analyzing the reporting error of public transport trips in the danish national travel survey using smart card data. *Transportation*, 1–30. https://doi.org/10.1007/s11116-024-10535-z

Shen, L., & Stopher, P. R. (2014). Review of GPS travel survey and GPS data-processing methods. *Transport Reviews*, *34*, 316–334. https://doi.org/10.1080/01441647.2014.903530

Shen, Y., Li, W., Xu, G., & Li, B. (2014). Spatiotemporal filtering of regional GNSS network's position time series with missing data using principle component analysis. *Journal of geodesy*, *88*, 1–12. https://doi.org/10.1007/s00190-013-0663-y

Siddique, J., & Belin, T. R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in medicine*, *27*, 83–102. https://doi.org/10.1002/sim.3001

Silber, H., Keusch, F., Breuer, J., Siegers, P., Beuthner, C., Stier, S., Gummer, T., & Weiß, B. (2021). Linking surveys and digital trace data: Insights from two studies on determinants of data sharing behavior. *SocArXiv Papers*. https://doi.org/10.31235/osf.io/dz93u

Smit, R., Zondag, B., & Willigers, J. (2021). Growth model 4: The new dutch national passenger transport model. *European Transport Conference*. https://significance.nl/wp-content/uploads/2022/03/2021-BZO-GM4-The-new-Dutch-national-passenger-transport-model.pdf

Stanley, K., Yoo, E.-H., Paul, T., & Bell, S. (2018). How many days are enough?: Capturing routine human mobility. *International journal of geographical information science: IJGIS*, *32*, 1485–1504. https://doi.org/10.1080/13658816.2018.1434888

Stedman, R. C., Connelly, N. A., Heberlein, T. A., Decker, D. J., & Allred, S. B. (2019). The end of the (research) world as we know it? understanding and coping with declining response rates to mail surveys. *Society & natural resources*, *32*, 1139–1154. https://doi.org/10.1080/08941920.2019.1587127

Stephens, S., Beyene, J., Tremblay, M. S., Faulkner, G., Pullnayegum, E., & Feldman, B. M. (2018). Strategies for dealing with missing accelerometer data. *Rheumatic diseases clinics of North America*, *44*, 317–326. https://doi.org/10.1016/j.rdc.2018.01.012

Stone, A. A., Schneider, S., Smyth, J. M., Junghaenel, D. U., Wen, C., Couper, M. P., & Goldstein, S. (2023). Shedding light on participant selection bias in ecological momentary assessment (EMA) studies: Findings from an internet panel study. *PloS one*, *18*, e0282591. https://doi.org/10.1371/journal.pone.0282591

Stopher, P., & Greaves, S. (2010). Missing and inaccurate information from travel surveys: Pilot results. *Institute of Transport and Logistics Studies Working*

Paper. https://ses.library.usyd.edu.au/bitstream/2123/19374/1/itls-wp-10-07.pdf

Stopher, P., & Shen, L. (2011). In-depth comparison of global positioning system and diary records. *Transportation research record*, *2246*, 32–37. https://doi.org/10.3141/2246-05

Storesund Hesjevoll, I., Fyhri, A., & Ciccone, A. (2021). App-based automatic collection of travel behaviour: A field study comparison with self-reported behaviour. *Transportation Research Interdisciplinary Perspectives*, *12*. https://doi.org/10.1016/j.trip.2021.100501

Struminskaya, B., Lugtig, P., Keusch, F., & Höhne, J. K. (2020). Augmenting surveys with data from sensors and apps: Opportunities and challenges. *Social science computer review*, 0894439320979951. https://doi.org/10.1177/0894439320979951

Stutz, P. (2019). *Enhancing validity long-term travel diary study gnss-data evaluate dose mobility daily commuting*.

Sun, B., Ma, L., Cheng, W., Wen, W., Goswami, P., & Bai, G. (2017). An improved k-nearest neighbours method for traffic time series imputation. *2017 Chinese Automation Congress (CAC)*, 7346–7351. https://doi.org/10.1109/CAC.2017.8244105

Susilo, Y. O., Liu, C., & Börjesson, M. (2019). The changes of activity-travel participation across gender, life-cycle, and generations in sweden over 30 years. *Transportation*. https://link.springer.com/article/10.1007/s11116-018-9868-5

Swenson, L. S., Grimwood, J. M., & Alexander, C. C. (1966). This new ocean: A history of project mercury. *4201*. https://books.google.nl/books?hl=nl&lr=&id=mHlBAAAAIAAJ&oi=fnd&pg=PR11&dq=this+new+ocean+a+history+of+project+mercury&ots=0g13zCcsQB&sig=qIbiwj20R2PFmMAxy6wp7JdV7R0

Tanaka, A., Tateiwa, N., Hata, N., Yoshida, A., Wakamatsu, T., Osafune, S., & Fujisawa, K. (2021). Offline map matching using time-expanded graph for low-frequency data. *Transportation Research Part C: Emerging Technologies*, *130*, 103265. https://doi.org/10.1016/j.trc.2021.103265

Thierry, B., Stanley, K., Kestens, Y., Winters, M., & Fuller, D. (2024). Comparing location data from smartphone and dedicated global positioning system devices: Implications for epidemiologic research. *American journal of epidemiology*, *193*, 180–192. https://doi.org/10.1093/aje/kwad176

Thomas, T., Geurs, K. T., Koolwaaij, J., & Bijlsma, M. (2018). Automatic trip detection with the dutch mobile mobility panel: Towards reliable multiple-week trip registration for large samples. *Journal of urban technology*, *25*, 143–161. https://doi.org/10.1080/10630732.2018.1471874

Tormene, P., Giorgino, T., Quaglini, S., & Stefanelli, M. (2009). Matching incomplete time series with dynamic time warping: An algorithm and an application to post-stroke rehabilitation. *Artificial intelligence in medicine*, *45*, 11–34. https://doi.org/10.1016/j.artmed.2008.11.007

Uğurel, E., Guan, X., Wang, Y., Huang, S., Wang, Q., & Chen, C. (2024). Correcting missingness in passively-generated mobile data with multi-task gaussian processes. *Transportation research. Part C, Emerging technologies*, *161*, 104523. https://doi.org/10.1016/j.trc.2024.104523

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/JSS.V045.I03

van Buuren, S. (2018, July 17). *Flexible imputation of missing data, second edition*. CRC Press.

van Buuren, S., & Oudshoorn, C. G. M. (2000). Multivariate imputation by chained equations. https://publications.tno.nl/publication/34618573/AMELNR/buuren-2000-multivariate.pdf

Verzosa, N., Greaves, S., & Ellison, R. (2017). Smartphone-based travel surveys: A review. https://ses.library.usyd.edu.au/handle/2123/19540

Wang, F., Wang, J., Cao, J., Chen, C., & Ban, X. J. (2019). Extracting trips from multi-sourced data for mobility pattern analysis: An app-based data example. *Transportation research. Part C, Emerging technologies*, *105*, 183–202. https://doi.org/10.1016/j.trc.2019.05.028

Wang, Z., He, S. Y., & Leung, Y. (2018). Applying mobile phone data to travel behaviour research: A literature review. *Travel behaviour & society*, *11*, 141–155. https://doi.org/10.1016/j.tbs.2017.02.005

Witlox, F. (2007). Evaluating the reliability of reported distance data in urban travel behaviour analysis". *Journal of Transport Geography*, *15*, 172–183.

Wojtusiak, J., & Mogharab Nia, R. (2021). Location prediction using GPS trackers: Can machine learning help locate the missing people with dementia? *Internet of Things*, *13*, 100035. https://doi.org/10.1016/j.iot.2019.01.002

Wolf, J., Oliveira, M., & Thompson, M. (2003). The impact of trip underreporting on VMT and travel time estimates: Preliminary findings from the california statewide household travel survey GPS study. *Transportation research record*, *1854*, 189–198. https://www.researchgate.net/profile/Marcelo_Simas_Oliveira/publication/228604174_The_impact_of_trip_underreporting_on_VMT_and_travel_time_estimates_preliminary_findings_from_the_California_statewide_household_travel_survey_GPS_study/links/55e458c508ae2fac4721ef49.pdf

Xie, P., Li, T., Liu, J., Du, S., Yang, X., & Zhang, J. (2020). Urban flow prediction from spatiotemporal data using machine learning: A survey. *An international journal on information fusion*, *59*, 1–12. https://doi.org/10.1016/j.inffus.2020.01.002

Yang, F., Yao, Z., Cheng, Y., Ran, B., & Yang, D. (2016). Multimode trip information detection using personal trajectory data. *Journal of Intelligent Transportation Systems*, *20*, 449–460. https://doi.org/10.1080/15472450.2016.1151791

Yang, Y., Jin, M., Wen, H., Zhang, C., Liang, Y., Ma, L., Wang, Y., Liu, C., Yang, B., Xu, Z., Bian, J., Pan, S., & Wen, Q. (2024). A survey on diffusion models for time series and spatio-temporal data. *arXiv [cs.LG]*. http://arxiv.org/abs/2404.18886

Yoo, E.-H., Roberts, J. E., Eum, Y., & Shi, Y. (2020). Quality of hybrid location data drawn from GPS□enabled mobile phones: Does it matter? *Transactions in GIS*, *90*, 187. https://doi.org/10.1111/tgis.12612

Zhang, B., Rasouli, S., & Feng, T. (2024). Social demographics imputation based on similarity in multi-dimensional activity-travel pattern: A two-step approach. *Travel behaviour & society*, *37*, 100843. https://doi.org/10.1016/j.tbs.2024.100843

Zhang, Y., & Thorburn, P. J. (2021). A dual-head attention model for time series data imputation. *Computers and Electronics in Agriculture*, *189*, 106377. https://doi.org/10.1016/j.compag.2021.106377

Zhao, F., Ghorpade, A., Pereira, F. C., Zegras, C., & Ben-Akiva, M. (2015). Stop detection in smartphone-based travel surveys. *Transportation Research Procedia*, *11*, 218–226. https://doi.org/10.1016/j.trpro.2015.12.019

Zhao, P., Jonietz, D., & Raubal, M. (2021). Applying frequent-pattern mining and time geography to impute gaps in smartphone-based human-movement data. *Geographical Information Systems*, 1–29. https://doi.org/10.1080/13658816.2020.1862126

Zhao, Z., Yin, L., Shaw, S.-L., Fang, Z., Yang, X., & Zhang, F. (2018). Identifying stops from mobile phone location data by introducing uncertain segments. *Transactions in GIS: TG*, *22*, 958–974. https://doi.org/10.1111/tgis.12332

Zheng, Y., Li, Q., Chen, Y., Xie, X., & Ma, W.-Y. (2008). Understanding mobility based on GPS data. *Proceedings of the 10th international conference on Ubiquitous computing*, 312–321. https://doi.org/10.1145/1409635.1409677

Zhou, H., Wang, H., Zhou, Y., Luo, X., Tang, Y., Xue, L., & Wang, T. (2020). Demystifying diehard android apps. *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*. https://doi.org/10.1145/3324884.3416637

Zhu, L., Boissy, P., Duval, C., Zou, G., Jog, M., Montero-Odasso, M., & Speechley, M. (2022). How long should GPS recording lengths be to capture the community mobility of an older clinical population? a parkinson's example. *Sensors*, *22*. https://doi.org/10.3390/s22020563

A

Appendix to Chapter 2

Table A.1 *Device registration by full sample characteristics*

	Unregistered (*N*=1228)	Registered (*N*=674)	p value
Age			$< 0.01^1$
N-Miss	6	0	
M (*SD*)	50.4 (19.0)	44.3 (17.1)	
Range	15.0 - 96.0	15.0 - 90.0	
AgeCat			$< 0.01^2$
N-Miss	6	0	
15-20	90 (7.4%)	63 (9.3%)	
21-30	150 (12.3%)	111 (16.5%)	
31-40	157 (12.8%)	117 (17.4%)	
41-50	186 (15.2%)	124 (18.4%)	
51-60	229 (18.7%)	118 (17.5%)	
61-70	208 (17.0%)	102 (15.1%)	
>70	202 (16.5%)	39 (5.8%)	
Gender			$< 0.77^3$
N-Miss	6	0	
Male	606	329	
Female	616	345	
Origin			$< 0.01^2$
N-Miss	6	0	
Dutch	961 (78.6%)	577 (85.6%)	
Non-Western	130 (10.6%)	40 (5.9%)	
Western	131 (10.7%)	57 (8.5%)	
Generation			$< 0.01^2$
N-Miss	6	0	
Dutch	961 (78.6%)	577 (85.6%)	

	Unregistered (*N*=1228)	Registered (*N*=674)	p value
First	149 (12.2%)	40 (5.9%)	
Second	112 (9.2%)	57 (8.5%)	
Marital Status			$< 0.01^2$
N-Miss	6	0	
Married	624 (51.1%)	355 (52.7%)	
Never married	410 (33.6%)	264 (39.2%)	
Divorced	121 (9.9%)	49 (7.3%)	
Widow/Widower	67 (5.5%)	6 (0.9%)	
Education			$< 0.01^4$
N-Miss	518	190	
Elementary School	63 (8.9%)	17 (3.5%)	
High School	282 (39.7%)	187 (38.6%)	
Vocational School	155 (21.8%)	59 (12.2%)	
University	136 (19.2%)	146 (30.2%)	
Graduate School	74 (10.4%)	75 (15.5%)	
Household Type			$< 0.01^2$
N-Miss	6	0	
One-Person HH	244 (20.0%)	92 (13.6%)	
Partners, no child	420 (34.4%)	234 (34.7%)	
Partners, child	473 (38.7%)	310 (46.0%)	
Single parent	79 (6.5%)	35 (5.2%)	
Other household	6 (0.5%)	3 (0.4%)	
Urbanicity			0.55^4
Very high	259 (21.2%)	144 (21.4%)	
High	316 (25.9%)	171 (25.4%)	
Moderate	239 (19.6%)	136 (20.2%)	
Slight	208 (17.0%)	128 (19.0%)	
Rural	200 (16.4%)	95 (14.1%)	
Province			0.20^2
N-Miss	6	0	
Groningen	37 (3.0%)	28 (4.2%)	
Friesland	46 (3.8%)	26 (3.9%)	
Drenthe	47 (3.8%)	12 (1.8%)	
Overijssel	75 (6.1%)	45 (6.7%)	
Flevoland	31 (2.5%)	13 (1.9%)	
Gelderland	149 (12.2%)	88 (13.1%)	
Utrecht	85 (7.0%)	59 (8.8%)	
Noord-Holland	206 (16.9%)	103 (15.3%)	
Zuid-Holland	256 (20.9%)	127 (18.8%)	
Zeeland	27 (2.2%)	13 (1.9%)	
Noord-Brabant	179 (14.6%)	115 (17.1%)	
Limburg	84 (6.9%)	45 (6.7%)	
Std. Income Pct.			$< 0.01^1$
N-Miss	20	5	

	Unregistered (*N*=1228)	Registered (*N*=674)	p value
Mean (SD)	55.4 (27.7)	64.3 (25.6)	
Range	0.0 - 100.0	0.0 - 100.0	
Has car			< 0.11[3]
N-Miss	6	0	
No	665 (54.4%)	341 (50.6%)	
Yes	557 (45.6%)	333 (49.4%)	
Has moped			< 1.00[3]
N-Miss	6	0	
No	1155 (94.5%)	637 (94.5%)	
Yes	67 (5.5%)	37 (5.5%)	
Has drivers license			< 0.01[3]
N-Miss	6	0	
No	281 (23.0%)	102 (15.1%)	
Yes	941 (77.0%)	572 (84.9%)	

Note. [1] Linear Model ANOVA [2] Fisher's Exact Test for Count Data with simulated p-value (based on 500 replicates)[3] Fisher's Exact Test for Count Data [4] Trend test for ordinal variables

B

Appendix to Chapter 3

B.1. Additional algorithms

Algorithm 5 Linear interpolation

```
Input: locs[1, ..., N] sorted on time
Output: locs′[1, ..., N], with gaps linearly interpolated
function interpolate(locs[1, ..., N])
    for all locs[i, i+1] $\rightarrow (loc1, loc2)$ ∈locs[] do
        if timeDiff($loc1_{time}, loc2_{time}$) $> time_{min}$ and haversineDist($loc1, loc2$) $> dist_{min}$ then
            $\Delta_{time} \leftarrow$ timeDiff($loc1_{time}, loc2_{time}$)S
            $\Delta_{long} \leftarrow (loc2_{long} - loc1_{long})/\Delta_{time}$
            $\Delta_{lat} \leftarrow (loc2_{lat} - loc1_{lat})/\Delta_{time}$
        end if
        for all $minute \in 1$ to $\Delta_{time}$ do
            def $newloc$
            $newloc_{time} \leftarrow loc1_{time} + minute$
            $newloc_{long} \leftarrow loc1_{long} + \delta_{long} \times minute$
            $newloc_{lat} \leftarrow loc1_{lat} + \delta_{lat} \times minute$
            Add $newloc$ to locs[]
        end for
    end for
     return locs′[]
end function
```

Algorithm 6 Stop Resolver

```
Input: locs[1, ..., N]
Output: locs′[1, ..., N], with stop id annotation
function resolveStops(locs[])
    $i \leftarrow 1$
    $j \leftarrow i + 1$
    while $i < N$ do
        while $j < N$ do
            $\Delta_{dist} \leftarrow$ haversineDist(locs[i], locs[j])
            if $\Delta_{dist} \geq dist_{max}$ then
                $\Delta_{time} \leftarrow$ timeDiff(locs[i], locs[j-1])
                if $\Delta_{time} \geq time_{min}$ then
                    Annotate locs[i, ..., j] with
                        stop id $\leftarrow i$
                end if
                $i \leftarrow j$
                break
            else
                $j \leftarrow j + 1$
            end if
        end while
    end while
    if timeDiff(locs[i], locs[j-1]) $\geq time_{min}$ then
        Annotate locs[i, ..., j] with
            stop id $\leftarrow i$
    end if
     return locs′[]
end function
```

B.2. Mixed models for personal covariates

Table B.1 *Percent bias in distance estimation by q and personal covariates*

	Dependent variable: ‘Distance % bias‘				
	Gender ref: Male	Age	Urbanicity ref: Rural	Ethnicity ref: Dutch	Education ref: Primary
Constant	−0.044**	−0.059**	−0.050**	−0.042**	−0.005
	(0.004)	(0.008)	(0.007)	(0.003)	(0.022)
q	0.810**	0.810**	0.812**	0.813**	0.819**
	(0.020)	(0.020)	(0.020)	(0.021)	(0.025)
Female	0.001				
	(0.005)				
Age		0.0003*			
		(0.0002)			
Urbanicity					
Slight			0.002		
			(0.009)		
Moderate			0.007		
			(0.009)		
High			0.013		
			(0.008)		
Very high			0.004		
			(0.009)		
Origin					
Western				−0.019	
				(0.011)	
Other				−0.007	
				(0.011)	
Education					
Secondary					−0.048*
					(0.023)
Trade School					−0.048
					(0.029)
Bachelors					−0.044
					(0.022)
Graduate					−0.038
					(0.023)
N_{sets}	60842	60842	60293	60293	41841
N_{days}	554	554	549		
N_{users}	184	184	183	183	131
$\sigma^2_{sets}/\sigma^2_{tot}$	29%	29%	29%	46%	46%
$\sigma^2_{days}/\sigma^2_{tot}$	55%	55%	56%		
$\sigma^2_{users}/\sigma^2_{tot}$	15%	15%	15%	54%	54%
LL	972.180	971.062	887.579	−4,062.347	−2,659.438

Note. Mixed effect models were used to account for clustered errors and total N.
*p<0.05; **p<0.01

7

Table B.2 *Percent bias in number of moves estimation by q and personal covariates*

	Dependent variable: 'Moves % bias'				
	Gender ref: Male	Age	Urbanicity ref: Rural	Ethnicity ref: Dutch	Education ref: Primary
Constant	-0.019^{**}	-0.026^{**}	-0.031^{**}	-0.020^{**}	0.005
	(0.004)	(0.009)	(0.008)	(0.003)	(0.022)
q	0.906^{**}	0.906^{**}	0.906^{**}	0.906^{**}	0.916^{**}
	(0.019)	(0.019)	(0.019)	(0.019)	(0.022)
Female	-0.007				
	(0.006)				
Age		0.0001			
		(0.0002)			
Urbanicity					
Slight			-0.003		
			(0.010)		
Moderate			0.023^{*}		
			(0.010)		
High			0.014		
			(0.010)		
Very high			0.008		
			(0.010)		
Origin					
Western				-0.016	
				(0.012)	
Other				-0.006	
Education					
				(0.012)	
Secondary					-0.035
					(0.022)
Trade School					-0.020
					(0.029)
Bachelors					-0.032
					(0.022)
Graduate					-0.025
					(0.022)
N_{sets}	58204	58204	57655	57655	39861
N_{days}	530	530	525	525	
N_{users}	182	182	181	181	130
$\sigma^2_{sets}/\sigma^2_{tot}$	38%	38%	38%	38%	55%
$\sigma^2_{days}/\sigma^2_{tot}$	48%	48%	48%	48%	
$\sigma^2_{users}/\sigma^2_{tot}$	14%	14%	14%	14%	45%
LL	$-2{,}953.724$	$-2{,}957.686$	$-2{,}971.189$	$-2{,}966.332$	$-3{,}638.624$

Note. Mixed effect models were used to account for clustered errors and total N.
*p<0.05; **p<0.01

Table B.3 *Percent bias in RoG estimation by q and personal covariates*

	Dependent variable: ‘RoG % bias‘				
	Gender ref: Male	Age	Urbanicity ref: Rural	Ethnicity ref: Dutch	Education ref: Primary
Constant	−0.069**	−0.099**	−0.074**	−0.070**	−0.026
	(0.005)	(0.009)	(0.008)	(0.004)	(0.026)
q	0.750**	0.749**	0.752**	0.752**	0.748**
	(0.021)	(0.021)	(0.021)	(0.021)	(0.025)
Female	−0.006				
	(0.006)				
Age		0.001**			
		(0.0002)			
Urbanicity					
Slight			−0.006		
			(0.011)		
Moderate			0.003		
			(0.011)		
High			0.011		
			(0.010)		
Very high			−0.001		
			(0.011)		
Origin					
Western				−0.017	
				(0.013)	
Other				−0.009	
				(0.013)	
Education					
Secondary					−0.055*
					(0.026)
Trade School					−0.074*
					(0.034)
Bachelors					−0.048
					(0.026)
Graduate					−0.048
					(0.026)
N_{sets}	60842	60842	60293	60293	41841
N_{days}	554	554	549	549	381
N_{users}	184	184	183	183	131
$\sigma^2_{sets}/\sigma^2_{tot}$	28%	28%	28%	28%	29%
$\sigma^2_{days}/\sigma^2_{tot}$	58%	58%	58%	58%	54%
$\sigma^2_{users}/\sigma^2_{tot}$	14%	14%	13%	13%	18%
LL	−165.341	−164.320	−292.465	−284.924	262.167

Note. Mixed effect models were used to account for clustered errors and total N.
*p<0.05; **p<0.01

C

Appendix to Chapter 4

C.1. Parameter selection

The initial simulation study was performed in order to determine the appropriate parameters described in the previous section. We selected the following two key travel metrics for evaluation of performance: total distance and number of stops. We compare the parameters on the basis ofRMSE and mean absolute bias (Bias), as well as on a set of metrics developed to assess the accuracy and directional bias of the imputed travel distance and number of periods spent moving. RMSE and Bias both assess the accuracy of the underlying imputed metric in absolute terms. Because different parameter sets generated either a significant upward bias or downward bias on total distance, we compared under- and overestimation separately.

Travel Periods Overestimated (TP ↑) reflects the percentage of 15-minute travel periods imputed that did not exist in the true data set. Conversely, Travel Periods Underestimated (TP ↓) reflects the percentage of the true number of periods that were spent in movement that were not reflected in the imputation. Travel Period Accuracy (TP Acc.) reflects the percentage agreement with the total count of moving/stationary periods between the true data and the imputed data.

Distance Overestimated (Dist ↑) reflects only the upward bias of the imputed values relative to the true distance. Distance Underestimated (Dist ↑) similarly reflects only the downward bias of the imputed values relative to the true distance. Both are expressed in kilometers.

Table C.1 shows the mean performance of each parameter option across the simulations for distance. Overall, a match buffer of 1 hour is preferred across most metrics and provides the smallest bias in distance. A high candidate specificity is generally preferred overall, although the overall bias is larger. Time window has less of a clear pattern when examining the mean performance. Most metrics prefer a higher

Table C.1 *Comparison among possible DTWBMI parameter values*

		Bias (km)	Dist ↑	Dist ↓	RMSE	TP Acc.	TP ↑	TP ↓
Cand. Spec.	Low	1.9	4.1	5.7	1.317	95%	16%	18%
	Medium	1.7	3.9	5.4	1.315	95%	15%	17%
	High	1.6	4.0	5.2	1.327	95%	15%	17%
Match Buffer	1 hour	1.0	3.9	4.2	1.317	95%	16%	14%
	4 hours	1.6	4.0	5.7	1.341	95%	15%	18%
	8 hours	2.5	4.0	6.4	1.301	95%	15%	20%
Time Window	< 1 hour	2.3	3.7	5.9	1.300	95%	14%	18%
	< 3 hours	1.6	3.9	5.5	1.320	95%	15%	17%
	No Window	1.3	4.3	4.9	1.340	95%	16%	17%
N Imps	1	2.0	3.8	5.5	1.306	95%	15%	17%
	3	1.5	4.1	5.4	1.326	95%	15%	17%
	5	1.6	4.0	5.4	1.325	95%	15%	17%
	10	1.7	4.0	5.4	1.322	95%	15%	17%

number of imputations, although the relationship between the average distance bias and number of imputations prefers a smaller number of imputations.

The performance of these metrics is dependent both upon the other metrics in the simulation model as well as characteristics of the data and its missingness. Figure C.1 shows mean absolute bias, RMSE and similarity values for the imputation of total distance traveled. The figure shows the interactions in these metrics across all parameter values for Candidate Specificity (CS), Match Buffer (MB), Time Window (TW), and Number of Imputations (N Imps). The dashed line demonstrates best relative performance.

We first consider absolute bias (AbsBias) in Figure C.1. MBMatch Buffers of one hour demonstrate the lowest absolute bias in meters, with a corresponding increase of absolute bias as the length of the match buffer increases. A low Candidate Specificity (CS) has an increased bias relative to medium and high when the match buffer increases. Imputing with a MB of 8 hours is less biased with a higher CS, while imputing with a MB of 4 hours is less biased with medium CS. Differences between the various time windows emerge across the different match buffer levels. As the length of the Match Buffer increases, the 1-hour Time Window has reduced performance with respect to bias. The number of imputations does not have a consistent relationship with the bias.

Next, we consider the distance over- and underestimated across the various parameters. As is shown in Figure C.1, the general relationship is that parameters combinations that decrease overestimation increase underestimation. Conversely, parameter combinations that decrease underestimation, increase overestimation. As the MB length increases, the underestimation of travel distance increases. However, the reverse trend is less clear, as a MB of 8 hours does not provide in aggregate a low level of distance overestimation, but rather ranges between 3 and 6 kilometers on average. The relationship of CS level to bias is unclear. An unrestricted Time Window provides less underestimation across all levels, while a 1-hour TW increases underestimation. A 1-hour TW does not, however, appear to offer the best performance with respect to overestimation, as this varies across the other parameters. Lastly, we identify no relationship between the number of imputations and the underestimation, but a slight increasing relationship between the number of imputations within the medium candidate specificity condition.

Travel Period Accuracy (TP Acc.) remains a relatively consistent 95% across most condition combinations. When the MB is 8 hours, an unrestricted TW seems to offer worse performance in this metric. In aggregate, a MB of 1 hour offers the best performance. No clear patterns emerge for Candidate Specificity or Number of Imputations.

Travel period over- and underestimation differ slightly from the interpretation of the total distance over- and underestimated. The relationship for MB is largely the same, with a modest increase in rate of travel period underestimation in the 4-hour and 8-hour conditions as compared to the 1-hour condition. Similarly, there is a reduction in aggregate in travel period overestimation for MBMatch Buffers of 4 and 8 hours relative to 1 hour. No clear pattern is evident for CS. On average, unrestricted TWTime Windows provide more overestimation and less underestimation, while the most restrictive time window condition provides less overestimation and more underestimation of the total number of travel periods. No clear pattern emerges for Number of Imputations.

C.2. Impact of number of own sets

Given the assumption that an individual's prior travel behavior often serves as the most appropriate imputation candidate for their own missing data, we executed an additional simulation study. This was done with the aim of evaluating the effects of increasing the availability of self-referential data sets, or "own sets".

We selected the full subset of 16 participants who had a minimum of four available sets. For each of these individuals, we randomly selected one set into which missing data was inserted. This missing data was introduced as a single contiguous block, initiating at a randomly chosen suitable location within the set. The missingness was introduced at five different levels, namely one hour, three hours, six hours, eight hours and twelve hours, thereby establishing five distinct gap lengths within the missingness condition.

In the own data condition, we created four levels: zero self-sets, one self-set, two self-sets and three self-sets. Each own data condition incorporated a base reference set consisting of 48 randomly selected sets from individuals with fewer than four datasets. This was done to ensure an ample quantity of imputation candidates. For each condition, the unused self-sets were replaced with additional reference sets from individuals having fewer than four data sets, in order to maintain a consistent number of total available candidates.

We conducted ten simulations wherein DTWBMI-HI, DTWBMI-LO and DTWBI were applied to the generated data set. These simulations were evaluated based on the performance criteria delineated in section 4.3.3 to determine their relative performance on travel distance imputation.

Table C.2 shows aggregate results across the own set condition. Both DTWBMI-HI and DTWBMI-LO show a reduction in absolute bias corresponding to an increase in the number of own sets. RMSE and TP ↑, TP ↓, and TP Acc. remain mostly stable. The distance overestimated increases slightly, while the distance underestimated decreases slightly for both methods. Median bias is excluded from this table, but was 0 across both methods and all conditions. Overall, it seems that increasing the number of own sets generally improves the performance of both imputation methods, with DTWBMI-LO consistently performing better than DTWBMI-HI based on the presented metrics.

Table C.3 shows the breakout across both gap length condition and number of self-sets available. As gap length increases, the performance boost for both DTWBMI methods increases in total bias reduction. Similar patterns are demonstrated within each level of missing data: Absolute bias reduces with the increase of the number of own sets, RMSE remains stable or shows evidence of a slight decrease, and Travel Period metrics remain largely stable. Overall, DTWBMI-LO outperforms DTWBMI-LO.

Although we expected DTWBMI-HI to outperform DTWBMI-LO in situations where it was advantageous to match with a greater specificity to the shape of the behavior, this simulation study does not provide evidence indicating that a small increase in

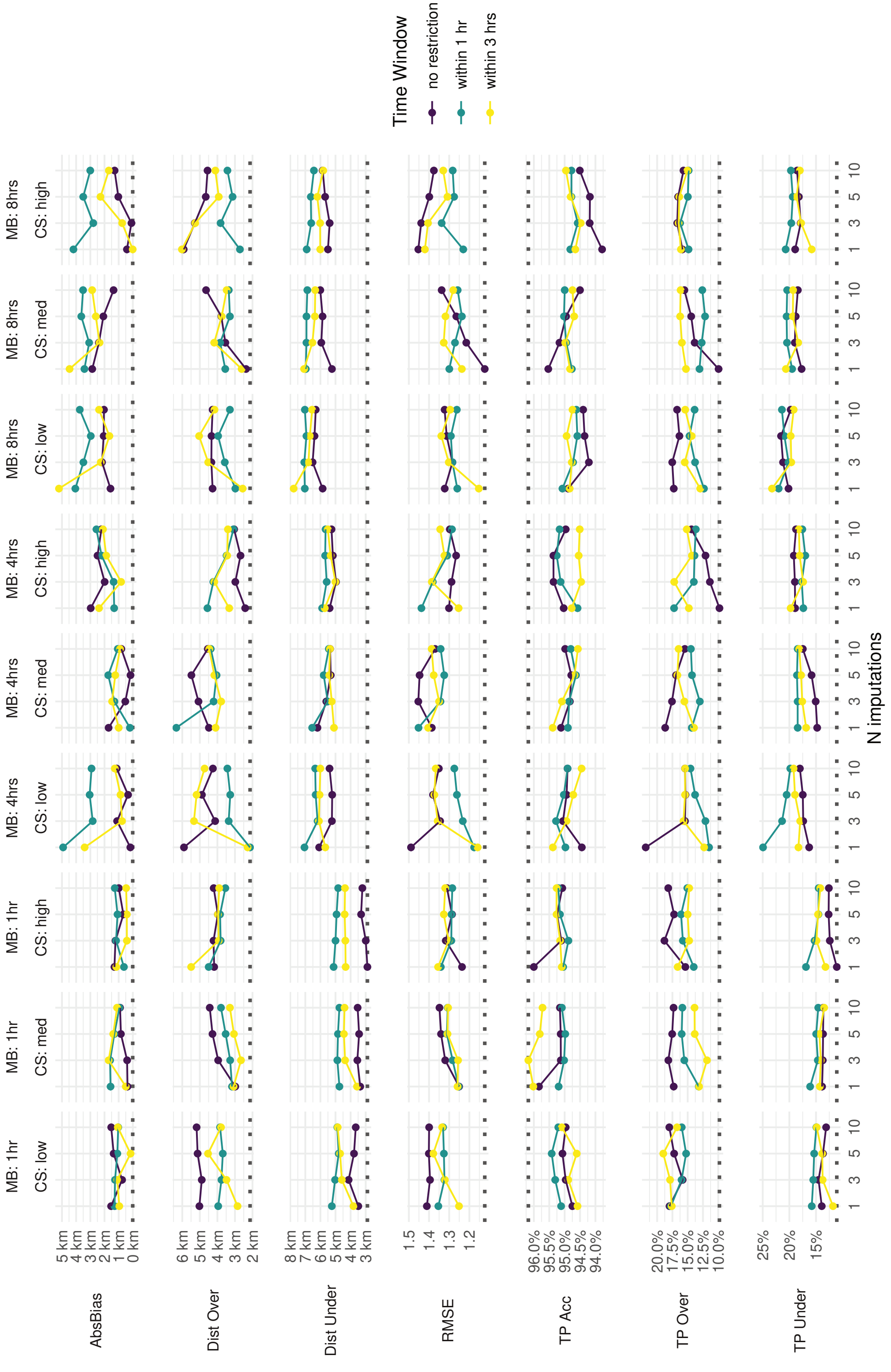


Figure C.1 *Visual comparison of DTWBMI parameters across six measures of performance. Match Buffer (MB) and Candidate Specificity (CS) are represented by facets. The Time Window parameter is represented by shape and color. The dotted gray line indicates the best relative performance across all simulation parameter combinations within each measure.*

Table C.2 *Method comparison across number of own reference sets*

Sets	Model	Bias (km)	RMSE	TP ↑	TP ↓	TP Acc.	Dist ↑	Dist ↓
0	DTWBMI-HI	5.6	33.16	14.0%	17.2%	94.1%	4.4	10.0
	DTWBMI-LO	3.9	24.88	11.4%	12.9%	95.6%	3.1	7.0
1	DTWBMI-HI	4.4	34.43	14.6%	16.4%	94.1%	5.2	9.6
	DTWBMI-LO	3.0	23.12	11.7%	12.1%	95.6%	3.4	6.4
2	DTWBMI-HI	3.8	32.69	13.9%	15.7%	94.2%	5.0	8.8
	DTWBMI-LO	2.2	24.36	12.1%	11.9%	95.6%	3.9	6.2
3	DTWBMI-HI	3.4	30.65	13.5%	16.6%	94.3%	5.1	8.5
	DTWBMI-LO	2.1	23.69	12.1%	12.1%	95.6%	3.8	6.0

the available historical data for a person would be sufficient to prefer the use of the high-information method to the low-information method.

Table C.3 *Method comparison across gap length and number of own reference sets*

Gap	Model	Sets	Bias	RMSE	TP ↑	TP ↓	TP Acc.	Dist ↑	Dist ↓
			(Km)					(Km)	(Km)
1hr	DTWBMI-HI	0	1.7	11.9	3.7%	5.7%	94.6%	0.4	2.1
		1	1.7	11.3	2.7%	4.9%	95.7%	0.2	2.0
		2	2.0	11.8	3.8%	5.6%	94.9%	0.3	2.3
		3	2.0	12.0	2.5%	5.8%	95.5%	0.2	2.2
	DTWBMI-LO	0	1.3	9.9	2.1%	2.7%	97.3%	0.1	1.4
		1	1.3	9.4	2.7%	2.3%	97.3%	0.1	1.4
		2	0.8	7.2	2.4%	1.8%	97.8%	0.2	1.0
		3	0.8	7.8	2.9%	2.1%	97.4%	0.2	1.0
3hr	DTWBMI-HI	0	2.8	19.4	7.1%	12.5%	95.3%	1.1	4.0
		1	2.9	18.5	8.5%	13.9%	94.6%	1.2	4.1
		2	2.7	17.0	6.0%	12.4%	95.6%	1.0	3.7
		3	2.4	19.0	7.3%	12.4%	95.1%	1.5	3.8
	DTWBMI-LO	0	1.5	12.5	6.8%	8.4%	95.9%	1.0	2.4
		1	0.6	9.7	6.5%	8.4%	96.0%	1.3	1.9
		2	0.5	11.0	8.4%	8.4%	95.8%	1.4	1.9
		3	0.6	10.9	7.9%	7.8%	95.8%	1.3	1.9
6hr	DTWBMI-HI	0	6.1	27.3	13.4%	18.4%	94.4%	2.6	8.6
		1	6.0	26.6	14.5%	19.1%	94.2%	2.6	8.6
		2	4.0	29.4	17.5%	18.1%	93.5%	4.3	8.3
		3	4.0	24.0	13.6%	18.2%	94.1%	3.3	7.2
	DTWBMI-LO	0	3.0	17.2	12.0%	14.7%	95.2%	2.3	5.3
		1	1.9	16.9	10.8%	12.7%	95.5%	2.5	4.4
		2	1.1	17.2	12.8%	12.1%	95.6%	3.0	4.1
		3	1.4	17.1	11.6%	12.8%	95.5%	2.8	4.2
8hr	DTWBMI-HI	0	3.2	40.6	22.4%	21.5%	93.9%	7.9	11.1
		1	0.2	50.6	22.1%	19.2%	93.6%	10.7	10.9
		2	1.3	44.9	20.0%	17.7%	94.0%	9.1	10.4
		3	2.0	38.7	21.6%	21.2%	94.1%	8.2	10.2
	DTWBMI-LO	0	1.2	26.7	17.7%	15.9%	95.4%	6.0	7.2
		1	0.9	25.2	19.5%	15.3%	95.0%	6.0	6.9
		2	0.1	27.6	17.5%	15.9%	95.1%	6.3	6.4
		3	0.6	26.8	18.2%	16.9%	95.2%	5.8	6.3
12hr	DTWBMI-HI	0	14.2	66.7	23.5%	27.7%	92.5%	9.8	24.0
		1	11.3	65.2	25.1%	24.8%	92.4%	11.1	22.4
		2	9.1	60.3	22.2%	24.7%	92.8%	10.5	19.6
		3	6.5	59.6	22.3%	25.5%	92.8%	12.3	18.8
	DTWBMI-LO	0	12.7	58.1	18.5%	22.8%	94.1%	6.2	18.8
		1	10.3	54.4	18.8%	21.8%	94.0%	7.1	17.4
		2	8.5	58.9	19.2%	21.2%	93.9%	8.9	17.4
		3	7.4	55.8	19.9%	20.7%	93.9%	9.1	16.5

D

Appendix to Chapter 5

Considerations for DTWBMI

Because respondents may have multiple gaps within a 24-hour period and because some gaps may occur at the beginning or end of the period, not all days can be imputed with DTWBMI. We require that the total sum of covered time before and after the gap is at least 110 minutes. We therefore preemptively exclude all data that occurs on respondent days which, even if successfully imputed, would still not constitute a complete day.[1]. It might be possible in the future to include these days so that they can become reference queries for imputation, even if they will lack sufficient information to be considered themselves.

In requiring that the set of potential data has a buffer of sufficient length on both sides, we restrict the set of potentially imputable days to 802 out of 5328. This is an unfortunate consequence of the division of the dataset into days. The restriction of continuous time into calendar days means that gaps occurring close to or across the 4am cutoff cannot have a buffer. In this case, there are two options: 1) retain the day-based format and implement DTWBMI with a one-sided buffer, or 2) perform DTWBMI prior to splitting the data on calendar days, dividing them after imputation. In this study, we implement the second method.

A secondary issue is that many of the long gaps, theoretically within the purview of DTWBMI, are interspersed with single measurements, which cannot function as a full matching buffer. These are potentially informative measurements, but they do not allow proper function of the imputation model. This issue is illustrated in Figure D.1, where the black line shows the gap time, and the color on either side of it shows the one hour match buffer. Here we will elect to combine intersecting gaps into a single gap, making note of the distance (if any) between successive single points.

[1] A complete day is 95% coverage during the 4am to 4am period, or 1368 minutes

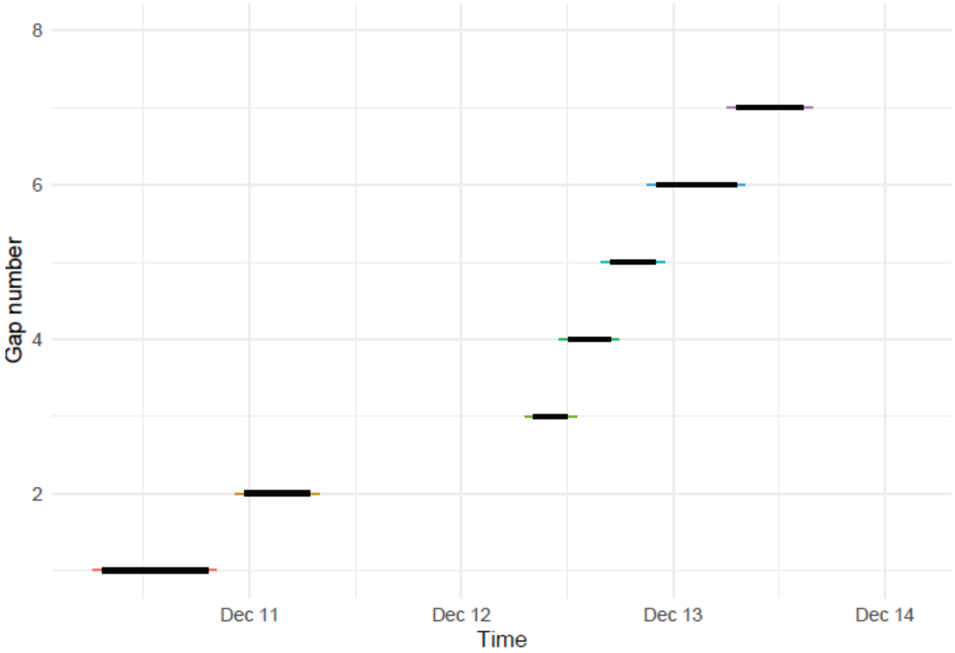


Figure D.1 *Buffering around single points to extend the imputation capabilities without data loss. The black line shows the gap time, and the color on either side of it shows the one hour match buffer.*

Data preparation steps

The data were aggregated from a set of timestamped geolocation coordinates to an imputable time series of travel behaviors in a workflow consisting of multiple sequential steps: 1) data cleaning, 2) partitioning into contiguous user-sets, 3) large gap assessment, 4) stop detection, 5) segmentation, 6) aggregation. Table D.1 provides a comprehensive overview of the parameters choices and reasoning therefore.

Data cleaning

All geolocations with a device-reported accuracy greater than 200 meters were filtered out. This is less conservative than many other studies, but 200 meters was selected in order to ensure that the maximum number of GNSS-provided points were maintained, especially during cold starts. Additionally, a test of empirical accuracy was carried out to establish the likely bounds of the reported accuracy. To do so, the distance between immediately preceding and immediately following points was calculated, provided the points occurred within five minutes' time of the selected point. While these points are also measured with error, we can compare points with a high reported accuracy to points with a low reported accuracy occurring subsequent and test the distance. Results indicated that internal studies suggested that iOS points reporting an accuracy under 200m demonstrated a 95% empirical *CI* = [0.18 m, 17.50 m], and for Android, 95% empirical *CI* = [1.59, 384.21]. Android's stated definition of the accuracy variable is a 68% confidence radius, while our analysis indicated a 68% empirical *CI* = [14.83, 49.91]. For these reasons, and because the following step allowed for greater precision during the filtering step, we opted for an upper boundary of 200 meters.

Following this, trajectory segments were flagged for manual inspection if any contained point was flagged as having an implausible speed. A speed variable was calculated both forward and backward, allowing for a two-second buffer to be accommodating of small timestamp irregularities that could result in two close locations occurring near-simultaneously and therefore registering with an implausible speed. The severity of each discrepancy was flagged with a severity category such that points with a speed >1000 kph were addressed first, followed by >500 kph, followed by 200 kph. Manual inspection was performed using a custom-built Shiny app that allowed for points to be plotted against a map as well as against the timestamp, and selected for filtering. The filtering process removed approximately 20k geolocations from the total set of 17.3m.

User-set segmentation, interpolation, and large gap assessment

For each user, the gap length between subsequent geolocations was calculated, which was used as a mechanism for establishing missingness at various levels. Locations from the same user separated by a gap of over 12 hours were subdivided into their own sets for all future steps. These sets were further split into sets on the points where the gap was greater than five minutes. At this point, a selection was made for the Listwise and Pairwise analyses on the basis of these fully-complete

Table D.1 *Justification for parameter and value sets used in analyses*

	Parameter: Definition	**Values**	**Reference/Justification**
Missingness	**Temporal resolution** τ: Length of the discrete interval evaluated to establish missingness yes or no	5 min	Currey 2023
	Max long gap time: maximum elapsed time gap before a respondent's set is divided into separate series.	12 hr	Practical limit still within theoretically sensible bounds.
	Max short gap time: maximum timespan over which interpolation may be performed	30 min	McCool 2022
	Max short gap distance: Maximum distance over which interpolation may occur	Inf	Previous cleaning steps made this unnecessary, but remains a consideration for uncleaned data.
Data	**Temporal aggregation resolution**: Size of the time window chosen for aggregation	15 min	Practical, ensuring at least two locations at the selected temporal resolution of 5 minutes.
	Min accuracy: Minimum device-labeled accuracy parameter over which the raw data are filtered	200 m	Accuracy >200m due to insufficient GNSS lock-ons or from celltower.
	Segmentation error: Top Down Time Ratio segmentation method error/stopping parameter or percent of points classified as moving/stopped respectively	40/200 m	Meratnia 2004
	Max speed flag: Geolocations with derived speed from previous location were visually inspected by overlaying the surrounding trajectory on a map and flagged as outliers or not	200 kph	Safi 2016
Stop detection	**Stop max radius**: Maximum distance a user can cover in a place to be considered a stop	100 m	Montoliu 2013, Cich 2015
	Stop min time: Minimum time that a user must stay in the same place to be considered a stop	3 min	Montoliu 2013, Cich 2015
	Stop merge distance: Maximum allowable distance between temporally adjacent stop center points to permit merging	100 m	Safi 2016, Montoliu 2013
	Small track max. locs: Maximum number of locations that may make up an implausible small track	1	Safi 2016
DTWBMI	**Matching buffer**: n time points before and after the gap that are used in establishing trajectory similarity	4	McCool 2024, Chakrabarti 2023
	Time window: Maximum allowable time difference between the query and donor set	1 hr	McCool 2024
	Donor candidates: Set of candidate donors from all complete cases with the highest similarity to the query, from which the donor is selected.	3	McCool 2024
	N imputations: The number of independent imputations creating the set of multiple imputations needed analysis	5	Van Buuren 2018 McCool 2024
	N chains: The number of iterations for imputing the data on the basis of the new set of imputed data	4	Set after evaluating trace line convergence on smaller set after 10 iterations.
MICE	**pmm donor candidates**: For variables imputed with predictive mean matching, number of candidate donors selected from	5	Van Buuren 2018
	N chains: Number of chained iterations for revisiting all variables	20	Assessed via convergence plots

sets. A person-day was assigned to each uninterrupted set of records, using 4am as a cutoff moment between days. Person-days with at least 1368 minutes (95% of a 24-hour day) of covered time were marked as fully complete, for use in the Pairwise analysis. Users with at least one complete day for each day of the week were selected for the Listwise deletion set.

Locations from the same user and separated by a gap greater than 30 minutes were flagged for interpolation. Interpolation was carried out using the stopdetection package in R, and built pseudo-records for the missing trajectory. Following the interpolation step, the data sets were again assessed and split on points where the gap exceeded five minutes. The Interpolation data set was selected at this point, and person-days were created within each complete set, and marked for use if the total coverage was at least 1368 minutes. The data set without further selection was marked for use as the basis of the DTWBMI imputation.

Stop detection and segmentation

Stop detection was carried out on each of the four data sets (Listwise, Pairwise, Interpolation, DTWBMI) using a spatiotemporal clustering algorithm (Montoliu, Blom, and Gatica- Perez 2013) as implemented by the stopdetection package in R. Radius parameters were set relatively short (100 m), as were the time parameters (3 minutes), in order to better discriminate between stops and short tracks, which was a focus of this paper. All points were subjected to a merging step that excluded short tracks between stops, and merged stops whose centers were within 100 meters of each other. This was done in order to reduce the impact of false-positive tracks.

Following stop detection, tracks and stops were segmented independently of each other using the R package topdowntimeratio which implements the time-sensitive segmentation algorithm from (Meratnia & de By, 2004). An error parameter of 40 meters was selected for tracks identified in the data, and an error parameter of 200 meters for stops. This step reduces the noise in the data due to measurement error prior to calculating distance estimates. Alternatives such as Kalman filters and mean/median smoothing were considered, but segmentation followed by summing across the segment Haversine distances was judged to offer superior results.

Aggregation

To facilitate comparison between spatially heterogeneous trajectories for imputation purposes, the data in the DTWBMI set were aggregated to 15-minute intervals, creating the following summative variables: total distance, travel distance, number of tracks, number of short tracks, and an additional variable for the id of short tracks contained in the 15-minute interval. Total distance for each interval was calculated on the basis of the TDTR adjusted latitudes and longitudes. Travel distance was calculated in the same way, but restricted to points identified as belonging to tracks. Track variables were less straightforward due to the lack of temporal alignment with track and stop designation. Where the 15-minute interval contained only points identified as either "stop", the number of tracks was set to zero. If the 15-minute interval contained one or more tracks, this was summed as well, but required special processing during the analysis stage to establish contiguous tracks versus separate tracks. For short tracks, where it was very likely that multiple would occur within a single 15-minute interval, the short track ID was preserved to be carried through to the imputation step, providing a manner of tracking individual short tracks. The result was a long-form data set, grouped by user-sets that could be imputed and joined, where each row represented travel behaviors occurring during a 15-minute interval.

Part V

Backmatter

E

Nederlandse samenvatting

Dit proefschrift gaat over wat er misgaat én wat er nog steeds kan als we reisgedrag meten met smartphone-apps in plaats van met traditionele dagboeken. Steeds meer onderzoekers en statistiekbureaus willen apps inzetten om de mobiliteit van mensen automatisch te registreren. Smartphones kunnen heel precies en bijna continu locaties vastleggen, en dat levert in theorie rijkere data op dan een papieren of online dagboek waarin mensen zelf hun reizen moeten onthouden en opschrijven.

In de praktijk blijken deze Smart Surveys echter allesbehalve probleemloos. De belangrijkste spelbreker is ontbrekende data: gaten in de GPS-metingen, dagen waarop de app niets registreert, deelnemers die halverwege afhaken, en systematische verschillen tussen typen telefoons. Als we die ontbrekende data negeren of op een naïeve manier opvullen, krijgen we vertekende schattingen van hoe ver mensen reizen, hoeveel trips ze maken, en hoe mobiliteit verschilt tussen groepen in de bevolking.

Dit proefschrift richt zich op één centrale vraag: hoe kunnen we op een verantwoorde manier omgaan met ontbrekende data in app-gebaseerde reisdagboeken, zodat de uitkomsten bruikbaar worden voor beleid en officiële statistiek? De nadruk ligt daarbij op smartphone-gebaseerde reisonderzoeken (Smartphone-Based Travel Surveys, oftewel SBTS), en op methoden om verschillende soorten gaten in de data te classificeren en te imputeren zonder de resultaten onherstelbaar te vertekenen.

Het proefschrift bestaat uit een inleidend hoofdstuk, drie methodologische kernhoofdstukken, en een afsluitend hoofdstuk waarin de belangrijkste inzichten samenkomen. In het eerste hoofdstuk wordt de context geschetst. Traditionele reisdagboekonderzoeken (Travel Diary Studies, TDS) kampen al jaren met onderrapportage en nonrespons. Deelnemers vergeten trips, hebben weinig zin om alles bij te houden of haken voortijdig af. Tegelijkertijd verwachten onderzoekers veel van smartphone-apps: automatische locatiemeting, minder belasting voor respondenten, meer detail, en mogelijk een betere dekking van korte en routinematige verplaatsingen.

Toch schuift met de overstap naar apps het probleem vooral van vragen naar algoritmen. In plaats van ”Hoeveel kilometer heeft u gereisd?” krijgen we miljoenen ruwe locatiemetingen die eerst vertaald moeten worden naar stops, verplaatsingen en afstanden. Daarbij is het idee van ”complete data” fundamenteel anders: sensoren leveren zelden een perfect aaneengesloten tijdlijn. Er zijn altijd hiaten: batterijoptimalisatie kan de app afsluiten, de locatiemetingen kunnen soms slecht ontvangen worden, of gebruikers kunnen de instellingen aanpassen. Hoofdstuk 1 maakt duidelijk dat ontbrekende data in smart surveys geen randprobleem is, maar de kern vormt van de methodologische uitdaging. De rest van het proefschrift bouwt voort op die observatie.

Hoofdstuk 2 beschrijft de ontwikkeling en eerste grootschalige inzet van de Statistics Netherlands Travel App, een app die is ontwikkeld door het Centraal Bureau voor de Statistiek (CBS) om reisgedrag in een landelijk representatieve steekproef te meten. Het hoofdstuk beschrijft de technische en praktische aspecten van de app en de bijbehorende back-end, en geeft een evaluatie van de datakwaliteit, zowel wat betreft representativiteit als meetkwaliteit.

Een steekproef van 1.902 personen werd uitgenodigd om een week lang hun mobiliteit via de app te laten registreren. Van de 674 deelnemers die enige vorm van data leverden, was de gemiddelde dekking slechts 8,2 uur per dag, oftewel ongeveer een derde van wat wenselijk is. Slechts 5 personen leverden zeven volledig gedekte dagen zonder grote gaten.

De data bevatten veel ontbrekende waarden: niet alleen korte gaps, maar ook lange periodes zonder metingen en complete dagen zonder data. Deze gaps bleken systematische patronen te vertonen: de datacompleetheid hing samen met persoonskenmerken zoals leeftijd én met apparaatkenmerken zoals besturingssysteem en merk. Dat betekent dat de gaten in de data niet willekeurig zijn, wat een bedreiging vormt voor de representativiteit. Hoofdstuk 2 geeft ook een eerste vergelijking met ODiN, de Nederlandse reisdagboekstudie: de app registreerde meer trips, vooral korte ritten, maar gaf onwaarschijnlijke verschillen in totale reisafstand ten opzichte van het traditionele dagboek. Dit wijst op een complex samenspel van meetfouten en ontbrekende data.

Hoofdstuk 2 toont daarmee dat smartphone-apps een enorm potentieel hebben, maar dat het probleem van ontbrekende data eerst goed begrepen en aangepakt moet worden voordat de data bruikbaar zijn voor officiële statistiek.

In Hoofdstuk 3 stellen wij de vraag hoe lang een gat mag zijn om nog behandeld te kunnen worden met interpolatie. In veel onderzoeken wordt ontbrekende GPS-data in feite genegeerd: elk gat wordt behandeld alsof iemand een rechte lijn heeft afgelegd tussen de laatste locatie voor het gat en de eerstvolgende locatie erna. Dit kan prima werken voor korte hiaten, maar wordt riskant als de gaten langer worden of op drukke momenten van de dag vallen. In dit hoofdstuk onderzoeken wij wanneer dit soort interpolatie een acceptabel niveau van meetfout introduceert, en vanaf welk punt de bias in belangrijke metingen van mobiliteitsgedrag (zoals totale afstand, aantal verplaatsingen, of radius of gyration) te groot wordt.

Met behulp van volledig geobserveerde data uit de CBS Travel App wordt een reeks simulaties uitgevoerd. In deze simulaties worden kunstmatig hiaten van verschillende lengtes en op verschillende tijdstippen aangebracht, vervolgens met lineaire interpolatie opgevuld, waarna wordt gekeken hoeveel de uitkomsten afwijken van de werkelijke waarden.

Uit deze simulaties blijkt dat de totale hoeveelheid ontbrekende data minder belangrijk is dan de lengte en het tijdstip van de gaten. Een gat van 30 minuten midden in de nacht kan vaak probleemloos worden geïnterpoleerd; hetzelfde gat tijdens de ochtendspits kan een volledige woon-werkrit maskeren. Ook blijkt er een praktische drempel te zijn: gaten korter dan ongeveer 10 minuten kunnen in veel situaties met eenvoudige interpolatie worden opgevuld, met een maximale bias van ongeveer 5% in kernstatistieken. Bij langere gaten neemt de bias niet lineair maar sterk versneld toe, vooral voor reisafstand.

Hoofdstuk 3 levert zo concrete richtlijnen op voor onderzoekers: korte gaten kunnen relatief veilig met eenvoudige methoden worden gevuld, maar langere gaten vragen om meer geavanceerde technieken. Deze drempels vormen de basis voor het onderscheid tussen korte en lange gaps in de rest van het proefschrift.

Korte gaten zijn dus nog beheersbaar met lineaire interpolatie; langere gaten niet. In Hoofdstuk 4 wordt daarom een nieuwe methode ontwikkeld voor het imputeren van lange gaten (tot ongeveer 10–12 uur) in mobiliteitstrajecten: Dynamic Time Warping-Based Multiple Imputation (DTWBMI).

Het idee achter de methode is intuïtief. Als je iemands bewegingen langer volgt, ga je patronen zien: dezelfde routes naar werk, vaste boodschappenrondjes, terugkerende bezoeken aan bepaalde locaties. Die patronen kunnen worden gebruikt om ontbrekende stukken in een dag te reconstrueren. Technisch gebeurt dat als volgt: rondom een lang gat worden de waarnemingen direct vóór en na het gat gebruikt als buffer. Met Dynamic Time Warping (DTW) worden in de data andere dagen of trajecten gezocht die qua patroon lijken op de geobserveerde buffers. Uit deze best passende referentietrajecten wordt de ontbrekende periode gehaald en gebruikt als donordata. Dit proces wordt herhaald om meerdere plausibele invullingen te genereren, zodat de onzekerheid rond de imputatie correct meegenomen kan worden.

Een belangrijk kenmerk is dat DTWBMI geen externe gegevensbronnen of kaartkoppeling nodig heeft; de methode werkt uitsluitend met tijdgestempelde locaties uit dezelfde studie. Uit uitgebreide simulaties blijkt dat vooral een low-information variant van DTWBMI verrassend goed presteert: met relatief beperkte input worden ook bij lange gaten de totale reisafstand en andere mobiliteitskenmerken met kleine bias gereconstrueerd. Terwijl lineaire interpolatie bij gaps van 10 uur al snel meer dan 10 kilometer aan afstand verkeerd inschat, blijft de gemiddelde absolute bias van DTWBMI rond de 0,6 kilometer.

Hoofdstuk 4 levert daarmee een nieuwe, praktisch toepasbare imputatiemethode op voor lange gaps in smartphone-gebaseerde mobiliteitsdata.

Hoofdstuk 5 presenteert een hiërarchisch framework voor ontbrekende data. Alle eerdere inzichten worden samengebracht in een driestappenframework voor ontbrekende data in app-gebaseerde reisdagboeken. In plaats van één uniforme methode voor alle soorten missingness te gebruiken, wordt onderscheid gemaakt tussen drie niveaus: korte gaps (tot 30 minuten), lange gaps (30 minuten tot 12 uur), en ontbrekende data op dagniveau (meer dan 12 uur, of volledige dagen zonder bruikbare data). Korte gaps worden behandeld met lineaire interpolatie op basis van de richtlijnen uit Hoofdstuk 3, met als doel het ruimtelijke pad te reconstrueren waar dat verantwoord kan, zonder onnodige bias te introduceren. Bij lange gaps wordt DTWBMI toegepast om op basis van temporele gedragspatronen de mobiliteit tijdens het gat te imputeren. Hier ligt de focus op het correct schatten van mobiliteitsgedrag over de tijd (afstand, aantal trips) in plaats van exacte routes. Bij volledig ontbrekende dagen is er geen bruikbare trajectinformatie meer en verschuift de aandacht naar het schatten van statistieken op populatieniveau. Met Multiple Imputation by Chained Equations (MICE) worden statistieken op dagniveau geïmputeerd op basis van demografische en contextuele variabelen en de patronen van respondenten die wél data hebben. Op deze manier wordt selectieve nonrespons gecorrigeerd, bijvoorbeeld wanneer bepaalde groepen (zoals ouderen of mensen die weinig reizen) systematisch minder geneigd zijn om de app te gebruiken.

Het framework is toegepast op de data uit de CBS Travel App veldtest van 2018. Zonder imputatie en bij strikte eisen aan compleetheid zouden slechts 5 personen overblijven met een volledige week, wat onvoldoende is voor zinvolle conclusies. Door de hiërarchische imputatie kunnen veel meer deelnemers bijdragen aan de schatting van weekstatistieken, dalen de betrouwbaarheidsintervallen aanzienlijk, en komen de geschatte reisafstanden en aantallen trips dichter in de buurt van die uit ODiN, met verschillen die inhoudelijk verklaarbaar zijn.

Hoofdstuk 5 laat zien dat een doordacht imputatieproces een dataset die op het eerste gezicht onbruikbaar lijkt, kan omvormen tot een bruikbare basis voor populatieschattingen van mobiliteit. De prijs die daarvoor wordt betaald is dat de zeer gedetailleerde ruimtelijke informatie gedeeltelijk verloren gaat in de laatste imputatiestappen, maar daar staat een betere schatting van reisgedragstatistieken tegenover.

Hoofdstuk 6 is het afsluitende hoofdstuk waarin alle bevindingen worden samengevat en vertaald naar praktische aanbevelingen voor verschillende groepen gebruikers.

Voor ontwerpers van app-gebaseerde travel surveys betekent dit onder meer dat app en analysemethode in nauwe samenwerking ontwikkeld moeten worden; algoritmen voor dataverwerking zijn geen bijzaak. Verder is het van belang de app te ontwerpen met het oog op bekende patronen van missingness, en kwaliteitsmonitoring in te bouwen om problemen vroegtijdig te signaleren. Ook kan het helpen om passieve metingen te combineren met gerichte actieve vragen aan respondenten, om gaten te identificeren en beter te begrijpen.

Voor data-analisten geldt dat verschillende methoden nodig zijn voor verschillende typen ontbrekende data: korte gaps kunnen anders worden behandeld dan lange gaps of ontbrekende dagen. Naïeve strategieën zoals lijstgewijze deletie of blind interpoleren over lange gaten leveren instabiele en sterk vertekende schattingen op en dienen vermeden te worden. Multiple imputation (op traject- en dagniveau) biedt de mogelijkheid om onzekerheid en selectiviteit expliciet te modelleren.

Voor nationale statistiekbureaus en beleidsmakers ten slotte is het belangrijk smartphone-gebaseerde surveys niet te zien als directe vervangers van traditionele methoden, maar als een aanvullende bron met eigen sterktes en zwaktes. Bij het steekproefontwerp moet rekening gehouden worden met hoge en systematische missingness; grotere steekproeven en extra methodologisch ontwikkelwerk zijn nodig. Nieuwe technologie verdient tijd: de eerste iteraties dienen vooral om methodologische lessen te trekken en standaarden te ontwikkelen voor datakwaliteit en imputatie.

Samenvattend laat dit proefschrift zien dat ontbrekende data weliswaar een groot probleem vormen in smartphone- gebaseerde reisonderzoeken, maar dat dit probleem met de juiste aanpak te beheersen is. Door verschillende soorten gaten te onderscheiden en voor elk type een geschikte imputatiemethode te kiezen, ontstaat een werkwijze die onderzoekers in de praktijk kunnen toepassen.

F

List of scientific publications

Publications

Akdeniz, D., Schmidt, M. K., Seynaeve, C. M., McCool, D., Giardiello, D., van den Broek, A. J., Hauptmann, M., Steyerberg, E. W., & Hooning, M. J. (2019). Risk factors for metachronous contralateral breast cancer: A systematic review and meta-analysis. *Breast*, *44*, 1–14. https://doi.org/10.1016/j.breast.2018.11.005

Bucher, H., Keusch, F., De vitiis, C., de Fausti, F., Inglese, F., van Tienoven, T. P., McCool, D., Lugtig, P., & Struminskaya, B. (2023). *Smart survey implementation workpackage 2: Research methodology deliverable M6: Review stage* (research rep.). European Commission. https://doi.org/10.13140/RG.2.2.20367.09126

Bucher, H., Keusch, F., Volk, J., Häufglöckner, L., Blanke, K., De Fausti, F., Inglese, F., Terribili, M., Perez, M., van Tienoven, T. P., Mccool, D., Lugtig, P., Struminskaya, B., Elevelt, A., De Groot, J., Kompier, M., Schouten, B., Klingwort, J., Van Den Heuvel, J., … Al Ibraheem, A. (2024). *Smart survey implementation - work package 2: Research methodology - deliverable M14: Smart baseline stage* (research rep.). European Commission. https://doi.org/10.13140/RG.2.2.15333.92640

Drooger, J., Akdeniz, D., Pignol, J.-P., Koppert, L. B., McCool, D., Seynaeve, C. M., Hooning, M. J., & Jager, A. (2015). Adjuvant radiotherapy for primary breast cancer in BRCA1 and BRCA2 mutation carriers and risk of contralateral breast cancer with special attention to patients irradiated at younger age. *Breast cancer research and treatment*, *154*(1), 171–180. https://doi.org/10.1007/s10549-015-3597-7

Fritz, M., Keusch, F., Volk, J., Häufglöckner, L., Blanke, K., De Vitiis, C., D'Amen, B., De Fausti, F., Inglese, F., Lorè, B. M., Pappagallo, A., Piccolo, F., Terribili, M., Perez, M., Van Tienoven, T. P., Lusyne, P., McCool, D., Lugtig, P., Struminskaya, B., … Holmøy, A. (2025, April 25). *Deliverable 2.3 smart advanced stage* (research rep.). ESSnet. Brussels, Belgium, European Commission.

Kramer, I., Schaapveld, M., Oldenburg, H. S. A., Sonke, G. S., McCool, D., van Leeuwen, F. E., Van de Vijver, K. K., Russell, N. S., Linn, S. C., Siesling, S., et al. (2019). The influence of adjuvant systemic regimens on contralateral breast cancer risk and receptor subtype. *JNCI: Journal of the National Cancer Institute*, *111*(7), 709–718. https://academic.oup.com/jnci/article-abstract/111/7/709/5304371

McCool, D., Lugtig, P., Mussmann, O., & Schouten, B. (2021). An app-assisted travel survey in official statistics: Possibilities and challenges. *Journal of official statistics*, *37*(1), 149–170. https://doi.org/10.2478/jos-2021-0007

McCool, D., Lugtig, P., & Schouten, B. (2022). Maximum interpolable gap length in missing smartphone-based gps mobility data. *Transportation*, 1–31.

van Rossum, A. G. J., Kok, M., McCool, D., Opdam, M., Miltenburg, N. C., Mandjes, I. A. M., van Leeuwen-Stok, E., Imholz, A. L. T., Portielje, J. E. A., Bos, M. M. E. M., van Bochove, A., van Werkhoven, E., Schmidt, M. K., Oosterkamp, H. M., & Linn, S. C. (2017). Independent replication of polymorphisms predicting toxicity in breast cancer patients randomized between dose-dense and docetaxel-containing adjuvant chemotherapy. *Oncotarget*, *8*(69), 113531–113542. https://doi.org/10.18632/oncotarget.22697

Conference papers

McCool, D., Schouten, B., & Lugtig, P. (2022). Dynamic time warping-based imputation for long gaps in gps mobility research. *European Transport Conference 2022 Association for European Transport (AET)*.

McCool, D., Schouten, B., Lugtig, P., Ròth, K., & Smeets, L. (2019). Smartphones and Always-On location data: Assessing the impact of a transition to mobile devices. *European Transport Conference 2019 Association for European Transport (AET)*. https://trid.trb.org/view/1730371

8

Selected presentations

McCool, D. *Assessing and addressing missingness mechanisms in Passively-Recorded location data*. European Survey Research Conference. Zagreb, Croatia, 2019, 16 7. https://www.europeansurveyresearch.org/conf2019/uploads/619/3076/24/ESRA_McCool_Assessing_Missingness_Mechanisms_in_Always_On_Location_Data_upload.pptx

McCool, D. *Improving transportation research with passively-collected location data*. Mobile Apps and Sensors in Surveys. Mannheim, Germany, 2019, April. https://daniellemc.cool/Improving_Transportation_Research

McCool, D. *Gaps small and large in app-based location tracking*. CBS Big Data Seminar. Online, 2020, December.

McCool, D. *Mind the gap - the missing data problem in longitudinal location data*. BigSurv20. Online, 2020, 13 11. https://www.youtube.com/watch?v=03D1Tf1Ve_s

McCool, D. *The tabi travel app – construction, results and commentary*. 2021, March 9.

McCool, D. *Multiple imputation with (annoyingly) spatiotemporal data*. Nederlandstalig Platform voor Survey Onderzoek (NPSO Innovatiedag). Den Haag, 2021, December.

McCool, D. *Dynamic time warping based imputation for imputing long gaps in mobility data*. European Transport Conference, Association for European Transport (AET). Milan, Italy, 2022, September.

McCool, D. *Dynamic time warping based multiple imputation for large gaps in time series*. Mobile Apps and Sensors in Surveys (MASS). Utrecht, Netherlands, 2022, June.

McCool, D. *Dynamic time warping based multiple imputation of long gaps in human trajectories*. Mobile Apps and Sensors in Surveys (MASS). Manchester, United Kingdom, 2023, June.

McCool, D. *Two methods for integrating smart surveys with traditional surveys*. 2025, March 12.

McCool, D., Lugtig, P., & Schouten, B. *Imputing missing mobility data in the 2018 smartphone travel diary study*. European Survey Research Association Conference 2023. Milan, Italy, 2023, July.

G

Curriculum Vitæ

Danielle McCool

Education

2012–2014	Msc Statistics and Methodology Utrecht University, Utrecht *Thesis:* Recurrent event model for population size estimation *Supervisor:* Marten Cruijf
2005–2008	Bsc Psychology/Philosophy Texas Woman's University, Denton, USA

Research and Industry

2025– present	Country Team Operator Survey of Health Ageing and Retirement in Europe
2024–2025	Postdoctoral Researcher Digital Data Donation Infrastructure
2023–2025	Postdoctoral Researcher ESSnet Project Smart Survey Implementation
2023–2023	Methodologist Centraal Bureau voor de Statistiek
2018–2023	PhD Candidate Methodology and Statistics, Utrecht University

Acknowledgements

I think as a statistician, maybe I'm not supposed to believe in luck, but what else would you call the one person in 726,818 who (by pure chance) flips heads 20 times in a row? I'm lucky to have had so many people in my life that have taken a chance on me, without whom this thesis would never have existed.

First and foremost, when it comes to people who took a chance on me, I will be forever grateful to Peter Lugtig and Barry Schouten.

It's impossible to imagine two better thesis supervisors than you two. You gave me the freedom to take the original idea of the PhD and run with it in whichever way I wanted. There were days I worried that it had gotten me into a bit of a mess, but no matter what, no matter how obscure my proposed solution, you were always supportive, you were always kind, and you were always helpful. I am grateful that you two gave me the chance in the beginning.

Peter, you're my goalpost for what it means to be a good professor. I ask myself all the time, "What would Peter do?" and usually the answer is something like 1. assume the best of everyone involved, 2. be humble and open to others' opinions, and 3. try to understand what really matters. Thank you for being you.

Barry, your capacity for finding ways for other people to shine is really something. You ask questions to which you already know the answer so that people get to feel clever in addressing them. It feels like you're always 10 steps ahead when it comes to truly understanding statistics, but you still let people find their own way. Re-reading old emails about the direction of my project is like becoming aware that I was being mentored by a prescient but benevolent god who knew I could only see the problems by getting there myself.

I would like to express my sincere thanks to my reading committee, Stef van Buuren, Dick Ettema, Ellen Hamaker, Florian Keusch, and Ton de Waal. I appreciate the time and effort invested in reviewing this work.

I'm grateful to so many people who took a chance on me at critical moments. Thanks Irene Klugkist; I've never forgotten that you were the deciding vote to let me into the program. If you hadn't looked at me and thought I could take it on, I wouldn't be here today. My life started in 2012. Thanks Maarten for taking a chance on me in supervising my Masters thesis what now feels like a lifetime ago. Thanks Vera Toepoel for letting me try my hand at writing. Thanks Dave Hessen for letting me attempt teaching for the first time. Thanks Peter van der Heijden for letting me try my first real consultation. Thanks Bella Struminskaya for giving me my first shot at undertaking real life (expensive) survey tasks. And thanks Laura Boeschoten for taking a chance on me now as a developer.

This thesis depended on a truly huge community of individuals who provided feedback, support, ideas, direction, and sometimes just pure pressure to complete it. I feel incredibly lucky to have been able to work alongside all you talented people.

For example, the community at (and around) Statistics Netherlands who developed, maintained, supported, or helped me analyze the smart surveys I worked on, including Victor Verstappen, Ole Mussmann, Jonas Klingwort, Yvonne Gootzen, Tim Schijvenaars, Tom Oerlemans, and Mike Vollebregt. Thanks also to the Primaire Waarneming group in Heerlen for allowing me to join in for a few months! I loved being there for the Keek op de Week with Deirdre Giesen, Jeldrik Bakker, Janelle van den Heuvel, Maaike Kompier, Vivian Meertens, Ger Snijkers, Lianne Tessers-Ippel, Marko Roos, and Jo Brouwers. And naturally, thanks to Vera for making my time with the group possible.

To those at Utrecht University, you're what make the place great. I hope we never lose the sense of togetherness that our department has.

Without a doubt, I owe more thanks than I can possibly write here to the unwavering support of the Data Quality group: Peter, Bella, Katharina, Angelo, Daniëlle, Camilla, Thijs, Laura, Anne, Niek, Charlotte, and Robin. Honestly, I kind of think of you like my work family except we get along much better and I completely trust each and every one of you. Coordinating ESRA with you all was a trip, and I hope I get to keep up our collaborations for a very long time.

Also a specific shout-out to the Missing Data group, including Stef, Gerko, Hanne, and Thom for all your help while I was preparing Chapter 4. The opportunity to bounce ideas off of you all while I was putting together what it would look like to use dynamic time warping for donor selection was totally invaluable.

And to ever-growing list of roommates Anne, Pia, Daan, Jeroen, Melanie, Angelo, Camilla, Anne, David, and Daniëlle for putting up with my nomadic lifestyle. As I migrated further down the C hallway 1.21 -> 1.19 -> 1.24, I felt more comfortable leaving my things in a wider array of cabinets, but you all were good sports about it. I really value our time sharing space together and I felt like you all really knew and accepted me. You got me through heavy times like when I wasn't sure how I was going to drag this beast across the threshold. Thank you, from the bottom of my sometimes-too-talkative heart. Thanks Pia and Melanie for agreeing to be my paranymphs. Although I have wisely selected two people who already know I'm a sentimental crier, I'd like to offer you two my apologies in advance.

A big, hearty thanks to the Thursday borrel crew. It ended up being a good platform for us all to talk about the impacts of heavy events, from elections to budget cut impacts. In no particular order: Thom, Pia, Özgür, the Camillas, Elena, Thijs, Edita, Alfons, Manuel, Timo, Florian, Jeroen, Pablo, Hanne, Pepijn, Ingrid, and Angelo.

To my other colleagues past and present in the department: thanks for talking shop with me. Trolling the coffee machine for people to bounce ideas off of was always successful, and I enjoyed being nosy about your projects during lunch. So thanks Qixiang, Hidde, Kyle, Boukje, Beth, Ayoub, Kirsten, Daniel, Noémi, Karlijn, Erik-Jan, Kevin, Oisín, Emmeke, Caspar, and Mahdi.

To my M&S Masters cohort: what a crazy chance we took together. I think it paid off exceptionally well. How often in life does one get such a fantastic opportunity to be surrounded by a bunch of smart, interesting people, all of whom are working towards the same set of goals, and who are 100% invested in wringing every last knowledge droplet out of our course load. We pulled each other up, made each other more powerful, and at the risk of sounding like a huge Boomer, I genuinely think that education can't get any better than the way we did it during the MSc. Whether we were working on Fundamentals around the Langeveld lunch table, SPSS syntax at the flexplekken, or R in the computer lab, we were always pulling each other upwards and onwards. Eva and Mariëlle, thank you for ensuring no one ever missed a deadline, Pascal thank you for being the powerhouse always pushing for different solutions, thanks Timo for the heads up on my American dance style, thanks Rob for being a good sport with my stupid joke to include your name as the Web survey consultant for McBurgers International, Sanne a ton of thanks for being a picture-perfect example of a great department Buddy (in the 7 or 8 times I've done it since, I've yet to live up to it, but I'm getting closer), thanks Sint Bente for the theedoosje, and for everyone for not laughing at my first ever attempt at a Sinterklaas gedicht. Laura, thanks for sharing job ads with me when we were both on the hunt for a PhD. I'm glad I didn't unsubscribe from the updates so another one could pass my way 11 years later.

Kirsten, Jolien, and Kees: for two years of my life, I think I spent 75% of my waking life in the company of at least one of you three. It was fun, and I miss it. Niek, Pascal, Joost, Teresa, working with you all on the Lobos board taught me a ton of things that have been unexpectedly helpful, like ensuring that meetings have agendas. I miss you all, too. To all of you in Lobos and YLL with whom I drank many beers over many a gezellig evening: Games at my place soon?

Matti and Eline, you have been fantastic friends, and I'm glad I've gotten to see your family grow. Even when it's a long time between us getting together, it somehow always feels like it was just yesterday. Thank you for all your support.

Nick and Mitchell, thank you for always being there to listen to me rant at length about things, and for looking at all the stupid papers I send to you. Thanks for when you let me explain statistics things to you because it's fun for me, and thanks for answering my Nth question of the month on git when Shawn is otherwise indisposed. Nick, I can't believe I have to come back in and add this part, but thanks for making my cover better. I don't think I'll ever forget that you turned a 5 minute "what do you think?" into a two-hour "Don't worry, I'll fix it" session. Thanks Maaike for the back-channel help there, and thanks Mitchell for whatever that one suggestion was, I'm sure it was great. Nick, I hope we never have to stop looking at Funda together, and Mitchell, I hope we get to stop looking at Funda together soon. You two are closer than friends, you're family, and I am so happy that I get to know and love you.

Thanks Joyce - Grandma 'Cool'. I love you and am so lucky to have a mother-in-law like you in my life. Thanks for everything that you do for us. Mom, I'm sorry you didn't make it to see me finish this whole thing up. I know you were proud of me.

Lastly, to my immediate family, thank you for bearing with me. Thank you Elly, your tenacity inspires me. I love that you go after what you want and make it real, even if it's a bit crazy like learning French and moving to Paris to study cooking. Thank you for holding up to my busy periods over these last six years. Serena, my firstborn, there's not a lot of kids who've been along for the entirety of their parents' academic journey from Bachelors to Doctorate, but we made it, you and I. Your empathetic, creative, lovely, warm self is going to make the world a better place. Thank you for being my kid. Shawn, to you I owe the biggest debt of gratitude of all. My partner, my soulmate, my best friend. Let's go on a vacation.