Collection Methods for Spatio-Temporal Micro-Scale Personal Movement Data

Konstantin Greger, Yuji Murayama
Division of Spatial Information Sciences
Graduate School of Life and Environmental Sciences
University of Tsukuba, Tsukuba, Japan

1 Introduction

The movements of individuals are defined and caused on the one hand by activities, routine and otherwise, and on the other hand by the constraints of the space they are moving within. Detailed knowledge about these movements can help solve many problems, from urban design, to transportation planning, retail expansion, disaster management, and so forth. Furthermore, a number of scientific examinations have shown the potential of statistically deriving activity patterns from existing such datasets.

In order to be useful for the aforementioned analyses these data should incorporate not only the actual spatio-temporal trajectory information, which indicates when the individuals have been at which geographic locations, but also semantic information, which bespeak the modes of transportation and the trip purposes, as well as socio-demographic attributes of the individuals (Fig. 1).

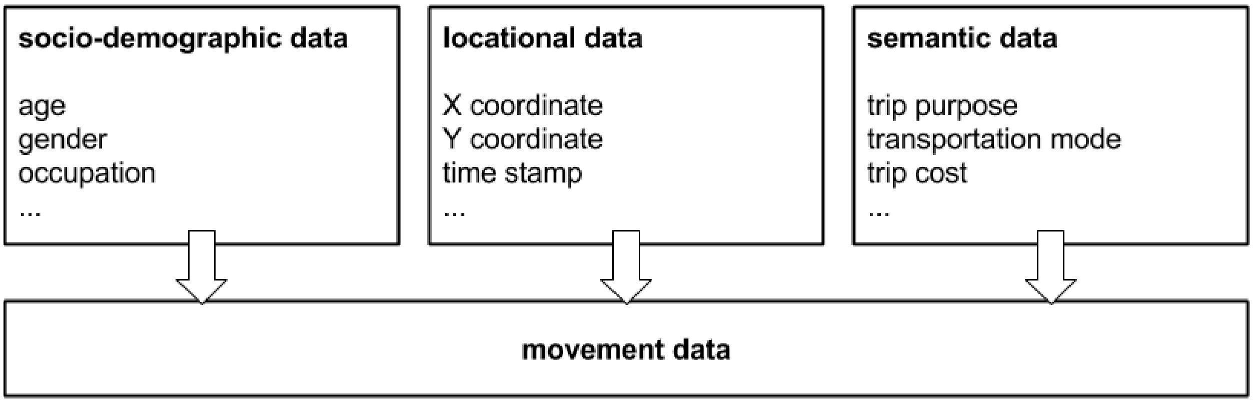


Figure 1: A complete movement dataset consists of three different dimensions of source data: socio-demographic, locational, and semantic.

In this report we investigate three different methodologies for collecting such micro-scale personal movement data: a questionnaire-based methodology, a methodology based on the use of GPS tracking devices, and finally a hybrid methodology, combining the two. We describe the data collection processes as well as the steps necessary to transform the data into formats that allow for their examination and statistical analyses, point out the advantages and shortcomings of the two traditional methodologies, and lastly show how the suggested hybrid methodology can overcome the latter. We do this by looking at previous studies and by designing an upcoming study that involves the collection of micro-scale movement data.

2 Questionnaire-based methodology

Travel diaries by the respondents have traditionally been used in many studies about urban and transportation planning. Examples are the study in Uppsala, Sweden, of 1971 (Hanson and Burnett 1981), the *UK National Travel Survey* (Department of the Environment 1994), the *Travel Tracker Survey* by the Chicago Metropolitan Agency for Planning (Chicago Metropolitan Agency for Planning 2008; Jiang et al. 2012), the 1999 *Mobidrive* study by the German Federal Ministry of Education and Research (Axhausen et al. 2002; PTV AG et al. 2000), and the *Persontrip* study by the Tokyo Metropolitan Area Transportation Planning Council (2013), which has been conducted every ten years since 1968. There is also a number of other studies that used similar data collection methods for smaller sample sizes and more specific applications (Doherty and Axhausen 1999; Doherty and Miller 1997; Ettema et al. 1995; Gärling et al. 1998; Mahmassani 1997). In this overview we focus mostly on the *2008 Tokyo Persontrip Study*.

2.1. Personal Information

All of the aforementioned studies collected data about the individuals as well as actual movement information. In the case of the *2008 Tokyo Persontrip* study the socio-demographic attributes comprised gender, age (in 5-year bins), occupation, occupation type, driver's license type, and car ownership (Tokyo Metropolitan Area Transportation Planning Council 2013). The German *Mobidrive* study performed a more far reaching collection of attributes, such as (amongst others) household size (in persons), existence of pets, car sharing habits, distance from home to closest public transportation facility, ownership of monthly passes for public transportation, presence of regular fixed-time commitments, etc. (PTV AG et al. 2000, pp.5–22).

In the questionnaire for the upcoming *Person Trip Tsukuba* study we are currently designing we decided to build upon the *2008 Tokyo Persontrip* model, but to extend it by a number of attributes that are of interest in the context of the study itself, which focuses on the main campus of University of Tsukuba and its immediate surroundings. Therefore we are using gender, age, family status, disabilities, living address, household size and structure, nationality, the duration of living in Tsukuba, Japanese language ability, occupation information, as well as a number of questions related to the status of students, researchers, teaching staff, and administrative staff, and questions regarding the individual's employment status and available modes of transportation.

2.2. Trip Information

In addition to the collection of the personal information the collection of the actual trip information is the central element of any study analyzing the movements of individuals.

Figures 2 and 3 show excerpts of the paper questionnaires that were used in the 1999 *Mobidrive* study in Germany and the 2008 *Tokyo Persontrip* study. Both ask specifically for the departure and arrival times of a trip, the sequence and duration of the usage of various modes of transportation, and the purpose of the trips. In addition, both forms ask for accompanying travelers and the cost involved, both in terms of parking fees and the usage of toll roads. Also, both forms require the respondent to enter the address of the departure and target locations, while the *Tokyo Persontrip* questionnaire also allows for the input of the name of landmarks or train station names.

These data can then be processed to derive trip chains and movement patterns of the sample individuals. Yet, this will only provide exact information about the whereabouts at certain points in time, where the persons are not moving but stay in one location. A team of researchers at the Center for Spatial Information Science (CSIS) at the University of Tokyo developed a methodology to synthesize the intermediate positions of the individuals into a dataset called *CSIS PFlow* (Sekimoto et al. 2011; Usui et al. 2009). This was achieved by the employment of routing algorithms, based on the various modes of transportation: walking or cycling; individual motorized transportation in cars, trucks and taxis; or use of public modes of transportation, such as trains or buses. This allowed for the creation of point positions for all sample individuals in 1-minute intervals.

This method of collecting trip data has several advantages and shortcomings, which we discuss in the following section. One of the greatest benefits of a paper questionnaire is the wide-spread ability among the population to fill it out, since all that is needed to do so is a pen. This makes it possible to collect data not only from larger sample sizes, but also eliminates all biases that can be introduced by the requirement of certain technologies or necessary devices. Yet, the fact that the questionnaires are mostly going to be filled out well after the actual trips took place, can introduce errors and generalizations in the data. An analysis of the departure and arrival time stamps of the *CSIS PFlow* dataset revealed that 88% of the stationary working activities supposedly started at round numbered minutes such as ":00", ":10", ":15" etc. 27% apparently started exactly at the full or half hour marks. In addition, erroneously filled-out fields in the questionnaire can produce wrong results in the analysis of the data. Also the entry of place names or addresses other than home or work locations can be challenging problems while filling out the questionnaire, since these are mostly not known to the sample individuals. Similarly, the translation of paper questionnaires into digital datasets by the analysts is quite error-prone, since it has to be manually performed by humans, who inevitably introduce mistakes and typing errors. An analysis of bicycle trips in the *CSIS PFlow* data showed a total of 127 trips longer than 50 km. The longest trip spans 143.8 km and was supposedly undertaken by a male office worker in his late fifties. The fact that this specific trip took only 15 minutes and the resulting average speed of 575.2 km/h makes it even more obvious that this data must be erroneous. Also, a cross-tabulation of the ages of the respondents and their occupations revealed some curious results (cf. Table 1). It is unclear at which step in the process these errors were introduced, i.e. while filling out the questionnaires or while translating them into a digital dataset.

In addition, while certainly useful and unique in its richness of information, the *CSIS PFlow* dataset has certain shortcomings on top of the aforementioned issues with the questionnaire-based collection of movement data. These originate mostly in the application of routing algorithms to interpolate the locations of individuals between stationary events. They will always assume that the people chose the shortest path from start to destination, which might not always be true in reality. Also while the routing of train passengers follows the actual train routes, it does not account for the actual departure and arrival times of trains and hence does no accurately represent the locations of the passengers.

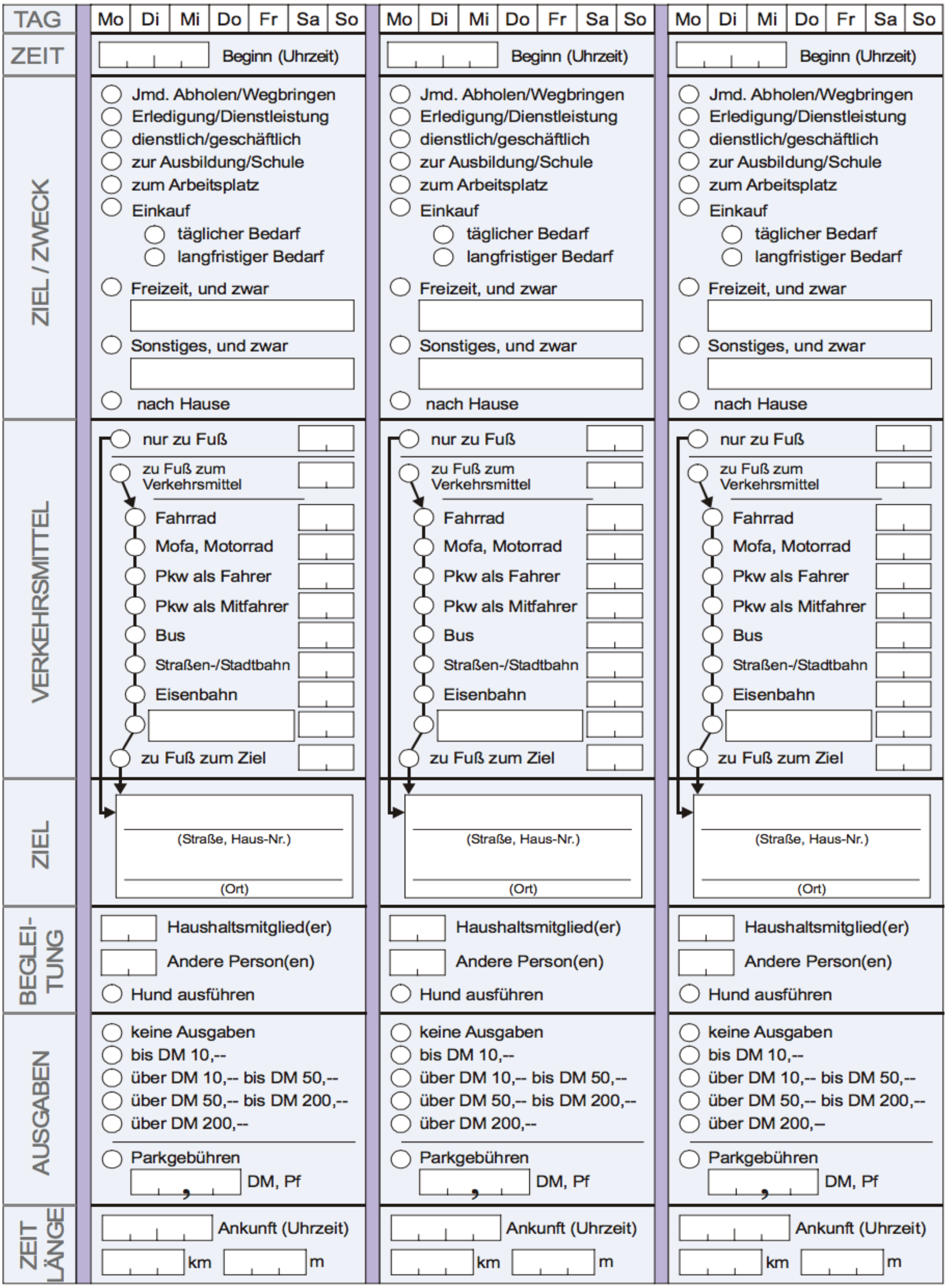
A similar method to collect the trip information is the use of online questionnaires, as we did in the case of the *Person Trip Tsukuba* study. This method ameliorates one of the major shortcomings of paper-based questionnaires, the error-prone translation of paper questionnaires into digital datasets, but instead introduces a possible bias due to the necessary availability of access to the online resources and knowledge about their usage. Furthermore, the problem of having to know the exact addresses and location names could be ameliorated by the usage of a web map that allows the sample individuals to select the locations in a more visual way.

We developed an online questionnaire loosely based on the 2008 *Tokyo Persontrip* paper questionnaire, but extended it by some transportation means and trip purposes which we deemed of interest for the study in Tsukuba (Fig. 4). The questionnaire is built upon the Google Drive platform for the design and publication of online forms and questionnaires. We chose this platform since it is usable for free, it allows for the necessary flexibilities in the questionnaire design while keeping the necessary development efforts to a minimum, and it also allows for an easy export of the collected data as Google Spreadsheets and CSV files.

Table 1: Cross-tabulation of the ages of the respondents of the 2008 Tokyo Persontrip study and their occupations.

Age	agriculture, forestry, fishery	production, industry	sales	services	transport & communication	security	business	professional & technical	administration	other occupation	kindergarten, elem., junior HS student	high school student	university student	housewife, househusband	unemployed	other	[undocumented code 91]	unknown
5-9											26,114							4
10-14											27,578							5
15-19	15	231	90	208	30	9	108	106	1	41	3,304	14,727	5,533	8	371	43	35	23
20-24	66	843	1,366	2,581	406	103	2,701	3,763	82	380		58	11,027	311	1,560	66	157	72
25-29	117	1,313	2,132	4,326	901	272	6,413	8,753	339	680		10	893	2,362	2,012	87	253	139
30-34	184	1,879	2,511	5,327	1,465	303	8,587	11,514	798	817		3	304	7,181	2,008	70	365	156
35-39	264	2,394	2,936	5,667	1,742	313	10,220	12,390	1,825	1,025		3	131	10,701	2,058	98	501	213
40-44	251	2,062	2,600	4,957	1,427	241	9,154	11,414	3,391	1,064		2	67	8,632	1,677	48	519	186
45-49	276	1,821	2,088	4,069	1,155	290	7,513	9,025	4,162	1,018		2	44	6,474	1,343	45	467	170
50-54	353	1,735	1,827	3,806	1,066	285	5,706	7,152	4,295	1,066		1	27	6,512	1,447	22	411	184
55-59	638	2,665	2,203	4,770	1,465	433	5,843	7,098	5,215	1,279		2	22	9,627	3,205	44	510	362
60-64	1,016	2,330	1,792	4,130	1,211	442	3,851	5,391	3,982	1,421		6	53	10,838	9,575	76	653	739
65-69	1,393	1,505	1,107	2,509	652	337	1,503	2,729	2,066	1,170		4	27	9,748	16,241	112	577	1,333
70-74	1,526	847	608	963	220	125	463	1,191	963	573		1	19	7,180	17,332	95	376	1,493
75-79	1,391	413	283	397	76	31	197	556	518	281			13	4,547	14,550	55	179	1,196
80-84	900	186	161	170	14	7	60	213	221	133		4	1	2,278	10,379	31	72	887
>85	460	84	71	120	11	2	22	65	108	89			1	807	9,375	55	34	597

Source: CSIS *PFlow* study dataset



TAG	Mo Di Mi Do Fr Sa So	Mo Di Mi Do Fr Sa So	Mo Di Mi Do Fr Sa So
ZEIT	□ Beginn (Uhrzeit)	□ Beginn (Uhrzeit)	□ Beginn (Uhrzeit)
ZIEL / ZWECK	○ Jmd. Abholen/Wegbringen ○ Erledigung/Dienstleistung ○ dienstlich/geschäftlich ○ zur Ausbildung/Schule ○ zum Arbeitsplatz ○ Einkauf ○ täglicher Bedarf ○ langfristiger Bedarf ○ Freizeit, und zwar □ ○ Sonstiges, und zwar □ ○ nach Hause	○ Jmd. Abholen/Wegbringen ○ Erledigung/Dienstleistung ○ dienstlich/geschäftlich ○ zur Ausbildung/Schule ○ zum Arbeitsplatz ○ Einkauf ○ täglicher Bedarf ○ langfristiger Bedarf ○ Freizeit, und zwar □ ○ Sonstiges, und zwar □ ○ nach Hause	○ Jmd. Abholen/Wegbringen ○ Erledigung/Dienstleistung ○ dienstlich/geschäftlich ○ zur Ausbildung/Schule ○ zum Arbeitsplatz ○ Einkauf ○ täglicher Bedarf ○ langfristiger Bedarf ○ Freizeit, und zwar □ ○ Sonstiges, und zwar □ ○ nach Hause
VERKEHRSMITTEL	○ nur zu Fuß □ ○ zu Fuß zum Verkehrsmittel □ ○ Fahrrad □ ○ Mofa, Motorrad □ ○ Pkw als Fahrer □ ○ Pkw als Mitfahrer □ ○ Bus □ ○ Straßen-/Stadtbahn □ ○ Eisenbahn □ ○ □ □ ○ zu Fuß zum Ziel □	○ nur zu Fuß □ ○ zu Fuß zum Verkehrsmittel □ ○ Fahrrad □ ○ Mofa, Motorrad □ ○ Pkw als Fahrer □ ○ Pkw als Mitfahrer □ ○ Bus □ ○ Straßen-/Stadtbahn □ ○ Eisenbahn □ ○ □ □ ○ zu Fuß zum Ziel □	○ nur zu Fuß □ ○ zu Fuß zum Verkehrsmittel □ ○ Fahrrad □ ○ Mofa, Motorrad □ ○ Pkw als Fahrer □ ○ Pkw als Mitfahrer □ ○ Bus □ ○ Straßen-/Stadtbahn □ ○ Eisenbahn □ ○ □ □ ○ zu Fuß zum Ziel □
ZIEL	(Straße, Haus-Nr.) (Ort)	(Straße, Haus-Nr.) (Ort)	(Straße, Haus-Nr.) (Ort)
BEGLEITUNG	□ Haushaltsmitglied(er) □ Andere Person(en) ○ Hund ausführen	□ Haushaltsmitglied(er) □ Andere Person(en) ○ Hund ausführen	□ Haushaltsmitglied(er) □ Andere Person(en) ○ Hund ausführen
AUSGABEN	○ keine Ausgaben ○ bis DM 10,-- ○ über DM 10,-- bis DM 50,-- ○ über DM 50,-- bis DM 200,-- ○ über DM 200,-- ○ Parkgebühren □ , □ DM, Pf	○ keine Ausgaben ○ bis DM 10,-- ○ über DM 10,-- bis DM 50,-- ○ über DM 50,-- bis DM 200,-- ○ über DM 200,-- ○ Parkgebühren □ , □ DM, Pf	○ keine Ausgaben ○ bis DM 10,-- ○ über DM 10,-- bis DM 50,-- ○ über DM 50,-- bis DM 200,-- ○ über DM 200,– ○ Parkgebühren □ , □ DM, Pf
ZEIT LÄNGE	□ Ankunft (Uhrzeit) □ km □ m	□ Ankunft (Uhrzeit) □ km □ m	□ Ankunft (Uhrzeit) □ km □ m

Figure 2: Excerpt of the paper questionnaire travel diary used in the German Mobidrive study.

Source: PTV AG et al. 2000, p.29

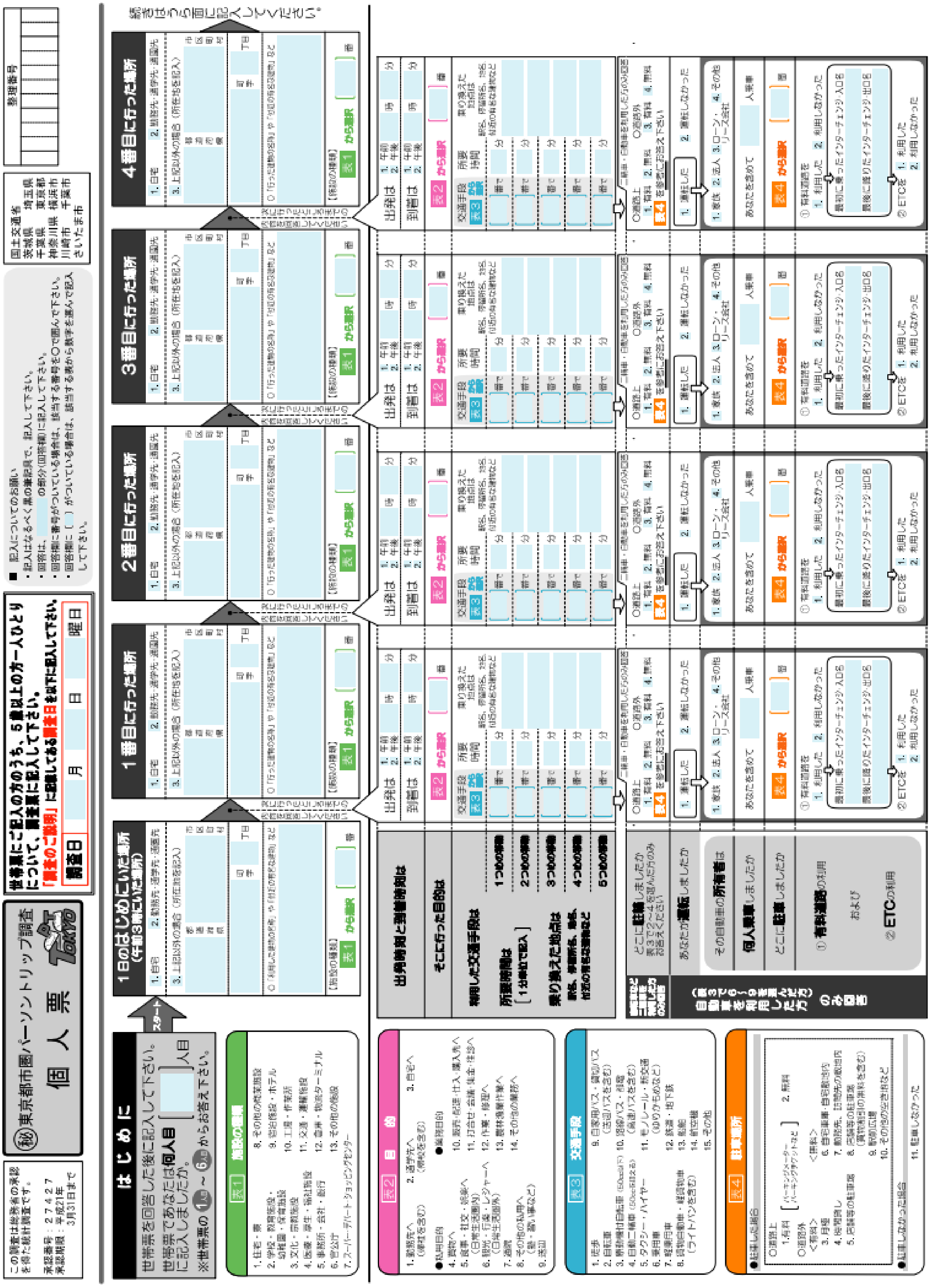

この調査は総務省の承認を得た統計調査です。
承認番号：27427
承認期限：平成21年3月31日まで
㊙東京都市圏パーソントリップ調査
個人票
Tokyo
世帯票にご記入の方のうち、5歳以上の方一人ひとりについて、調査票に記入して下さい。
「調査のご説明」に記載してある調査日を以下に記入して下さい。
調査日 月 日 曜日
記入についてのお願い
国土交通省 茨城県 埼玉県 千葉県 東京都 神奈川県 横浜市 川崎市 千葉市 さいたま市
整理番号
はじめに
世帯票を回答した後に記入して下さい。
世帯票であなたは何人目に記入しましたか。 人目
※世帯票の1人目～6人目からお答え下さい。
スタート
1日のはじめにいた場所（午前3時にいた場所）
1番目に行った場所
2番目に行った場所
3番目に行った場所
4番目に行った場所
1. 自宅 2. 勤務先・通学先・通園先 3. 上記以外の場合（所在地を記入）
表1 から選択
出発時刻と到着時刻は
出発は 到着は 1. 午前 2. 午後 時 分
そこに行った目的は
表2 から選択
利用した交通手段は
所要時間は［1分単位で記入］
乗り換えた地点は
駅名、停留所名、駐車場名など
1つめの手段 2つめの手段 3つめの手段 4つめの手段 5つめの手段
交通手段 表3 から選択
どこに駐車しましたか
あなたが運転しましたか
1. 運転した 2. 運転しなかった
その自動車の所有者は
1. 家族 2. 法人 3. ローン・リース会社 4. その他
何人乗車しましたか
人乗車
どこに駐車しましたか
表4 から選択
① 有料道路の利用 および ② ETCの利用
1. 利用した 2. 利用しなかった
最初に乗ったインターチェンジ 入口名
最後に降りたインターチェンジ 出口名
表1 施設の種類
表2 目的
表3 交通手段
表4 駐車場所
続きは裏面に記入して下さい。

Figure 3: Excerpt of the paper questionnaire travel diary used in the 2008 Persontrip Tokyo study.

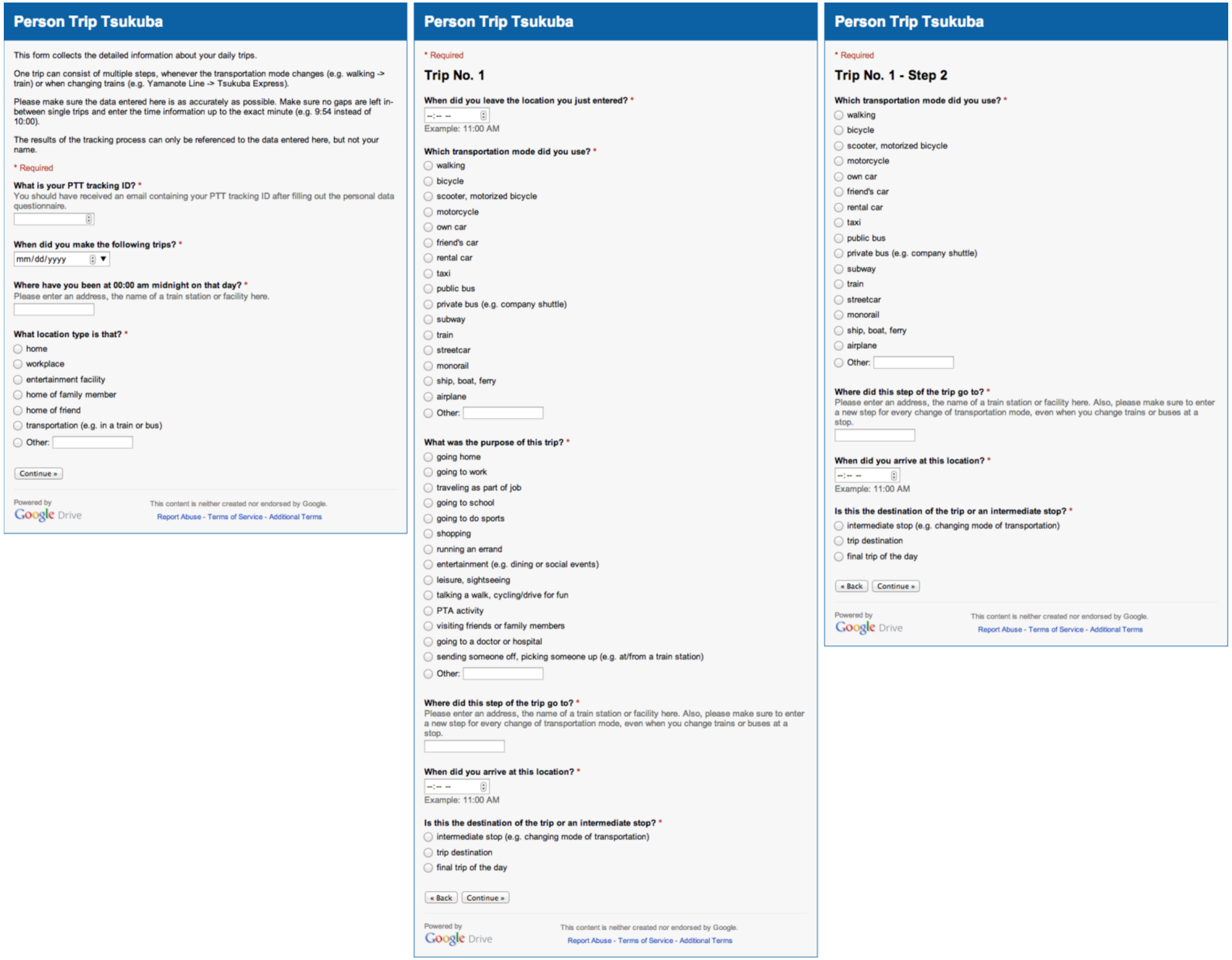
Source: Tokyo Metropolitan Area Transportation Planning Council 2013, p.29

3 GPS tracking-based methodology

Just like in the case of the data collection using questionnaires, the GPS tracking-based methodology requires the collection of personal information upfront. Also, the study design needs to enable a link between the personal information data collected by questionnaire and the trip information data collected by a GPS tracking device. This can be achieved by assigning unique identifiers for all sample individuals.

This GPS tracking-based methodology of collecting trip data also has several advantages and shortcomings. One of the biggest advantages is that it does not require the sample individuals to keep detailed diaries of the trips they undertook. Instead the tracking device will save their actual locations at any time of the day in bespoke time intervals (e.g. 1 minute, or 5 minutes). Yet, this automated localization also introduces error sources and presents potential for erroneous data output. One of the reasons is the precision of non-military GPS devices, which varies typically between two and 50 meters, depending on the structure of the surroundings and the visibility of enough satellites to perform a precise location triangulation. This visibility can be hampered by thick vegetation, such as in forests, or a high building density, such as in highly urbanized city centers. Furthermore, inside buildings and other built-up structures, such as road and subway tunnels, the localization by GPS is impossible. Other disadvantages of the use of GPS tracing devices are the limited availability of these devices among the population, which can create a bias in the selection of sample individuals, the fact that the sample individuals will have to make sure to carry the tracking device at all times, and the need for maintenance of these devices, especially in terms of battery power. The first and second point can somewhat be ameliorated by the growing use and spread of smartphones, which often contain the hardware necessary to perform a location tracking, either by GPS or by a triangulation of mobile phone cell information. Various software products, so called apps, for most mobile operating systems (i.e. Apple iOS, Google Android, Windows Phone, BlackBerry OS) are also available, some of them even for free.

For the purpose of comparing the results of the GPS tracking-based data collection methodology with those of the questionnaire-based data, we performed a brief sample study. One important aspect in selecting an app for this type of study is a number of criteria: the availability of the respective smartphone app on a large number of mobile operating systems; the possibility to export the collected data from the app to a bespoke format; and the cost of the app. After a thorough screening of a large number of available alternatives we decided to use the app *Moves* (ProtoGeo Oy 2014), since it is available on both iOS and Android, which together make up 99% of the Japanese smartphone market (Kantar Worldpanel ComTech 2014), and it allows to export the data as either GPX file or as a JSON string via an external free service on the website moves-



Person Trip Tsukuba

This form collects the detailed information about your daily trips.

One trip can consist of multiple steps, whenever the transportation mode changes (e.g. walking -> train) or when changing trains (e.g. Yamanote Line -> Tsukuba Express).

Please make sure the data entered here is as accurately as possible. Make sure no gaps are left in-between single trips and enter the time information up to the exact minute (e.g. 9:54 instead of 10:00).

The results of the tracking process can only be referenced to the data entered here, but not your name.

* Required

What is your PTT tracking ID? *
You should have received an email containing your PTT tracking ID after filling out the personal data questionnaire.

When did you make the following trips? *
mm/dd/yyyy

Where have you been at 00:00 am midnight on that day? *
Please enter an address, the name of a train station or facility here.

What location type is that? *

○ home
○ workplace
○ entertainment facility
○ home of family member
○ home of friend
○ transportation (e.g. in a train or bus)
○ Other:

Continue »

Powered by Google Drive

This content is neither created nor endorsed by Google.
Report Abuse - Terms of Service - Additional Terms

Person Trip Tsukuba

* Required

Trip No. 1

When did you leave the location you just entered? *
--:-- --
Example: 11:00 AM

Which transportation mode did you use? *

○ walking
○ bicycle
○ scooter, motorized bicycle
○ motorcycle
○ own car
○ friend's car
○ rental car
○ taxi
○ public bus
○ private bus (e.g. company shuttle)
○ subway
○ train
○ streetcar
○ monorail
○ ship, boat, ferry
○ airplane
○ Other:

What was the purpose of this trip? *

○ going home
○ going to work
○ traveling as part of job
○ going to school
○ going to do sports
○ shopping
○ running an errand
○ entertainment (e.g. dining or social events)
○ leisure, sightseeing
○ talking a walk, cycling/drive for fun
○ PTA activity
○ visiting friends or family members
○ going to a doctor or hospital
○ sending someone off, picking someone up (e.g. at/from a train station)
○ Other:

Where did this step of the trip go to? *
Please enter an address, the name of a train station or facility here. Also, please make sure to enter a new step for every change of transportation mode, even when you change trains or buses at a stop.

When did you arrive at this location? *
--:-- --
Example: 11:00 AM

Is this the destination of the trip or an intermediate stop? *

○ intermediate stop (e.g. changing mode of transportation)
○ trip destination
○ final trip of the day

« Back Continue »

Powered by Google Drive

This content is neither created nor endorsed by Google.
Report Abuse - Terms of Service - Additional Terms

Person Trip Tsukuba

* Required

Trip No. 1 - Step 2

Which transportation mode did you use? *

○ walking
○ bicycle
○ scooter, motorized bicycle
○ motorcycle
○ own car
○ friend's car
○ rental car
○ taxi
○ public bus
○ private bus (e.g. company shuttle)
○ subway
○ train
○ streetcar
○ monorail
○ ship, boat, ferry
○ airplane
○ Other:

Where did this step of the trip go to? *
Please enter an address, the name of a train station or facility here. Also, please make sure to enter a new step for every change of transportation mode, even when you change trains or buses at a stop.

When did you arrive at this location? *
--:-- --
Example: 11:00 AM

Is this the destination of the trip or an intermediate stop? *

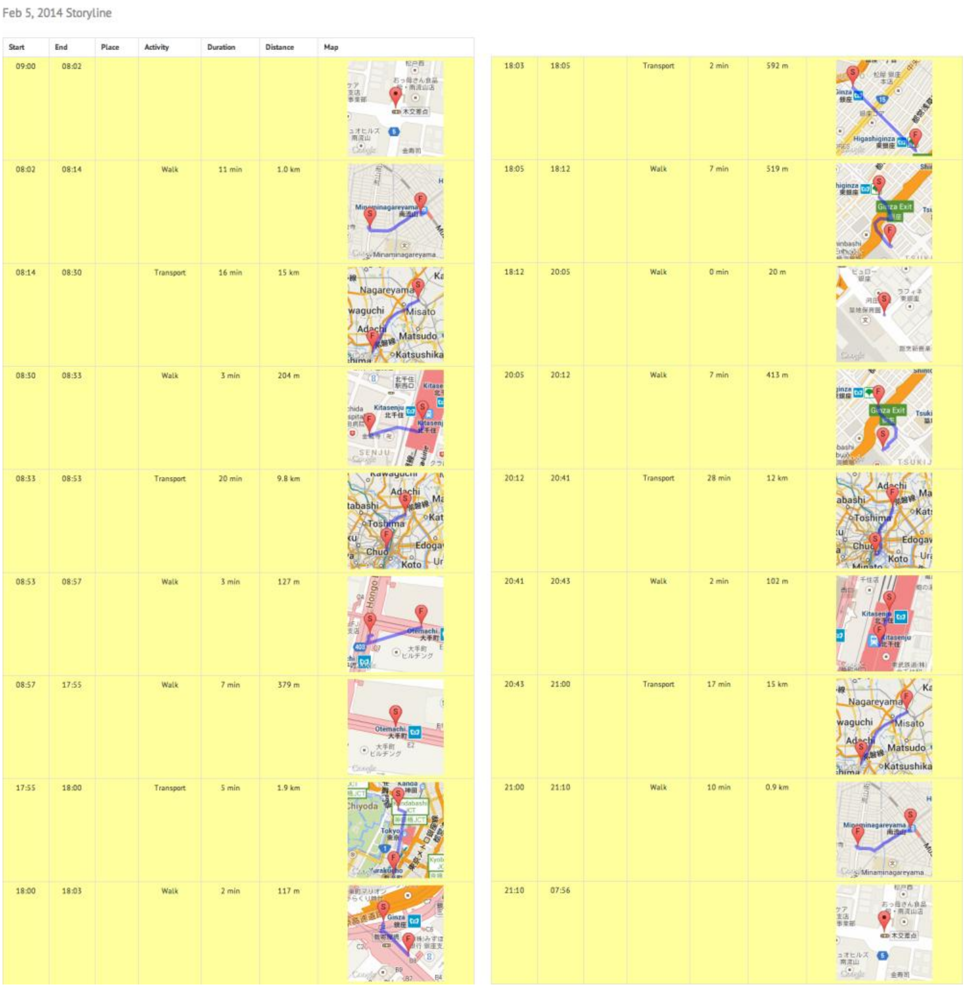
○ intermediate stop (e.g. changing mode of transportation)
○ trip destination
○ final trip of the day

« Back Continue »

Powered by Google Drive

This content is neither created nor endorsed by Google.
Report Abuse - Terms of Service - Additional Terms

Figure 4: Excerpt of the online questionnaire travel diary used in the Person Trip Tsukuba study.



Feb 5, 2014 Storyline

Start	End	Place	Activity	Duration	Distance	Map
09:00	08:02					
08:02	08:14		Walk	11 min	1.0 km	
08:14	08:30		Transport	16 min	15 km	
08:30	08:33		Walk	3 min	204 m	
08:33	08:53		Transport	20 min	9.8 km	
08:53	08:57		Walk	3 min	127 m	
08:57	17:55		Walk	7 min	379 m	
17:55	18:00		Transport	5 min	1.9 km	
18:00	18:03		Walk	2 min	117 m	
18:03	18:05		Transport	2 min	592 m	
18:05	18:12		Walk	7 min	519 m	
18:12	20:05		Walk	0 min	20 m	
20:05	20:12		Walk	7 min	413 m	
20:12	20:41		Transport	28 min	12 km	
20:41	20:43		Walk	2 min	102 m	
20:43	21:00		Transport	17 min	15 km	
21:00	21:10		Walk	10 min	0.9 km	
21:10	07:56					

Figure 5: Data collected by the Moves app, visualized on the website moves-export.com.

In order to be able to assess the quality of the data involved and also the feasibility of using export.com (Harris 2014). The only disadvantage of the *Moves* app is its price at ¥ 300.

In order to be able to assess the quality of the data involved and also the feasibility of using the exported data in an automated data processing and analysis workflow, we collected data from one sample individual on seven consecutive days from February 1st to 7th, 2014, using the *Moves* app on an Apple iPhone 5 on iOS 7. The results were rather mixed: for five of the seven days the app failed to collect any data, that results in a 71% loss rate. On the other hand, on the remaining two days the app provided complete and very accurate data (cf. Fig. 5).

The app uses an internal algorithm to assign the most likely mode of transportation between “walking”, “cycling”, “running”, and “transport”, which comprises both motorized and public transportation. On the two successfully logged sample days all transportation modes were detected correctly as “walking” and “transport”. While the GPS tracks of the longer “walking” segments showed very accurate location data, the shorter “walking” segments, especially during train changes inside station buildings, and “transport” segments on subway routes showed some deviations. Yet, the locations of the trip start and end point were detected very accurately in all cases. Phases of stationarity, where the sample individual does not move significantly in space (e.g. while at home or in the office) are not detected by the algorithm and are generally represented as “walking” trips, but when exported as a GPX file the movement data is automatically split into separate tracks at these stationarity events. Figure 6 shows the GPX data for one of the data collection days visualized on Google Maps. The three different colors denote the three trips “from home to work” (red), “from work to sport” (green), and “from sport to home” (blue).

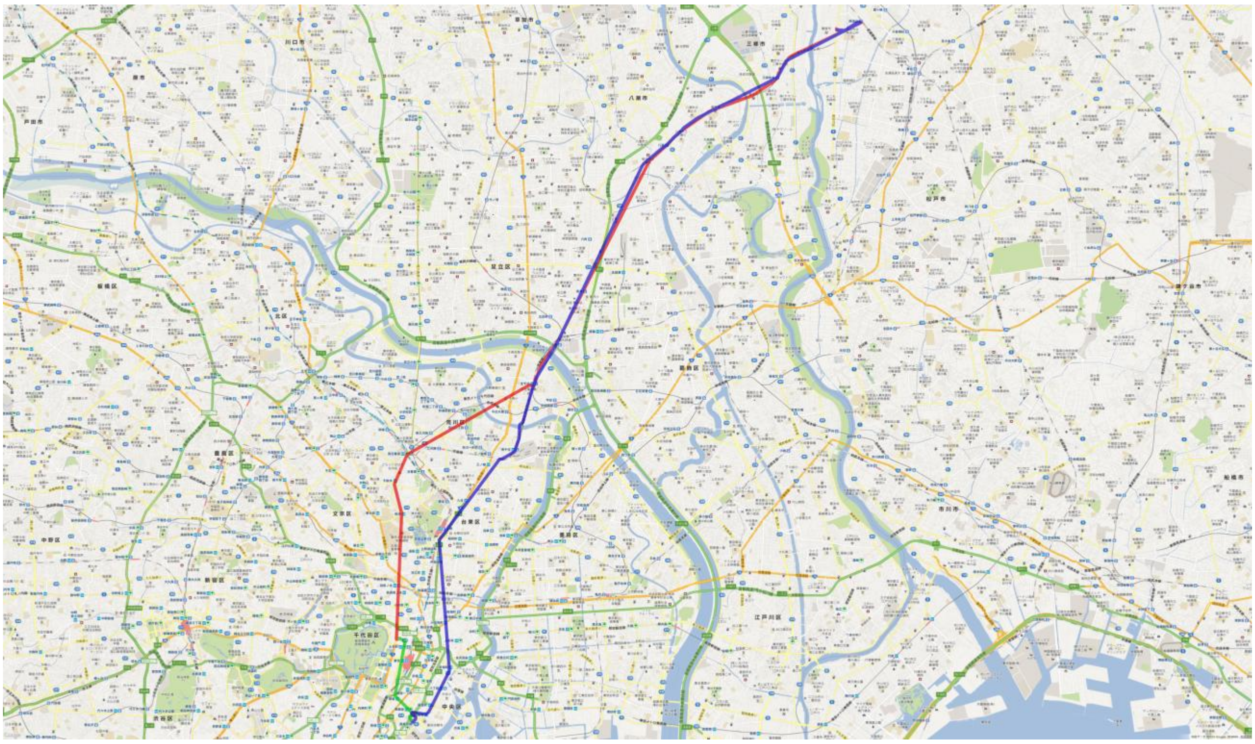


Figure 6: Data collected by the Moves app, exported as GPX file via moves-export.com and visualized on Google Maps.

The JSON data exported by the moves-export.com website provides a lot more data, both in terms of granularity and attributes:

```
[{
  "segments": [
    {
      "startTime": "20140204T000000Z",
      "place": {
        "id": 17663939,
        "location": {
          "lon": 139.8963221498,
          "lat": 35.8362447408
        },
        "type": "unknown"
      },
      "endTime": "20140204T230258Z",
      "type": "place"
    },
    {
      "startTime": "20140204T230258Z",
      "activities": [
        {
          "startTime": "20140204T230258Z",
          "distance": 974,
          "duration": 662,
          "trackPoints": [
            {
              "time": "20140204T230258Z",
              "lon": 139.8963221498,
              "lat": 35.8362447408
            },

... shortened for readability ...

            {
              "time": "20140204T231400Z",
              "lon": 139.903093202,
              "lat": 35.8378296714
            }
          ],
          "steps": 1230,
          "endTime": "20140204T231400Z",
          "activity": "wlk"
        },
        {
          "startTime": "20140204T231400Z",
          "distance": 15261,
          "duration": 991,
          "trackPoints": [
            {
              "time": "20140204T231400Z",
              "lon": 139.903093202,
              "lat": 35.8378296714
            },

... shortened for readability ...

            {
              "time": "20140204T233030Z",
              "lon": 139.8048985312,
              "lat": 35.7491724105
            }
          ],
          "endTime": "20140204T233031Z",
          "activity": "trp"
        },

... shortened for readability ...

        {
          "startTime": "20140204T235358Z",
          "distance": 127,
          "duration": 188,
          "trackPoints": [
            {
              "time": "20140204T235358Z",
              "lon": 139.7636590335,
              "lat": 35.6867170874
            },

... shortened for readability ...

            {
              "time": "20140204T235706Z",
              "lon": 139.7645131989,
              "lat": 35.6868279162
            }
          ],
          "steps": 254,
          "endTime": "20140204T235706Z",
          "activity": "wlk"
        }
      ],
      "endTime": "20140204T235706Z",
      "type": "move"
    },

... shortened for readability ...

    {
      "startTime": "20140205T121041Z",
      "place": {
        "id": 17663939,
        "location": {
          "lon": 139.8963221498,
          "lat": 35.8362447408
        },
        "type": "unknown"
      },
      "endTime": "20140205T225621Z",
      "type": "place"
    }
  ],
  "date": "20140205"
}]
```

In order to be able to process this data we developed a short Python script to convert the data from the JSON format into CSV files:

```
# -*- coding: utf-8 -*-
"""
processMovesExportJSON.py

Script to process the JSON output of moves-export.com into a structured,
2-dimensional table for export in CSV or database tables.

Author: Konstantin Greger
"""

import json
import datetime

# initialization
UTCadjust = 9                                   # local timezone of data (e.g.: UTC+9 for JST)
csvSeparator = ";"                              # separator to use in output
inputFileName = "jsonstoryline_20140206.json"   # path and filename of the input JSON string
outputFileName = "storyline_20140206.csv"       # path and filename of the CSV output file

inputFile = open(inputFileName)
data = json.load(inputFile)
json = data[0]['segments']
inputFile.close()

outputFile = open(outputFileName, "w")
outputString = ("ID","tripID","subtripID","trackpointID","type","mode","lon","lat",↓
                "timestamp","origin")
outputFile.write(csvSeparator.join(outputString) + "¥n")

# parse data from JSON string into CSV format
ID = 1
tripID = 1
for segment in json:
    if segment['type'] == "place":
        # stationarity event
        subtripID = 1           # dummy value
        trackpointID = 1
        stype = segment['type']
        mode = segment['type']  # dummy value
        lon = segment['place']['location']['lon']
        lat = segment['place']['location']['lat']
        timestamp = datetime.datetime.strptime(str(segment['startTime']), "%Y%m%dT%H%M%SZ")
        # adjust UTC timestamp by timezone offset
        timestamp += datetime.timedelta(hours = UTCadjust)
        if ID == 1:
            # special treatment for a day's first dataset
            timestamp = datetime.datetime.strptime(data[0]['date'], "%Y%m%d")
        timestamps = timestamp.strftime("%Y-%m-%d %H:%M:%S")
        outputString = (str(ID),str(tripID),str(subtripID),str(trackpointID),str(stype),↓
                        str(mode),str(lon),str(lat),str(timestamps),"d")
        outputFile.write(csvSeparator.join(outputString) + "¥n")
        ID += 1
        trackpointID += 1
        # synthesize intermediate stationary timesteps
        endtimestamp = datetime.datetime.strptime(str(segment['endTime']), "%Y%m%dT%H%M%SZ")
        endtimestamp += datetime.timedelta(hours = UTCadjust)
        while timestamp < endtimestamp:
            timestamp += datetime.timedelta(seconds = 1)
            if timestamp >= datetime.datetime.strptime(data[0]['date'], "%Y%m%d") + ↓
            datetime.timedelta(days = 1):
                break               # stop at 23:59:59
        timestamps = timestamp.strftime("%Y-%m-%d %H:%M:%S")
        outputString = (str(ID),str(tripID),str(subtripID),str(trackpointID),str(stype),↓
                        str(mode),str(lon),str(lat),str(timestamps),"s")
        outputFile.write(csvSeparator.join(outputString) + "¥n")
        ID += 1
        trackpointID += 1
elif segment['type'] == "move":
    # actual movement
    stype = segment['type']
    subtripID = 1
    for activities in segment['activities']:
        mode = activities['activity']
        trackpointID = 1
        for trackpoints in activities['trackPoints']:
            lon = trackpoints['lon']
            lat = trackpoints['lat']
            timestamp = datetime.datetime.strptime(str(trackpoints['time']), ↓
                       "%Y%m%dT%H%M%SZ")
            # adjust UTC timestamp by timezone offset
            timestamp += datetime.timedelta(hours = UTCadjust)
            if timestamp >= datetime.datetime.strptime(data[0]['date'], "%Y%m%d") + ↓
            datetime.timedelta(days = 1):
                break               # stop at 23:59:59
            if trackpointID == 1:
                prevTimestamp = timestamp
                prevLon = lon
                prevLat = lat
            if ID == 1:
                # special treatment for a day's first dataset
                timestamp = datetime.datetime.strptime(data[0]['date'], "%Y%m%d")
            else:
                if timestamp > prevTimestamp + datetime.timedelta(seconds = 1):
                    # save data for next timestep
                    nextLon = lon
                    nextLat = lat
                    nextTimestamp = timestamp
                    # synthesize intermediate movement timesteps
                    secsDiff = (nextTimestamp - prevTimestamp).seconds
                    lonDiff = nextLon - prevLon
                    latDiff = nextLat - prevLat
                    for i in range(1, secsDiff):
                        lon = prevLon + ((lonDiff / secsDiff) * i)
                        lat = prevLat + ((latDiff / secsDiff) * i)
                        timestamp = prevTimestamp + datetime.timedelta(seconds = i)
                        if timestamp >= datetime.datetime.strptime(data[0]['date'], ↓
                        "%Y%m%d") + datetime.timedelta(days = 1):
                            break               # stop at 23:59:59
                        timestamps = timestamp.strftime("%Y-%m-%d %H:%M:%S")
                        outputString = (str(ID),str(tripID),str(subtripID),↓
                                        str(trackpointID),str(stype),str(mode),↓
                                        str(lon),str(lat),str(timestamps),"s")
                        outputFile.write(csvSeparator.join(outputString) + "¥n")
                        ID += 1
                        trackpointID += 1
                    # restore data for next timestep
                    lon = nextLon
                    lat = nextLat
                    timestamp = nextTimestamp
            if timestamp >= datetime.datetime.strptime(data[0]['date'], ↓
            "%Y%m%d") + datetime.timedelta(days = 1):
                break               # stop at 23:59:59
            timestamps = timestamp.strftime("%Y-%m-%d %H:%M:%S")
            outputString = (str(ID),str(tripID),str(subtripID),str(trackpointID),str(stype),↓
                            str(mode),str(lon),str(lat),str(timestamps),"d")
            outputFile.write(csvSeparator.join(outputString) + "¥n")
            trackpointID += 1
            ID += 1
            prevLon = lon
            prevLat = lat
            prevTimestamp = timestamp
        subtripID += 1
tripID += 1
if timestamp >= datetime.datetime.strptime(data[0]['date'], "%Y%m%d") + ↓
datetime.timedelta(days = 1):
    break               # stop at 23:59:59
```

This creates data like that shown below for the exemplar day of data collection. These can then be imported into software such as R, SPSS, or Excel, and various database systems such as PostgreSQL, MySQL, or Microsoft SQL Server. The source code is publicly available on the corresponding author's GitHub repository at https://github.com/kogreger/moves-export. Since the *Moves* app doesn't collect the location in defined time steps, the time differences between the locations vary widely. In our script we amended this shortcoming by synthesizing intermediate locations in one second intervals. As can be seen from the excerpt the locations that represent actual collected data are marked by a “d” in the last column, while the synthesized locations are marked by an “s”.

```
1;1;1;1;place;place;139.89632215;35.8362447408;2014-02-05 00:00:00;d
2;1;1;2;place;place;139.89632215;35.8362447408;2014-02-05 00:00:01;s
3;1;1;3;place;place;139.89632215;35.8362447408;2014-02-05 00:00:02;s
4;1;1;4;place;place;139.89632215;35.8362447408;2014-02-05 00:00:03;s

... shortened for readability ...

28980;2;1;1;move;wlk;139.89632215;35.8362447408;2014-02-05 08:02:58;d
28981;2;1;2;move;wlk;139.896256503;35.8363162546;2014-02-05 08:02:59;s
28982;2;1;3;move;wlk;139.896190856;35.8363877684;2014-02-05 08:03:00;s
28983;2;1;4;move;wlk;139.896125209;35.8364592821;2014-02-05 08:03:01;s
28984;2;1;5;move;wlk;139.896059562;35.8365307959;2014-02-05 08:03:02;d
28985;2;1;6;move;wlk;139.896161588;35.8365582664;2014-02-05 08:03:03;d
28986;2;1;7;move;wlk;139.896142279;35.8365385632;2014-02-05 08:03:04;s
28987;2;1;8;move;wlk;139.89612297;35.8365188601;2014-02-05 08:03:05;s
28988;2;1;9;move;wlk;139.896103661;35.8364991569;2014-02-05 08:03:06;s
28989;2;1;10;move;wlk;139.896084352;35.8364794538;2014-02-05 08:03:07;s
28990;2;1;11;move;wlk;139.896065043;35.8364597507;2014-02-05 08:03:08;s
28991;2;1;12;move;wlk;139.896045734;35.8364400475;2014-02-05 08:03:09;d

... shortened for readability ...

29641;2;1;662;move;wlk;139.903104141;35.8378352982;2014-02-05 08:13:59;s
29642;2;1;663;move;wlk;139.903093202;35.8378296714;2014-02-05 08:14:00;d
29643;2;2;1;move;trp;139.903093202;35.8378296714;2014-02-05 08:14:00;d
29644;2;2;2;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:01;d
29645;2;2;3;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:02;s
29646;2;2;4;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:03;s
29647;2;2;5;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:04;s
29648;2;2;6;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:05;s
29649;2;2;7;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:06;s
29650;2;2;8;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:07;s
29651;2;2;9;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:08;s
29652;2;2;10;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:09;s
29653;2;2;11;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:10;s
29654;2;2;12;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:11;s
29655;2;2;13;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:12;s
29656;2;2;14;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:13;s
29657;2;2;15;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:14;s
29658;2;2;16;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:15;s
29659;2;2;17;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:16;s
29660;2;2;18;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:17;s
29661;2;2;19;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:18;s
29662;2;2;20;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:19;s
29663;2;2;21;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:20;s
29664;2;2;22;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:21;s
29665;2;2;23;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:22;s
29666;2;2;24;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:23;s
29667;2;2;25;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:24;s
29668;2;2;26;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:25;s
29669;2;2;27;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:26;s
29670;2;2;28;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:27;s
29671;2;2;29;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:28;s
29672;2;2;30;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:29;s
29673;2;2;31;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:30;s
29674;2;2;32;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:31;s
29675;2;2;33;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:32;s
29676;2;2;34;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:33;s
29677;2;2;35;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:34;s
```

```
29678;2;2;36;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:35;s
29679;2;2;37;move;trp;139.902985443;35.8377751304;2014-02-05 08:14:36;d

... shortened for readability …
```

This data can then also be visualized on a map, as shown in Figure 7. Here we also visualized two of the attributes inherent in the data: the left hand chart shows mode of transportation, where green represents “walking” and red “transport”; the right hand chart shows the differences between collected locations in black and synthesized locations in red. It is obvious how the higher speed on the train (locations from 29,643 in the CSV data) results in greater distances between collected locations and also synthesized locations compared to the “walking” segment (locations from 28,980 in the CSV data). Also, the straight lines between very few captured locations are representations of the synthesizing due to missing data as a result of the use of underground transportation.

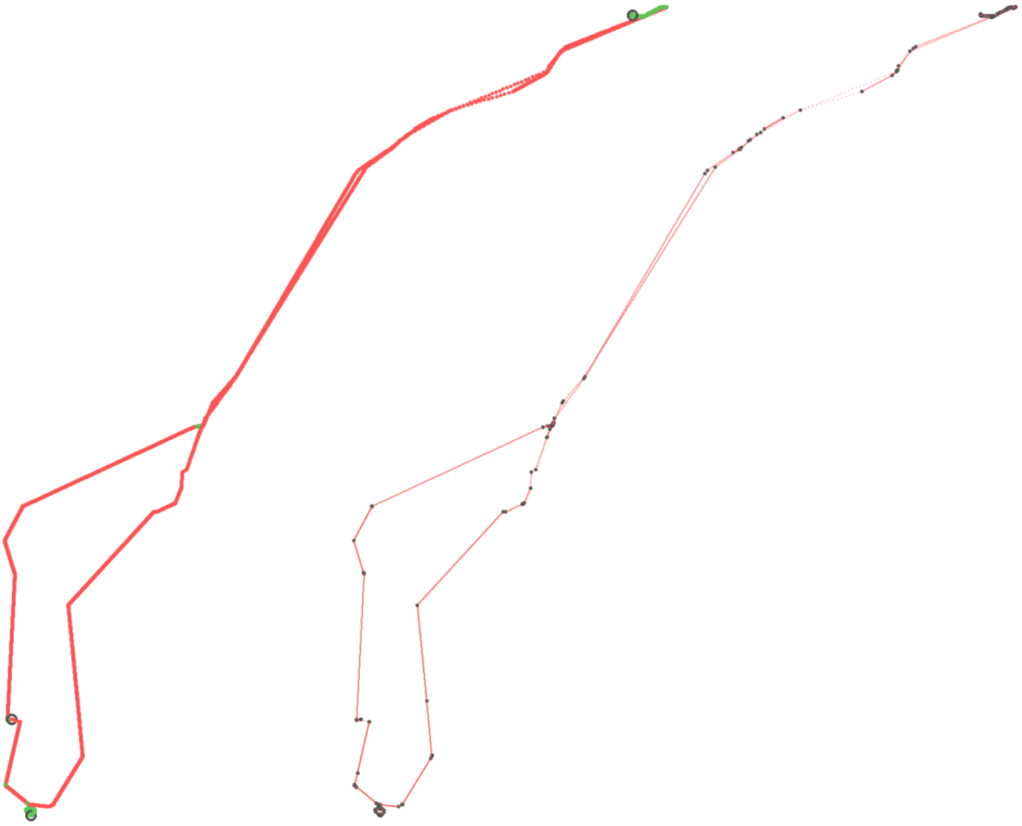


Figure 7: Visualization of synthesized CSV output of the processed JSON string regarding the mode of transportation (left: green = walking, red = transport) and collected vs. synthesized locations (right).

It is obvious how the GPS tracking-based data collection methodology simplifies the process of collecting the location component of the movement data. The sample individual doesn't need to take care of collecting the data while traveling or remembering details about visited places and the accompanying times. Yet, at the same time it is obvious that most of the semantic components such as the trip purpose are missing. On the other hand advanced data analysis methods already allow for the automated extraction of some additional attributes. An example is the ability of the *Moves* app to derive the mode of transportation from the collected data.

4 Hybrid methodology

In this report we suggest a hybrid methodology for the collection of movement profiles. This combines the benefits of the GPS tracking-based methodology regarding the collection of locational data, and those of the questionnaire-based methodology regarding the collection of the semantic components as shown in Figure 8.

The idea is to collect the data in an automated fashion as introduced in chapter 3 using a GPS device that allows for the export of the collected data in a standardized data format (e.g. GPX, JSON, CSV, etc.). This data then needs to be preprocessed, which can already derive some of the semantic components from the locational data. An example is the mode of transportation that the *Moves* app is able to detect from the point locations and movements speeds. The final step is an online questionnaire that presents the preprocessed locational data and possibly automatically derived semantic attributes to the sample individual, who then needs to fill out the missing attributes. At this step it is important to also allow for the correction, amendment or deletion of the preprocessed data, since errors in these automated processes are possible.

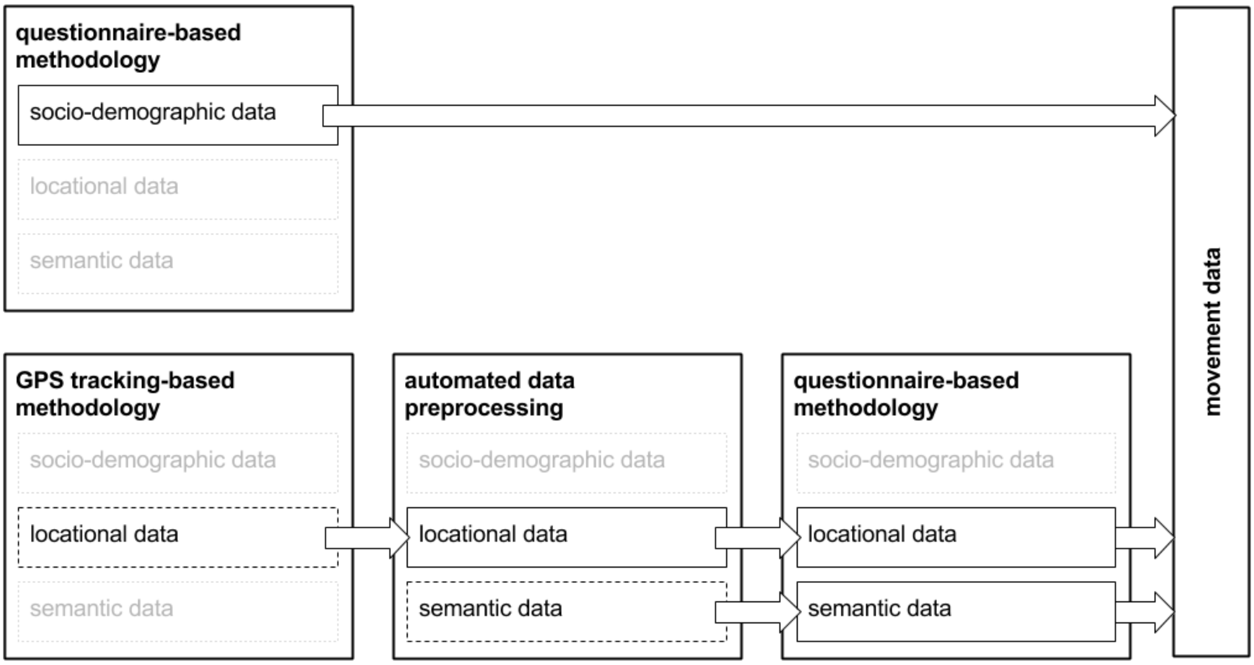


Figure 8: Schematic representation of the suggested hybrid data collection methodology.

Such a hybrid movement data collection system is currently in development at the Division for Spatial Information Science at the University of Tsukuba under the supervision of the authors.

5 Conclusion

In this paper we discussed several methodologies that can be used in the context of the collection of movement data: questionnaire-based (both paper-based and online), GPS tracking-based, and a novel hybrid approach. Table 2 sums up their advantages and disadvantages. It is obvious how none of the existing methodologies can produce satisfactorily precise data across all three dimensions of movement data: socio-demographic, locational, and semantic. Therefore only a combination of the GPS tracking-based methodology for the easy and precise collection of locational data, a preprocessing step for data cleansing, synthesizing of intermediary data and the automated generation and derivation of some semantic attributes, and lastly an online questionnaire for the collection of missing semantic data and the correction, amendment and deletion of erroneous data from the previous automated steps.

Acknowledgements

We thank the Center for Spatial Information Science at the University of Tokyo for providing the 2008 *PFlow* dataset for the Tokyo Metropolitan Area by the CSIS People Flow Project Office within the scope of joint research No. 405.

Table 2: Comparison of advantages and disadvantages of the trip data collection methodologies presented in this paper.

Collection methodology	Advantages	Disadvantages
Paper-based questionnaire	• wide-spread ability to use it • large sample size possible • unbiased sample possible • no technical devices necessary	• *post-hoc* data collection can introduce errors and omissions • possible generalizations and simplifications when filling out • entry of place names or addresses other than home or work locations can be challenging • error-prone translation of paper questionnaires into digital datasets • only location information about stationarity events, in-between movements have to be synthesized additionally
Online questionnaire	• possible usage of a web map to allow the selection of locations in a more visual way • no translation of paper questionnaires into digital datasets necessary	• *post-hoc* data collection can introduce errors and omissions • possible generalizations and simplifications when filling out • entry of place names or addresses other than home or work locations can be challenging • only location information about stationarity events, in-between movements have to be synthesized additionally • possible bias due to necessary access to online resources and knowledge about their usage
GPS tracking-based methodology	• simplification of the process of collecting the location component of the movement data • no need to take care of data collection (e.g. remembering visited places and the accompanying times) while traveling • data processing allows for the synthesizing of missing intermediary data • advanced data analysis methods allow for the automated extraction of some additional attributes	• most of the semantic components are missing (e.g. trip purpose) • low precision of locational data at high movement speeds or while traveling underground, indoors or beneath vegetation cover
Hybrid methodology	• easy and precise collection of locational data • preprocessing step allows for data cleansing, synthesizing of intermediary data and the automated generation and derivation of some semantic attributes • online questionnaire for the collection of missing semantic data, correction, amendment and deletion of erroneous data from the previous automated steps	• development effort necessary

References

Axhausen K W, Zimmermann A, Schönfelder S, Rindsfüser G, and Haupt T 2002 Observing the rhythms of daily life: A six-week travel diary. Transportation 29:95–124

Chicago Metropolitan Agency for Planning 2008 Chicago Regional Household Travel Inventory: Final Report.

Department of the Environment 1994 National travel survey: 1991/93. London, HMSO

Doherty S T, and Axhausen K W 1999 A unified framework for the development of a weekly household activity-travel scheduling model. In: Brilon W, Huber F, Schreckenberg M, Wallentowitz H (eds) Traffic and Mobility Simulation - Economics - Environment. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 35–56

Doherty S T, and Miller E J 1997 Tracing the household activity scheduling process using a one week computer-based survey. Austin, TX

Ettema D F, Borgers A W J, and Timmermans H J P 1995 Competing risk hazard model of activity choice, timing, sequencing, and duration. *Transportation Research Record* 1493:101–109

Gärling T, Kalén T, Romanus J, Selart M, and Vilhelmson B 1998 Computer simulation of household activity scheduling. *Environment and Planning A* 30:665–679

Hanson S, and Burnett K O 1981 Understanding complex travel behavior: Measurement issues. In: Stopher PR, Meyburg AH, Brög W (eds) New horizons in travel-behavior research. Lexington Books, Lexington, Mass, pp 207–230

Harris N 2014 Moves Export. WWW document, http://www.moves-export.com/

Jiang S, Ferreira J, and Gonzalez M C 2012 Discovering urban spatial-temporal structure from human activity patterns. ACM Press:95

Kantar Worldpanel ComTech 2014 Kantar Worldpanel ComTech December 2013. WWW document, http://www.kantarworldpanel.com/dwl.php?sn=news_downloads&id=399

Mahmassani H S 1997 Dynamics of commuter behaviour: Recent research and continuing challenges. In: Stopher PR, Lee-Gosselin M (eds) Understanding travel behaviour in an era of change, 1st ed. Pergamon, Oxford, OX, UK ; Tarrytown, N.Y., U.S.A, pp 279–313

ProtoGeo Oy 2014 Moves. WWW document, http://www.moves-app.com/

PTV AG, Fell B, Schönfelder S, and Axhausen K W 2000 Mobidrive questionnaires.

Sekimoto Y, Shibasaki R, Kanasugi H, Usui T, and Shimazaki Y 2011 PFlow: Reconstructing People Flow Recycling Large-Scale Social Survey Data. *IEEE Pervasive Computing* 10:27–35

Tokyo Metropolitan Area Transportation Planning Council 2013 パーソントリップ調査とは (About the Persontrip Study). WWW document, http://www.tokyo-pt.jp/person/index.html

Usui T, Kanasugi T, Sekimoto Y, Minami Y, Shibasaki R, and Nanako A 2009 Realization and implementation of Tokyo Metropolitan Area person-trip data spatio-temporal interpolation by a People Flow Analysis Platform. In *Papers and Proceedings of the Geographic Information Systems Association*. 541–54