Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 52 (2015) 600 – 605

Procedia
Computer Science

The 6th International Conference on Ambient Systems, Networks and Technologies (ANT 2015)

Diary Survey Quality Assessment Using GPS Traces

Ali Raza[a], Luk Knapen[a], Katrien Declercq[a], Tom Bellemans[a], Davy Janssens[a], Geert Wets[a]

[a]School of Transportation Sciences, Hasselt University, Wetenschapspark 5 bus 6 3590 Diepenbeek, Belgium

Abstract

Surveying diaries to collect evidence about travel behavior is known to be susceptible to erroneous reporting. Many respondents do not exactly remember details about executed trips and activities. This paper describes a diary collection project aimed at tracking travel behavior in an electric vehicle pilot project. GPS traces corresponding to the collected diaries have been registered. Dedicated interactive software was used to align reported activity and trip timing with the GPS recordings. The difference between each original diary and the corresponding corrected version, has been quantified. The number of modifications as well as the distribution of their magnitudes have been analyzed. Time needed for correction has been recorded. Those results can serve to plan future data collection efforts and lead to specific recommendations to avoid error and data cleaning time.

Peer-review under responsibility of the Conference Program Chairs
Keywords: Prompted recall technique; Stop Detection; Schedule Alignment; Data Analysis on Corrections

1. Introduction - Context

Electric vehicles are considered as an answer to the ever increasing adverse effects of fuel emissions. The impact of the use of electric vehicles on the travel behavior of individuals was studied in the DATA SIM project. For this purpose activity travel diaries were filled out by the persons participating in the project. The data collected via travel diaries for the DATA SIM project is referred to as the iMove data. The participants were provided with an electric vehicle. Travel diaries were collected with the help of a smartphone application known as SPARROW. The respondents filled in the details of their activities and trips in the application. GPS traces of the movements were also recorded by this application. The consistency of the travel diaries was checked with the help of GPS traces The errors in the data were then removed with the help of schedule alignment software. After the removal of errors, the set of applied corrections were analyzed. With the help of this analysis, various problems in the active data collection techniques could be highlighted.

* Corresponding author. Tel.: +32-11-269-111 ; fax: +32-11-269-199.
E-mail address: syed.raza@student.uhasselt.be

Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2015.05.045

2. Data Collection Process

2.1. SPARROW application

Data collection took place through a smartphone application designed specifically for this purpose known as SPARROW. With the help of this application, the users could fill in the details of their activities and trips. The application could also record the GPS tracks of the individuals.

2.2. Data Entry and description of errors in the diaries

The users could enter the schedules/diaries as per their convenience. The time at which data were entered was recorded. This helped in calculating the time lapse between data entry time and the stated activity/trip start time.

Due to this degree of freedom, users filled the diaries in advance as well as a number of days after performing an activity/trip. This led to many errors because the entered data by the users differed from the GPS tracks recorded by the application. The reason behind this could be not correcting the diaries after changing their plans. Some of the errors included missing out on trips/activities because of not remembering. This probably in some instances led to the entry of fake data as well. In some cases respondents did not enter the missing trips. In the application, respondents could enter more than one activity/trip in the same time slot. This overlap and holes were displayed to the users with a red outline around the time and trips/activities box. Since the correction of these overlaps was not obligatory, some of the respondents did not correct the errors. Lack of education regarding the use of the application also played a role in the huge amount of errors in the data. Another type of error found in the travel diaries included the missing GPS points. The reason for this could be switching off the application or battery drainage.

3. Data Cleaning Tool and Process

3.1. Qgis Schedule Alignment tool

A plugin for the Quantum GIS software was used to correct the diaries. It can display the GPS points on an underlying map. The plugin is capable of showing both the original diary filled in by the users as well as the upgraded version. The coordinates on the map show the location of the user at a particular point in time. Different activity types and transport modes were differentiated from each other by different colors. The GPS points of trips spread out on the road networks since the user was on the move while the GPS points of activities occurred in the form of clusters since the user did not move while performing an activity. Deletion or addition of an activity/trip along with the adjustments in start/end times of the activities/trips could be done in the software.

3.2. Data Correction

The corrections were made by comparing GPS tracks with the entered data. In order to define the correction process an examples is discussed below. Figure 1(a) shows the data as entered by the participant along with the recorded GPS traces. The respondent claimed that he made an electric-car trip, performed a non daily shopping activity, again made an electric-car trip, performed a social visit activity and immediately after that started a home activity. GPS points show that the user has forgotten to report an electric-car trip (black) between social visit and home activity since green points are shown in the form of clusters as well as movements which means the user was performing a home activity while he was moving. The user has also specified wrong start/end times for home (green), non-daily shopping (yellow) and social visit (purple) activities. Figure 1(b) shows the corrected version of the diary.

3.3. Time recording results for the interactive work

The number of corrections was counted automatically using a software that compares the original and corrected diaries. The total number of corrections, additions and deletions for trips amount up to 3837, 1105 and 1557 respectively. Correcting the diaries for 33 persons took 70 hours and 15 minutes. The average time required to correct a single diary was 2 hours and 22 minutes.

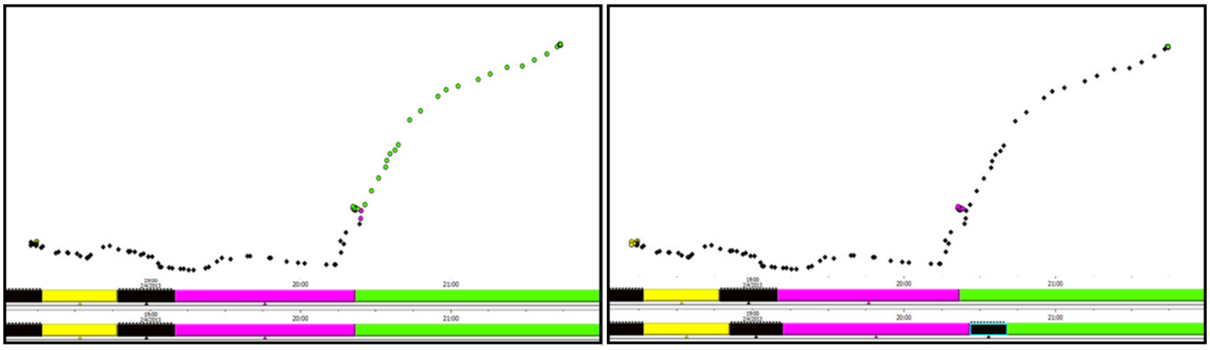

Fig. 1. (a) Diary (b) Corrected diary (The width of activities corresponds to the duration)

4. Quantitative Analysis of Data Cleaning

4.1. Extraction of Evidence - Quantifying Differences

The data recorded using the SPARROW smartphone app were collected in a relational database and exported via an XML format to the *Qgis Schedule Alignment* tool for interactive adaptation to the corresponding recorded GPS trace. The adapted data became available in the same XML format which allows for automatic comparison. The combination of a trip towards a location and the activity conducted at that location, is called an episode. Each episode in a diary got a unique identifier when it was first recorded by the SPARROW smartphone app. Newly added episodes entered during the cleaning process received a unique identifier too. The XML representations raw and cleaned diaries have been compared using the standard unix `diff` tool. Post-processing the `diff` output lead to episode *addition*, *deletion* end *modification* records in a relational database, ready for statistical analysis. Each such record specifies the original and adapted (if applicable) values for travel mode, activity type and timing.

4.2. Data Analysis Process

For all persons for whom the activity and trip data were corrected, files were created with data before and after correction.

For activities/trips, differences between corrected and original data were calculated for start-time, end-time and duration. Additions, modifications and deletions were distinguished and linked to socio-demographic variables (gender, age-class, profession and diploma).

4.3. Analysis for Activities and trips

4.3.1. Number of modifications, additions and deletions per person per day

For activities, the number of corrections per type (modification, addition, deletion) according to gender, age-class, profession, diploma and batch was analyzed first. For each person per day, there were 3 to 4 corrections, 1 addition and 2 to 3 deletions. The corrections were 3.4 for men and 4.3 for women, and almost the same number of additions (0.8 and 0.7 respectively) and deletions (2.1 and 2.5 respectively). The number of corrections are higher in the age-group 40-49 compared to the 30-39 group (4.0 and 3.0 respectively). Deletions are 2.9 for blue-collar and 1.8 for white-collar. With regards to diploma, there seem to be less corrections for those with a university diploma (2.8) as compared to 4.0 and 3.6 for those with a secondary school and high school diploma. The number of additions is roughly the same and the number of deletions is 2.7 for secondary school and university and 1.5 for high school.

For trips, the number of corrections is 3 on average, the number of additions is slightly above 1 and the number of deletions is 0.4. There is no apparent difference between men and women. The corrections are 2.5 for age 30-39 and 3.4 for age 40-49. Similarly the number of additions is 0.8 for 30-39 and 1.6 for 40-49. For the number of deletions there is no difference.

	Min	Q1	Median	Q3	Max	-10 to 10 minutes
Start-time	-765(<-12h)	-2	0	1	193	84%
End-time	-382(<-6h)	0	1	8.5	3469(>57h)	69%
Duration	-333(<-5h)	-1	2	19	3482(>58h)	60%

Table 1. Difference between Corrected and Original times for activities (in minutes)

	Min	Q1	Median	Q3	Max	-10 to 10 minutes
Start-time	-353(<-5h)	-2	1	3	50	87%
End-time	-333(<-6h)	-3	0	2	86	84%
Duration	-97	-4	-2	1	139	88%

Table 2. Difference between Corrected and Original times for trips (in minutes)

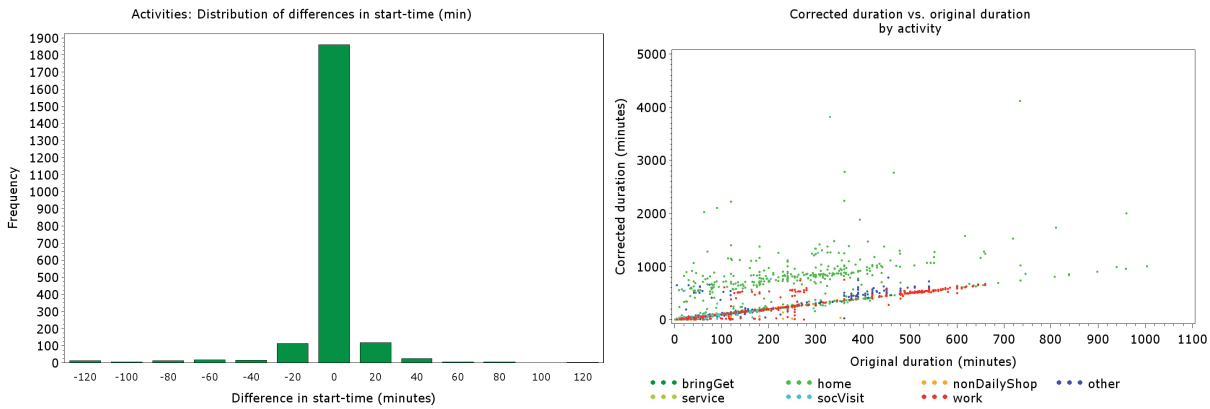

Fig. 2. (a) Activities: Difference in start times (b) Activity Duration-Corrected vs Original

4.3.2. Distributions of the differences in start-time, end-time and duration

The differences (corrected-original) were calculated (in minutes) for start-time, end-time and duration. For each, the distribution of the values was analyzed and summarized in table 1.

It can be seen from table 1 that the minimum and maximum corrections are rather extreme. However, the main bulk of the corrections is in a much smaller range. This is why we limited the range from -120 to 120 minutes for Figure 2(a). By doing so we dropped 1.22% of the observations only. Further analysis showed that most corrections were between -10 and 10 minutes. The percentages are in the last column in the table.

In order to obtain further insight in the correction process, we made some graphs. The graph of the differences in end-time vs. the differences in start-time showed a clear concentration around x=0 and y=0 meaning that often only one of the two times was changed.

The graph of the corrected duration vs. the original duration (Figure 2(b)) showed a clear concentration of points around y=x, meaning no change in duration, but there seems to be a parallel line above this line. Further investigation learned that these observations mainly come from people who claimed to have made trips but where the GPS did not show trips. In these cases the home activities were prolonged, which explains the parallel line (shifted over about 700 min) in the graph. Another possibility is that these persons left their GPS at home during the trip they claim to have made. It was impossible to find out which of both options was correct.

The graph was refined by using different colors for gender, age-class, profession, diploma, batch and activity-type. None of these revealed a clear trend, except that by activity type, which confirmed that the large changes in duration showing the parallel line were mainly for home-activities.

Table 2 summarizes the distributions of the differences in start-time, end-time and duration for trips.

The values are much less extreme than for the activities which is also reflected in the percentage of corrections between -10 and 10 minutes. Also for trips similar graphs were made as for activities. For trips, the graph of corrected duration vs. original duration did not show the parallel line, which makes sense since it was due to home-activities mainly. In those cases trips were deleted and no times were changed.

4.4. Discussion

Time recording was performed for the cleaning process. It turns out that solving inconsistencies in data reported by survey respondents, is a very time consuming task.

Allowing the respondent to enter inconsistent data results in a large cleaning cost and uncertainty because the interactive cleaning depends partially on the analyst judgment. On the other hand, this study does not give any insight into the effect of prohibiting the respondent to enter inconsistent data. One can argue that this method detects errors in an early stage of the data collection process where all information required for correction is still available. Unfortunately, the behavior of a respondent forced to supply consistent information while possibly being under time pressure, is not known.

5. Related Work

There are a lot of discrepancies between the self reported trips and GPS traces. Wolf et al.[1] described the results of several household travel surveys carried out in the USA from 2001 to 2004. According to Wolf et al.[1] the rate of trips missed ranged between 11% to 81%. Similarly in a study carried out by Murakami et al.[2], it was found out that there is a considerable difference between the self reported distances and the distances recorded by GPS.

P. Kelly et al.[3] describes an approach which assessed the errors in self reported travel diaries with the help of wearable cameras. The results suggest that there is an under reporting of trips by 14.6% and average trip duration is over reported by 12.1%.

An approach to assess the accuracy of Sydney household travel survey has been described by Stopher et al.[4] Accuracy of the survey was checked by comparing the results of a conventional paper and pencil face-to-face interview technique with the GPS based prompted recall approach. The travel times and travel distances were over-reported while the number of trips were under-reported by 7.4%.

According to Safi et al.[5], an increase of 4.5% in trips was observed while using a smarpthone application with post-processing analysis. Similarly Bohte et al.[6] emphasize that the number of trips missed are fewer while using a GPS-based prompted recall approach. Pereire et al.[7] reported an increase in accuracy of the duration of activities particularly for *work* activities. Abdulazim et al.[8] states that the burden on respondents during data collection should be minimal. The GPS based collection procedures are less burdensome for users as compared to the traditional data collection techniques.

Both Chaix et al.[9] and Hurvitz et al.[10] focus on the problem of diary and travel registration in health research. Hurvitz et al.[10] describes a campaign for data collection about physical movement GPS and accelerometer data are recorded automatically; respondents are provided with a smartphone app to record the diary (locations visited, travel mode, food intake etc.). The authors explicitly report that capturing the diary is problematic and there are discrepancies between the recorded and reported trip times.

6. Conclusions and Recommendations

Trip/activity data collection procedure is error prone. Moving from paper and pencil procedure to computer aided programs and smartphone applications was done to reduce errors and save time. A smartphone application was utilized to collect iMove data. Even with the use of a smartphone application, corrective diary alignment was required. Correction of diaries proved to be time consuming. Even after data correction, the data might not be completely accurate since the corrector does not know the whereabouts of the respondents and the corrections made are based on assumptions derived from the GPS traces. The proposed use of a prompted recall technique in the collection of data will reduce the burden on respondents as well as increase the quality of the data. The *prompted recall* technique refers to the annotation of stop locations using the GPS traces.

References

1. Wolf, J.. Applications of new technologies in travel surveys. In: *International Conference on Transport Survey Quality and Innovation*. Costa Rica; 2004, .
2. Murakami, E., Wagner, D.P.. Can using global positioning system (GPS) improve trip reporting? *Transportation Research Part C: Emerging Technologies* 1999;**7**(23):149 – 165. URL: `http://www.sciencedirect.com/science/article/pii/S0968090X99000170`. doi:http://dx.doi.org/10.1016/S0968-090X(99)00017-0.
3. Kelly, P., Doherty, A., Mizdrak, A., Marshall, S., Kerr, J., Legge, A., et al. High group level validity but high random error of a self-report travel diary, as assessed by wearable cameras. *Journal of Transport & Health* 2014;**1**(3):190 – 201. URL: `http://www.sciencedirect.com/science/article/pii/S2214140514000358`. doi:http://dx.doi.org/10.1016/j.jth.2014.04.003.
4. Stopher, P., FitzGerald, C., Xu, M.. Assessing the accuracy of the sydney household travel. *Transportation* 2007;**2007**(34):723–741. doi:10.1007/s11116-007-9126-8.
5. Safi, H., Mesbah, M., Ferreira, L.. Smartphone-assisted travel surveys: a smart way for transport planning. Sydney, Australia: CAITR (Conference of Australian Institutes of Transport Research); 2014, .
6. Bohte, W., Maat, K.. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the netherlands. *Transportation Research Part C: Emerging Technologies* 2009;**17**(3):285 – 297. URL: `http://www.sciencedirect.com/science/article/pii/S0968090X08000909`. doi:http://dx.doi.org/10.1016/j.trc.2008.11.004.
7. Pereira, F., Carrion, C., Zhao, F., Cottrill, C.. The future mobility survey: Overview and preliminary evaluation. In: *Proceedings of the Eastern Asia Society for Transportation Studies*; vol. 9. Taipei: Eastern Asia Society for Transportation Studies; 2013, .
8. Abdulazim, T., Abdelgawad, H., Habib, K.N., Abdulhai, B.. Using smartphones and sensor technologies to automate the collection of travel data. *TRB Research Record* 2013;.
9. Chaix, B., Meline, J., Duncan, S., Merrien, C., Karusisi, N., Perchoux, C., et al. {GPS} tracking in neighborhood and health studies: A step forward for environmental exposure assessment, a step backward for causal inference? *Health & Place* 2013;**21**(0):46 – 51. URL: `http://www.sciencedirect.com/science/article/pii/S1353829213000099`. doi:http://dx.doi.org/10.1016/j.healthplace.2013.01.003.
10. Hurvitz, P., Vernez Moudon, A., Kang, B., Saelens, B.. Emerging technologies for assessing physical activity behaviors in space and time. *Frontiers in Public Health* 2014;**2**. doi:10.3389/fpubh.2014.00002.