Using GPS, GIS, and Accelerometer Data to Predict Transportation Modes

RUBEN BRONDEEL[1,2,3], BRUNO PANNIER[4], and BASILE CHAIX[1,2]

[1]Institut National de la Santé et de la Recherche Médicale, UMR_S 1136, Pierre Louis Institute of Epidemiology and Public Health, Research Team in Social Epidemiology, Paris, FRANCE; [2]Sorbonne Universités, Université Pierre et Marie Curie Univ Paris 06, UMR_S 1136, Pierre Louis Institute of Epidemiology and Public Health, Research Team in Social Epidemiology, Paris, FRANCE; [3]Ecole des Hautes études en Santé Publique School of Public Health, Rennes, FRANCE; and [4]IPC Medical Centre, Paris, FRANCE

ABSTRACT

BRONDEEL, R., B. PANNIER, and B. CHAIX. Using GPS, GIS, and Accelerometer Data to Predict Transportation Modes. *Med. Sci. Sports Exerc.*, Vol. 47, No. 12, pp. 2669–2675, 2015. **Introduction**: Active transportation is a substantial source of physical activity, which has a positive influence on many health outcomes. A survey of transportation modes for each trip is challenging, time-consuming, and requires substantial financial investments. This study proposes a passive collection method and the prediction of modes at the trip level using random forests. **Methods**: The RECORD GPS study collected real-life trip data from 236 participants over 7 d, including the transportation mode, global positioning system, geographical information systems, and accelerometer data. A prediction model of transportation modes was constructed using the random forests method. Finally, we investigated the performance of models on the basis of a limited number of participants/trips to predict transportation modes for a large number of trips. **Results**: The full model had a correct prediction rate of 90%. A simpler model of global positioning system explanatory variables combined with geographical information systems variables performed nearly as well. Relatively good predictions could be made using a model based on the 991 trips of the first 30 participants. **Conclusions**: This study uses real-life data from a large sample set to test a method for predicting transportation modes at the trip level, thereby providing a useful complement to time unit-level prediction methods. By enabling predictions on the basis of a limited number of observations, this method may decrease the workload for participants/researchers and provide relevant trip-level data to investigate relations between transportation and health. **Key Words:** PHYSICAL ACTIVITY, ACTIVE TRANSPORT, PASSIVE DATA COLLECTION, MACHINE LEARNING, RECORD COHORT STUDY, FRANCE

Physical activity has a positive influence on several health outcomes, such as obesity, cardiovascular health problems, depression, and certain cancers (15,36,39). Active transportation modes, such as walking, biking, and public transport, represent a substantial source of physical activity (27,28). However, reliably assessing the use of transportation modes has proven challenging (9,11,12), thereby hindering the study of the relation between transportation and physical activity. Self-reported measures of the use of transportation modes are prone to memory biases (3). Short trips, especially walking trips, tend to be underreported. Moreover, the time spent during car trips tends to be underreported, whereas the time spent in public transport tends to be exaggerated.

Using objective measurements using accelerometers or global positioning system (GPS) receivers is useful to overcome some of these issues. These devices can, in theory, register the spatial location and body movements of participants over several days. The difficulty lies in transforming the raw data into qualitative trip information, such as the transportation modes used or the departure and arrival locations of each trip.

One approach used in transportation sciences is to perform a so-called GPS-based prompted recall survey, i.e., using information derived from GPS receivers to prompt participant recall (32,38). Using this approach, GPS and accelerometer data are first collected. The departure and arrival points (in space and time) of each trip are then identified by detecting the activity places, i.e., the places visited

Address for correspondence: Ruben Brondeel, M.Sc., UMR_S 1136, Faculté de Médecine Saint-Antoine, 27 Rue Chaligny, 75012 Paris, France; E-mail: Ruben.Brondeel@iplesp.upmc.fr.

Submitted for publication November 2014.
Accepted for publication May 2015.
Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.acsm-msse.org).

0195-9131/15/4712-2669/0
MEDICINE & SCIENCE IN SPORTS & EXERCISE®

DOI: 10.1249/MSS.0000000000000704

by the participant for which a function can be identified such as a residence, workplace or shop. One technique to identify departure and arrival points involves manually segmenting the trips with geographical information systems (GIS) (31). Another approach is to apply algorithms that identify the departure and arrival points of trips on the basis of the raw GPS data (33), as conducted in the RECORD GPS study. Finally, the resulting information is verified and data on the transportation mode in each trip are collected via phone or Internet recall surveys with the participants (3,11,12). Combining device and survey data, the memory bias, and social desirability bias in survey data are reduced by the objective measures. With this approach, information derived concerning trips using the manual processing or automatic algorithms is completed with the survey information. Such GPS-based prompted recall surveys can be performed either at the end of the observation period or on a daily basis during this period (1,32); the latter method is useful for reducing memory biases.

More recently, SenseCam, a camera worn around the neck that takes pictures at regular intervals or when triggered by imbedded sensors, has been suggested to improve data collection of daily activities including trips (6,16,30). Pictures are then used to identify transportation modes or other trip characteristics.

To obtain high-quality data using these approaches, a substantial investment from both participants and research teams is required. In the RECORD GPS study, in which we performed a complete mobility survey for an observation period of 7 d, a research assistant was often able to survey only one participant per day (the entire process included the preparation, the survey, and entering the data into the application). Using SenseCam is likely to be even more burdensome, as research assistants must code all photographs. The time and cost investments required for data collection strongly limit the number of participants, whereas the burden on participant limits the extent of the remainder of the survey.

Therefore, researchers have developed algorithms to predict transportation modes on the basis of device data and sometimes on a limited number of survey items (18,20). Most of these algorithms designed to recognize modes consider short periods (time units) ranging from 1 to 60 s. These algorithms sometimes use sliding windows to optimize the prediction for a given unit using the information from one or more previous and subsequent time frame units. In addition to transportation modes, certain classifications take into account body posture (including lying, sitting, standing, etc.) or household activities. Classifications in these algorithms are based on criteria-based methods, machine learning (such as random forests, support vector machine, and Bayesian network), and probability methods (such as fuzzy logic and multinomial regression) (20).

A smaller number of detection methods, such as the present one, uses trips or trip stages (parts of trips made by a single transportation mode) as the prediction level. These methods first segment the data into trips and activity places and then predict the transportation mode for each trip. This additional step of segmenting the data into trips is an obvious drawback compared with time unit prediction methods. However, trips are meaningful units in behavioral and transportation sciences when analyzing transport-related issues. For example, when studying physical activity associated with the use of public transport, the walking distance required to travel to a train or bus is more important than the physical activity needed during the actual use of these modes. These types of research questions therefore must be addressed at the trip level, thereby making prediction models at the trip level complementary to prediction models at the time unit level.

The present study does not address the segmentation of trips process (algorithms are available for this first step (33)) but rather focuses on transportation mode detection. The aim of this study was to construct an algorithm, building on passive data collection methods that reduce the burden of work for both respondents and research teams. The approach should yield reliable predictions of the transportation mode used at the trip level, which could reduce the time required for the mobility survey or even allow researchers to avoid it completely. We propose a method based on random forests to predict transportation modes at the trip level.

METHODS

Population. As previously described in detail, the RECORD participants were recruited during preventive health checkups in 2007–2008 and 2011–2013, born between 1928 and 1978, and resided at baseline in 112 municipalities of the Île-de-France Paris region (5,7,13,34). In the second wave of the study (8,26), after undergoing a medical checkup and filling computerized questionnaires at the IPC Medical Centre (10,23), 410 individuals were invited to participate in the RECORD GPS study (9), of which 247 subjects agreed to participate. Nine participants abandoned the study, and data collection failed for two participants, thereby yielding a final participation and completion rate of 57.6% (n = 236). A written informed consent was obtained from all participants. The RECORD GPS study was approved by the French Data Protection Authority.

Data collection procedures. The recruitment was guided using a standardized recruitment form. Participants wore a BT-Q1000XT GPS (QStarz) and a GT3X+ accelerometer (ActiGraph) on the right hip with a dedicated elastic belt for the recruitment day and seven additional days, all day long from the time of waking up until bedtime. The participants completed a travel diary to report their activity places over 7 to 8 d, each time with arrival and departure times.

Using a GIS-based Python language algorithm (33) to assess the GPS data, we identified the sequence of activity places for each participant and, consequently, the departure and arrival times of trips between these places. The algorithm automatically uploaded the history of visits to places into the electronic survey application. As previously described (9), this information and the travel diary were then used for

the prompted recall survey conducted during a phone call (10). This procedure resulted in the observation of 7425 trips for 236 participants.

Measures. During the survey, participants reported a chronological sequence of transportation modes for each trip. For modeling purposes, this information was coded into a transportation mode variable consisting of four categories: "walking" (i.e., only walking), "bicycle," "private motorized," and "public transport." When both walking and another transportation mode were sequentially used within a trip, the nonwalking mode was attributed to the trip. We excluded 96 trips with two or more nonwalking modes because they could not be attributed to the mutually exclusive categories of modes required to perform the comparison and there were not enough trips with each combination of two nonwalking modes to define additional categories.

The random forests method is able to use a large variety of variables as predictors of the outcome of interest. However, because the aim of the study is also to lower the burden for researchers, we only used predictors that are relatively easy to define, such as GPS and accelerometer variables, GIS variables that require only standard data, and seven simple survey questions.

The accelerometer recorded the acceleration on three axes for each 5-s epoch or period during the trip. We used both the standard filter and a low-frequency extension filter (37), as implemented in the ActiLife software. The optional low-frequency extension filter extends the lower end of the filter, which is useful for example when processing the data of people who move slowly. On the basis of the raw accelerations obtained with these two filtering approaches, we estimated for each epoch 1) the number of footsteps taken (ActiLife software), 2) the energy expenditure calculated from activity counts and participant weight based on the Sasaki and Freedson equation (29), 3) whether moderate-to-vigorous physical activity (MVPA) was performed (29), and 4) whether the participant was sedentary during the epoch (22). We aggregated these time unit data at the trip level. To capture a maximum of relevant information, we derived standard measures of central tendency (i.e., mean and median) and measures of dispersion (i.e., SD, minimum, maximum, 10th and 90th percentiles). On the basis of the accelerometer data, the accelerations at each of the three axes separately, the number of steps taken, MVPA, sedentary time, and energy expenditure in kilocalories were aggregated in this way. In addition, we calculated the total number of steps taken, the number of MVPA epochs, the number of sedentary epochs, and total energy expenditure for each trip. We also determined the percentage of epochs that were characterized sedentary or MVPA. Each of these variables was calculated for both accelerometer filters.

Every 5 s, the GPS device registered the position coordinates (i.e., latitude, longitude, and elevation), speed, and the following three indicators of the quality of the observation: horizontal, vertical, and positional dilution of precision (HDOP, VDOP, and PDOP, respectively). To derive the summary values described earlier, only the good-quality observations (HDOP < 6, VDOP < 7, PDOP < 8) were retained (9) for the aggregation of time-unit observations at the trip level. GPS observations were determined to be valid, invalid (high dilution of precision), or missing (less than three satellites in view). On average, 27% of GPS observations were missing and 1.5% of the existing observations were invalid. The distribution of potential GPS data points across these three categories provides information on the circumstances of the trips (e.g., underground public transport, tunnel, high buildings). To capture this trip characteristic, the total number of GPS observations, number of valid GPS observations, percentage of valid GPS observations among recorded observations, and percentage of valid GPS observations relative to the maximum number of observations (including missing ones) were included in the model.

On the basis of the GPS data and geographical information on the street network provided by the National Geographic Institute, four distance measures between the departure and arrival points of each trip were calculated, as follows: the straight line distance, the shortest walking distance following the street network, the shortest street network distance by car, and the map-matched distance. The latter distance is based on the most likely route taken by the participant derived by projecting the GPS data points onto the street network (35). These four distance measures and their combination provide complementary information to differentiate between alternative transportation modes. For example, for two trips for which the shortest distance by car would be the same, a difference in the shortest walking distance could add information to differentiate between motorized and nonmotorized transport. Speed measures were calculated on the basis of these distance measures. The GIS was also used to determine whether the residence and the departure and arrival points of each trip were inside or outside the Paris inner city. All geographic calculations were conducted with Python scripts for ArcGIS 10.1. The administrative files of the study provided the sex and age of the participants. During phone call interviews, it was recorded whether the participant possessed a car, bicycle, motorbike, driving license, or public transport pass. Supplemental Digital Content 1 provides an overview of the variables used in the prediction model (see Table, Supplemental Digital Content 1, Overview of 170 predictors used in the random forest models, http://links.lww.com/MSS/A549).

Statistical analysis. We used random forests to predict the transportation mode of each trip (among four possible modes). The random forests method (4) is based on the decision tree method. Decision trees classify data into groups in subsequent steps, each time searching for the feature that best differentiates the group into consideration (branch). To obtain better generalizability, the random forests method adds two sources of randomness to the simple decision tree method and repeats the process a large number of times, thereby resulting into a forest of decision trees. The first source of randomness consists of considering only a random subsample

TABLE 1. Observed and predicted number of trips with each transportation mode.

	Observed			
Predicted	**Walking**	**Bicycle**	**Private Motorized**	**Public**
Walking	3010	59	229	76
Bicycle	6	115	1	1
Motorized	107	26	2565	112
Public	35	13	65	909

n = 7329.

of the explanatory variables in the definition of each knot of the trees. Secondly, for each tree, only a random subsample of the observations (the trips in our case) is used. Predictions are obtained from each tree for the data not used to grow the tree (so-called out-of-bag data). Finally, a forest prediction of the transportation mode is obtained for each trip as the majority of the tree predictions that were derived when the corresponding trip was out-of-bag. A forest is evaluated on the prediction error rate, in our case, the percentage of trips for which the mode has been wrongly predicted. Regarding missing values, we attributed the median value or the modal value to the corresponding observations for continuous or categorical variables, respectively. All analyses were performed using R with the "randomForest" package (24).

RESULTS

Among the 7329 trips retained for the analyses, 43.1% of the trips were made by walking, 2.9% were with a bike, 39.0% relied on a private motorized vehicle, and 15.0% relied on public transport. The median duration of a trip was 15 min (interdecile range, 3–61 min).

A first forest was grown on the full data set of 7329 trips with all 170 variables. The model had an overall error rate of 10.0% and specific error rates of 4.7% for walking, 46.0% for biking, 10.3% for private motorized transport, and 17.2% for public transport. Table 1 cross-tabulates the observed versus the predicted number of trips for each mode.

The overall error rate was relatively low, but the error rate was larger for the modes with a lower number of trips, such as bicycle or public transport use. When minimizing the overall error rate, classification methods favor precision in the categories with a greater number of observations over precision in the categories with a lower number of observations (25). When interested in greater precision for the smaller categories, the majority-vote-prediction rule can be weighted by the inverse of the probability of belonging to a category. This method greater penalizes the decision rule for mistakes in smaller categories. Growing a random forest using this method, the error rate for the prediction of "bicycle" and "public transport" dropped to 16.9% and 12.8%, respectively. The error rate for the larger categories ("walking" and "private motorized") rose to 14.4% and 19.9%, respectively. The overall error rate rose to 16.4%.

The importance of the source of information (accelerometer, GPS, or GPS/GIS data) was then evaluated using separate forests grown with only the respective subsamples of variables (Table 2). The overall error rate for the forest with only the accelerometer variables was 17.7%. The overall error rates for the forests with GPS variables only and GPS/GIS variables only were 17.6% and 11.6%, respectively. Interestingly, the latter error rate was thus not markedly higher than the error rate of the full model (10.0%).

To mimic a study in which participant and trip data are used to predict modes for subsequent trips, forests were grown on the basis of the first 5, 10, 20, 30, 40, 50, 100, 150, and 200 participants. These forests were evaluated by using the prediction error rates for subsequently observed participants (Table 3). A model based on the first five participants (143 trips) yielded a prediction error rate of 28% for the other 231 participants (7187 trips). The overall error rate dropped and then stabilized when at least 30 participants were used to grow the forest (991 trips). The error rates for transportation modes with a larger number of trips were relatively small even for the model based on only a few participants. The gain in prediction quality was relatively small when including additional participants (i.e., more than 30 individuals) in the model. For the transportation modes with a small number of trips, the error rate was high in models with few participants, and it dropped relatively slowly. The reduction in the error rates became negligible only when including more than 50 participants.

DISCUSSION

Main results. When using the data of all participants, the random forest correctly predicted the transportation mode in 90.0% of the trips. This is comparable with the prediction rates found in studies that made predictions at the time-unit level. Ellis et al. (17) have reported prediction rates of 89.8% and 91.9% (depending on the method) when using random forests to predict five different modes for units of 1 min on the basis of GPS and accelerometer data. Using 1-s units,

TABLE 2. Error rates (%) of models considering only a subset of the explanatory variables.

	Accelerometer	**GPS**	**GPS/GIS**	**Accelerometer + GPS**	**Accelerometer + GPS/GIS**
Overall	17.7	17.6	11.6	12.1	10.6
Walking	12.3	7.3	5.6	4.9	4.9
Bicycle	66.2	52.1	51.2	57.3	49.3
Motorized	15.1	14.4	11.3	12.8	10.6
Public	31.0	49.4	21.9	22.4	19.5

No. of trees in each forest = 1000; n = 7329.

TABLE 3. Error rates (%) of the predictions from models based on a limited number of participants.

	Overall	**Walking**	**Bicycle**	**Private Motorized**	**Public**	***n***
First 5	28.0	7.4	100.0	16.1	95.5	143
First 10	17.0	5.8	85.3	21.6	21.2	298
First 20	15.2	4.6	95.9	14.4	28.4	630
First 30	13.9	4.1	79.5	14.1	26.2	991
First 40	13.7	4.5	81.4	12.6	26.5	1340
First 50	13.0	5.0	57.7	13.0	24.9	1639
First 100	13.6	5.0	64.1	14.3	24.7	3280
First 150	12.9	4.5	64.6	10.8	27.2	4757
First 200	10.2	4.7	61.7	7.6	15.4	6261

No. of trees in each forest = 1000; n = number of trips; total number of trips = 7329.

Feng and Timmermans (18) have found a prediction rate of approximately 90% for eight modes using a Bayesian Belief Network Model using GPS, accelerometer, and survey data.

Few studies addressing mode prediction at the trip or trip-stage level have been reported. Gong et al. (19) and Chen et al. (14) have yielded prediction rates of 79.1% and 82.6%, respectively, in two New York–based studies using a step-by-step algorithm. Other work attempting to predict modes at the level of trip stages (unimodal components of trips) have used more complex strategies. For example, Kohla et al. (21) have used time unit-level detection of walking stages within trips to further segment the trips into trip stages. The nonwalking modes were then identified. Multinomial logistic regression yielded a prediction rate of 80%.

The prediction rates found in the present study are within the range of those reported in the aforementioned studies, which is promising for future applications of the method. However, it is difficult to compare the performance of our algorithm with those of previous studies. Most of these models relied on relatively small convenience samples or scripted/controlled travel behavior data collections (in which participants are asked to follow a specific itinerary with a specific mode). Models based on controlled data to predict activity modes are less generalizable and less apt to predict real-life data (2,16); the same may be expected for the prediction of transportation modes. In contrast, small convenience samples might lack some variety, and they do not represent the relative importance of the different categories well. Because the size of the categories influences the overall prediction rate, these overall rates are not easily comparable between studies. More studies are required to compare the different prediction methods in the same context, with the same quality of data and the same choice of categories (21).

Importantly, we found that the method differentiated between public and private motorized transport well. Additional analysis of these two categories only (not reported) indicated that the highest predictive variables were "possessing a car" (survey), "proportion of valid GPS observations among all possible (including observed and missing) observations," and "possessing a public transport pass." The findings suggest that these indicators that are not always included in published models may be of particular interest. However, it must be kept in mind that the public transport system is particularly well served in Paris and that these variables may have a different predictive contribution in other settings.

Testing trees grown on the data of various numbers of participants enabled us to evaluate the predictive performance of the algorithm for data collected later (i.e., to understand from how many participants detailed mode data should be collected to make reliable predictions using less detailed data). When using no more than 30 participants, the overall prediction rate for the remaining 206 participants was 86.1%. This observation shows that data on a relatively small number of participants can provide valuable information on a much larger data set. However, prediction models based on less than 30 participants displayed poor performance for the mode categories with the fewest trips. To limit the number of participants required to grow the random forest, oversampling the categories with the fewest trips or participants could be considered.

The approach of collecting limited data from the context in which one is willing to make predictions to build a prediction model contrasts with pretrained models (i.e., prediction models trained on data from a different context). Pretrained models are considerably less expensive because no preliminary data collection is required in each particular context. However, pretrained models are less well adapted to the specific context of interest. It can be expected that the optimal set of variables and thresholds of variables used to differentiate transportation modes vary between different contexts. Further studies are required to compare pretrained and same-context prediction models and then determine whether the extra effort of preliminary data collection yields a significant improvement in prediction quality.

In previous studies, it has been argued that accelerometer data may enhance the predictive power of a model for transportation modes, especially concerning trips with frequent missing GPS data values (e.g., during subway use) (18). We found relatively good prediction rates from an accelerometer data-only model. However, we noted only a small increase in prediction rates when including accelerometer data in the GPS/GIS model, which may be attributable to our study design in which observation units represented trips rather than time units, as applied in most previous studies. The indicators associated with GPS data (proportion of invalid or missing data, dispersion of speed throughout the trip) are possibly more informative in the trip-level models than in a time unit-level approach, thus rendering it less useful to also consider accelerometer data.

Strengths and limitations. The algorithm of imputation of transportation modes developed in our study was relatively accurate, using a combination of GPS and GIS data processed with algorithms, travel diaries, and a phone prompted recall survey. However, the preparation of the survey and difficulties to contact some of the participants by phone proved to be a bottleneck in the data collection process, thereby causing delays between the device data collection and the survey for a median period of 17 d. This delay very likely led to memory bias in identification of activity places and transportation modes, despite the information available to prompt participant recall. Our prediction method proved to be convenient to implement and reliable compared with the results of previous studies. This prediction method can be easily adapted to a different study context, and the explanatory variables used to grow the random forest can be selected depending on the available information. In our approach, the prediction model was accurate because it was constructed on data obtained from the same population for which the predictions were made. To obtain this context specificity and the ability to select the set of locally available variables, one must conduct a preliminary data collection to adequately train the model. The duration

of this learning phase depends on the complexity of the prediction (i.e., the number of categories of the outcome and especially the number of observations in the smallest categories). Importantly, our work demonstrated that a fairly short-term learning phase is sufficient for adequate predictions. When adapting this methodology, data collection for a limited number of participants could include techniques such as a system of survey of modes and activity locations (if not too burdensome for the participants) or the SenseCam methodology. The extra burden on the participants during this learning phase could be compensated for by reducing the amount of data collected in other parts of the data collection process or reducing the number of observation days per participant.

Compared with a time unit-level prediction method, predictions at the trip level provide less detailed information. However, as trip-level data are useful in transportation sciences and behavioral sciences, a trip-level prediction method has some interesting advantages over time unit-level prediction methods. First, information on the entire trip can be used to derive predictors, such as quantification of the intratrip variability in GPS and accelerometer indicators (e.g., speed or acceleration) and summaries of the GPS data quality. Second, a trip-level method is more parsimonious in the number of predictions made. Because only one prediction per trip is required, the method allows for more participants and more observation days per participant in the model. In this RECORD GPS study, 7329 trips were observed for 236 participants and 1647 observation days. Given 12 $h \cdot d^{-1}$ of observations, a 5-s window approach would yield more than 14 million predictions, while a 1-min window approach would yield nearly 1.2 million predictions. Modeling this number of predictions would require a very high computational time. For large-scale studies with 1000 participants or more, time unit predictions would therefore require high performance computing. Finally, time unit-level models also model the data at activity places and must include activity mode categories in the model, which may reduce the quality of the overall prediction. In conclusion, we do not argue that a trip-level method is better than a unit-level method, although it does provide researchers with a valid alternative to address a large number of research questions.

A clear limit of our proposed mode detection algorithm is that its application requires data segmented into trips because the present algorithm was intended to be a complement of another trip segmentation algorithm that we commonly use in our studies (33). Moreover, it should be emphasized that the use of an algorithm of mode detection at the time unit level (e.g., min) would also require the application of a second algorithm to derive coherent information on the mode(s) used at the trip level.

This method is inappropriate for trips with multiple transportation modes. In our study, we observed 1.3% of multimodal trips (comprising more than one nonwalking mode), and we excluded them to train the prediction model. The model predicted one of the two modes for 99% of these trips. Depending on the application of this method and the proportion of multimodal trips in the study area, this limitation may be problematic and may provide an argument for the use of more advanced prediction methods that segment trips into trip stages and impute the corresponding modes (38).

Finally, it should be kept in mind that any mode prediction algorithm will have a certain error rate. In specific circumstances, researchers may want to collect more accurate data on modes for each trip. Although using SenseCam in addition to GPS receivers is useful to obtain an accurate criterion for validating algorithms, we argue that wearable cameras are too intrusive and the corresponding data are too burdensome to process to permit data collection across a large sample size. In that case, combining GPS data collection with the use of a GPS-based prompted recall mobility survey may represent a feasible option to derive accurate trip-level data.

CONCLUSIONS

This study is one of the first to use real-life data from a relatively large and diverse sample to test a prediction method for transportation modes. The approach uses a trip-level model, thereby rendering the application more convenient for subsequent application in a variety of transportation or behavioral study designs. This method could improve future data collection processes by decreasing the workload for both participants and researchers and providing relevant data to investigate the relation between transportation and health.

The authors thank the following partners from the funding institutions: Pierre Arwidson, Nadine Asconchilo, Annette Gogneau, Colette Watellier, Yasmina Baaba, Mélani Alberto, Christelle Paulo, Anne-Eole Meret-Conti, Cédric Aubouin, Benoît Kiéné, Hélène Pierre, Sophie Mazoué, John Séraphin, and Ivan Derré.

The RECORD GPS study was supported by the INPES (National Institute for Prevention and Health Education), the Ministry of Ecology (DGITM), CERTU (Centre for the Study of Networks, Transport, Urbanism, and Public constructions), ARS (Health Regional Agency) of Ile-de-France, STIF (Ile-de-France Transport Authority), the Île-de-France Regional Council, RATP (Paris Public Transport Operator), and DRIEA (Regional and Interdepartmental Direction for Equipment and Planning).

The provision of financial support does not, in any way, infer or imply endorsement of the research findings by any agency.

The authors declare no conflict of interest.

The results of the present study do not constitute endorsement by the American College of Sports Medicine.

REFERENCES

1. Auld J, Williams CA, Mohammadian AK, Nelson PC. An automated GPS-based prompted recall survey with learning algorithms. *J Transport Lett*. 2009;1(1):59–79.
2. Bastian T, Maire A, Dugas J, et al. Automatic identification of physical activity types and sedentary behaviors from triaxial accelerometer: laboratory-based calibrations are not enough. *J Appl Physiol (1985)*. 2015;118(6):716–22.
3. 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. *Transport Res C Emer*. 2009;17(3):285–97.

4. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
5. Brondeel R, Weill A, Thomas F, Chaix B. Use of healthcare services in the residence and workplace neighbourhood: the effect of spatial accessibility to healthcare services. *Health Place*. 2014; 30:127–33.
6. Carlson JA, Jankowska MM, Meseck K, et al. Validity of PALMS GPS scoring of active and passive travel compared with SenseCam. *Med Sci Sports Exerc*. 2015;47(3):662–7.
7. Chaix B, Bean K, Daniel M, et al. Associations of supermarket characteristics with weight status and body fat: a multilevel analysis of individuals within supermarkets (RECORD study). *PLoS One*. 2012;7(3):e32908.
8. Chaix B, Kestens Y, Bean K, et al. Cohort profile: residential and non-residential environments, individual activity spaces and cardiovascular risk factors and diseases—the RECORD Cohort study. *Int J Epidemiol*. 2012;41(5):1283–92.
9. Chaix B, Kestens Y, Duncan S, et al. Active transportation and public transportation use to achieve physical activity recommendations? A combined GPS, accelerometer, and mobility survey study. *Int J Behav Nutr Phys Act*. 2014;11(1):124.
10. Chaix B, Kestens Y, Perchoux C, Karusisi N, Merlo J, Labadi K. An interactive mapping tool to assess individual mobility patterns in neighborhood studies. *Am J Prev Med*. 2012;43(4):440–50.
11. Chaix B, Meline J, Duncan S, et al. Neighborhood environments, mobility, and health: towards a new generation of studies in environmental health research. *Rev Epidemiol Sante Publique*. 2013; 61(3 Suppl):S139–45.
12. Chaix B, Meline J, Duncan S, et al. GPS tracking in neighborhood and health studies: s step forward for environmental exposure assessment, a step backward for causal inference? *Health Place*. 2013;21:46–51.
13. Chaix B, Simon C, Charreire H, et al. The environmental correlates of overall and neighborhood based recreational walking (a cross-sectional analysis of the RECORD study). *Int J Behav Nutr Phys Act*. 2014;11(1):20.
14. Chen C, Gong H, Lawson C, Bialostozky E. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: lessons learned from the New York City case study. *Transport Res A Pol*. 2010;44(10):830–40.
15. de Nazelle A, Nieuwenhuijsen MJ, Anto JM, et al. Improving health through policies that promote active travel: a review of evidence to support integrated health impact assessment. *Environ Int*. 2011;37(4):766–77.
16. Ellis K, Godbole S, Chen J, Marshall S, Lanckriet G, Kerr J. Physical activity recognition in free-living from body-worn sensors. In: *Proceedings of the 4th International SenseCam and Pervasive Imaging Conference*; 18–19 November, 2013; San Diego (CA): ACM; 2013. pp. 88–9.
17. Ellis K, Godbole S, Marshall S, Lanckriet G, Staudenmayer J, Kerr J. Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms. *Front Public Health*. 2014;2:36.
18. Feng T, Timmermans HJ. Transportation mode recognition using GPS and accelerometer data. *Transport Res C Emer*. 2013; 37:118–30.
19. Gong H, Chen C, Bialostozky E, Lawson CT. A GPS/GIS method for travel mode detection in New York City. *Comput Environ Urban*. 2012;36(2):131–9.
20. Gong L, Morikawa T, Yamamoto T, Sato H. Deriving personal trip data from GPS data: a literature review on the existing methodologies. *Procd Soc Behv*. 2014;138:557–65.
21. Kohla B, Meschik M, Gerike R, Sammer G, Hössinger R, Unbehaun W. A new algorithm for mode detection in travel surveys: mobile technologies for activity—travel data collection and analysis. In: Rasouli S, Timmermans HJP, editors. *Mobile Technologies for Activity-Travel Data Collection and Analysis*. Hershey (PA): IGI Global; 2014. pp. 134–51.
22. Kozey-Keadle S, Libertine A, Lyden K, Staudenmayer J, Freedson PS. Validation of wearable monitors for assessing sedentary behavior. *Med Sci Sports Exerc*. 2011;43(8):1561–7.
23. Leal C, Bean K, Thomas F, Chaix B. Multicollinearity in the associations between multiple environmental features and body weight and abdominal fat: using matching techniques to assess whether the associations are separable. *Am J Epidemiol*. 2012;175(11):1152–62.
24. Liaw A, Wiener M. Classification and regression by random forest. *R News*. 2002;2(3):18–22.
25. Lin WJ, Chen JJ. Class-imbalanced classifiers for high-dimensional data. *Brief Bioinform*. 2013;14(1):13–26.
26. Perchoux C, Kestens Y, Thomas F, Van Hulst A, Thierry B, Chaix B. Assessing patterns of spatial behavior in health studies: their socio-demographic determinants and associations with transportation modes (the RECORD Cohort study). *Soc Sci Med*. 2014;119:64–73.
27. Rissel C, Curac N, Greenaway M, Bauman A. Physical activity associated with public transport use—a review and modelling of potential benefits. *Int J Environ Res Public Health*. 2012;9(7):2454–78.
28. Sahlqvist S, Song Y, Ogilvie D. Is active travel associated with greater physical activity? The contribution of commuting and non-commuting active travel to total physical activity in adults. *Prev Med*. 2012;55(3):206–11.
29. Sasaki JE, John D, Freedson PS. Validation and comparison of ActiGraph activity monitors. *J Sci Med Sport*. 2011;14(5):411–6.
30. Shen L, Stopher PR. Using SenseCam to pursue "ground truth" for global positioning system travel surveys. *Transport Res C Emer*. 2014;42:76–81.
31. Southward EF, Page AS, Wheeler BW, Cooper AR. Contribution of the school journey to daily physical activity in children aged 11–12 years. *Am J Prev Med*. 2012;43(2):201–4.
32. Stopher PR, Collins A, editors. Conducting a GPS prompted recall survey over the internet. In: *The 84th Annual Meeting of the Transportation Research Board*; 9–13 January, 2005; Washington (DC).
33. Thierry B, Chaix B, Kestens Y. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. *Int J Health Geogr*. 2013;12(14).
34. Van Hulst A, Thomas F, Barnett TA, et al. A typology of neighborhoods and blood pressure in the RECORD Cohort study. *J Hypertens*. 2012;30(7):1336–46.
35. Velaga NR, Quddus MA, Bristow AL. Developing an enhanced weight-based topological map-matching algorithm for intelligent transport systems. *Transport Res C Emer*. 2009;17(6):672–83.
36. Wanner M, Gotschi T, Martin-Diener E, Kahlmeier S, Martin BW. Active transport, physical activity, and body weight in adults: a systematic review. *Am J Prev Med*. 2012;42(5):493–502.
37. Wanner M, Martin BW, Meier F, Probst-Hensch N, Kriemler S. Effects of filter choice in GT3X accelerometer assessments of free-living activity. *Med Sci Sport Exerc*. 2013;45(1):170–7.
38. Wolf J, Oliveira M, Thompson M. Impact of underreporting on mileage and travel time estimates: Results from global positioning system-enhanced household travel survey. *Transp Res Record*. 1854;2003:188–98.
39. Xu H, Wen LM, Rissel C. The relationships between active transport to work or school and cardiovascular health or body weight: a systematic review. *Asia Pac J Public Health*. 2013;25(4):298–315.