A novel methodology for identifying environmental exposures using GPS data

Andreea Cetateanu[1], Bogdan Luca[3], Andrei Alin Popescu[3], Angie Page[4], Ashley Cooper[4], Andy Jones[2]

[1] School of Environmental Sciences, University of East Anglia, Norwich, Norfolk, NR4 7TJ, UK.
[2] Norwich Medical School, University of East Anglia, Norwich, Norfolk, NR4 7TJ, UK.
[3] School of Computing Sciences, University of East Anglia, Norwich, Norfolk, NR4 7TJ, UK.
[4] Centre for Exercise, Nutrition and Health Sciences, School for Policy Studies, University of Bristol, 8 Priory Road, Bristol BS8 1TZ.

Correspondence to: Andreea Cetateanu, a.cetateanu@gmail.com

Tel: 00 44 (0)2075942637

Contact address: School of Public Health, Imperial College London, St Mary's Campus, Medical School Building, Office G39, Norfolk Place, W2 1PG, London

Abstract

Aim: While studies using GPS (Global Positioning Systems) have the potential to refine measures of exposure to the neighbourhood environment in health research, one limitation is that they do not typically identify time spent undertaking journeys in motorised vehicles when contact with the environment is reduced. This paper presents and test a novel methodology to explore the impact of this.

Methods: Using a case study of exposure assessment to food environments, an unsupervised computational algorithm is employed in order to infer two travel modes: motorised and non-motorised, on the basis of which trips were extracted. Additional criteria are imposed in order to improve robustness of the algorithm.

Results: After removing noise in the GPS data and motorised vehicle journeys, 82.43% of the initial GPS points remained. After comparing a sub-sample of trips classified visually of motorised, non-motorised and mixed mode trips with the algorithm classifications, it was found that there was an agreement of 88%. The measures of exposure to the food environment calculated before and after algorithm classification were strongly correlated.

Conclusion: Identifying non-motorised exposures to the food environment makes little difference to exposure estimates in urban children but might be important for adults or rural populations who spend more time in motorised vehicles.

Keywords: global positioning systems, food environments, travel mode, unsupervised algorithm

A novel methodology for identifying environmental exposures using GPS data

1. Introduction

A recent criticism of many neighbourhood and health studies has been that they have not adequately taken into account the actual exposures to the environment that individuals experience in their daily activity patterns (Kestens et al., 2010). Rather, they tend to assume exposures based on home and sometimes school or work locations. There are also studies that infer exposures from travel surveys or diaries, but these provide subjective declarative data based on participants' recall of where they visited (Chaix et al., 2012), and it has been reported that trip underreporting occurs (Bricka et al., 2012; Stopher et al., 2007; Wolf et al., 2003b). There is also a third type of study that uses passive tracking of study participants, which yields objective data. To this end GPS (Global Positioning Systems) are increasingly being used to measure daily activity spaces and investigate behaviours that relate more closely to health outcomes of interest (Kerr et al., 2011).

GPS is a satellite-based global navigation system that provides an accurate location of any point on the Earth's surface (Krenn et al., 2011). It thus provides a means to objectively assess the spatial location of individuals in the environment or people's behaviours while moving in the environment. Outdoor GPS relies on being able to receive a signal from four or more satellites in order to triangulate a person's position, and a GPS data point will typically consist of a time stamp and longitude, latitude and altitude coordinates. This daily mobility is of particular interest in environment–health research, as both a potential source of transportation-related physical activity and as a measure of exposure to certain geographic environments (Chaix et al., 2012), such as food environments. However, such multi-place measures must be carefully constructed in order to make sure true exposures of interest are assessed.

Whilst logging travel patterns using GPS measurements has become commonplace, managing the considerable volume of GPS data collected and extracting meaningful outcome values is difficult. GPS technologies are still developing, with associated different qualities of GPS software and hardware, and even if the device is working at peak performance there will always be some spatial error in the accuracy of location recording (Kerr et al., 2011), which differs based on conditions and type of GPS receiver used. Location errors can emerge from factors such as satellite propagation delays or precision of the device, and signal loss due to slow location detection (initialization and start-up, whereby the GPS receiver needs some time to first acquire signals from satellites) or ground cover such as trees.

Additional to technical or usability issues, other issues that arise with GPS data are related to how it is interpreted when extracting environmental exposures of interest. For example, in studies investigating exposures to the retail food environment and linking them to health-related outcomes, researchers may be interested only in GPS points that represent on-foot or slow cycling trips, as people within moving vehicles would have a lesser opportunity to access food outlets to purchase food without the vehicle stopping and them getting out. This consideration has typically been ignored in the literature, in part because of some of the problems inherent in identifying the travel modes of study participants. For example, GPS points that in reality represent a car slowing down at intersections, traffic calming measures or due to the presence of other traffic may be wrongly interpreted as walking because they register low speeds. Those studies that have attempted to make such differentiations typically use either crude criteria (such as identifying walking as GPS points under a certain speed threshold) (Wheeler et al., 2010), or they clean GPS data manually (Harrison et al., 2014), which can be very time consuming.

To date a small number of researchers have attempted to produce more robust algorithms for cleaning GPS data and extracting useful information such as travel mode from it (Auld et al., 2009; Carlson et al., 2015; Chao et al., 2010; Feng and Timmermans, 2013; Lin et al., 2013; Schuessler and Axhausen, 2009; Zheng et al., 2008). Whilst there is no uniform standard across disciplines, most methods have several commonalities among them. They typically each attempt to split the raw GPS data into smaller relevant segments (i.e. journeys or trips) on which further analysis is carried out (e.g. determining transport mode for each trip). Usually some form of pre-processing is carried out to remove outliers and de-noise the data, after which a main algorithm is applied for analysis, and subsequently post-processing is used to further improve classification accuracy. These main algorithms can be classified into machine learning approaches and criteria-based approaches. In turn, machine learning approaches can be divided into supervised and unsupervised methods.

Criteria based methods are based on expert chosen rules (e.g. speeds below a certain threshold are considered walking) to analyse trips. These are the simplest approaches and have been successfully used in various papers (Cho et al., 2011; Chung and Shalaby, 2005), but they are usually biased by the expert's expectations and experience and do not perform well on datasets on datasets other than those which they have been developed.

Supervised methods (Chao et al., 2010; Feng and Timmermans, 2013; Zheng et al., 2008) rely on manually classified data in order to make inferences about unknown data. In such cases, supervised classifier models such as decision trees are trained using the features (e.g. average speed, maximum speed, acceleration etc.) extracted from the data and the known class labels. The new data is then classified using the trained model. A particular drawback of such methods is the requirement for training data, which is usually obtained by manual classification and can hence be time consuming and costly to generate. A further limitation is that models trained on one dataset may perform poorly when applied to a different one.

Unsupervised methods overcome this disadvantage by not relying on training data for predictions. They rather infer transportation modes based on the structure and the characteristics of the input data, in some cases aided by expert-defined rules, e.g.(Schuessler and Axhausen, 2009). For example the work of Lin, et al. (2013) assumes that each transport mode generates speeds from a certain distribution. They use raw GPS data to estimate the parameters of these distributions and conduct statistical tests to determine the differences between these distributions across different segments. Based on these inferred differences, they then use hierarchical clustering to group trips into major groups which correspond to transport modes. Unreliable trips are classified based on proximity to relevant locations such as bus stops. Most of these methods are data intensive and require additional information, such as relevant landmark positions, and would not work as well for studies that do not have such information available.

The method presented here (which will be called Trans-Mod) falls in the category of unsupervised methods and is applied on the PEACH (Personal and Environmental Associations with Children's Health) dataset containing the GPS locations of a sample of children in Bristol. The development and testing of the methodology presented in this paper arose from the need to extract only trips not in a motorised vehicle from the PEACH dataset in order to be able to estimate exposure to the food environment and calculated associations with health outcomes such as diet and weight status (results not presented here). They key requirement was to identify times when children were inside a vehicle and those when they were not, as it is assumed that the ability of children to access food outlets will be limited when they are in a vehicle. A model known as a Hidden Markov Model (HMM) (Murphy, 2012) was used to model the differences in speeds from raw GPS data generated by two travel modes: non-motorised (walking or slow cycling) and in a motorised vehicle. HMMs have been previously used (Reddy et al., 2010) to determine travel modes using the information provided by mobile phones (accelerometer and GPS data). However, the method presented here has very low input data requirements, namely just the registered timestamp of each GPS point and the distance between two consecutive points, on the basis of which speed can be easily calculated. The present paper investigates how accurately the method presented here differentiates between motorised and non-motorised travel modes, and if the post-processing exposure estimates of exposure to the food environment differ to those before processing.

2. Methods

2.1.Dataset

The dataset used in developing the model presented here was obtained from PEACH, a study undertaken in Bristol, UK, which investigates how the environment can influence physical activity and dietary behaviours in children. Characteristics of the PEACH study sample have been described in more detail elsewhere (Lachowycz et al., 2012; Wheeler et al., 2010). In brief, this dataset provides up to 7 days of GPS data recorded in the morning (8am-9am), evening (3pm-10pm), and at weekends (8am-10pm). In total, 688 children in their first year of secondary school wore a Garmin Foretrex 201 GPS receiver recording data at 10-s intervals (epochs). The GPS has limited battery life, and participants were asked to switch the GPS on at the end of school, and off at bedtime. Research staff charged the units after the first two days of use.

GPS data from this study was used to measure personal exposure the food environment. Measures of the food environment exposure were computed in a Geographical Information System (GIS) (ArcGIS 10.0 (ESRI Inc, Redlands, CA, USA)) using the UK Ordnance Survey Points of Interest (PoI) dataset (OrdnanceSurvey, 2011), a dataset that includes the precise location of 21 categories of food outlets. The location of all food outlets in the Points of Interest data were mapped and grouped into three categories, based on evidence in the literature (Cetateanu and Jones, 2014; Gustafson et al., 2012; Liese et al., 2007), as well as fieldwork visits made by the authors to a sample of outlets falling within each category. The categories chosen were ‘food outlets where people can purchase healthy food’ which was computed to include markets, grocers, organic stores, supermarket chains and independent supermarkets; ‘food outlets where people can purchase unhealthy food’ including bakeries, delicatessens, confectioners, convenience stores and newsagents; and ‘food outlets where people can purchase fast food’ (fast food outlets, takeaways, fast food delivery services that also have an eat in option, and fish and chip shops).

The exposures were calculated as the percentage of the measurement period time spent outdoors in the vicinity (for the purposes of this study we choose within 50 meters) of different retail food outlet types, merged into three categories: time spent near healthy food outlets, time spent near unhealthy food outlets and time spent near fast food outlets. For the purposes of analysis, patterns of exposure during all the time periods (morning, evening, weekend) measured in PEACH were combined. This was done because the amount of time spent in the vicinity of food outlets was generally small, particularly before school. The denominator for these percentages was the total period (1 hour in the morning, 7 hours in the evening, 14 hours in the weekend) rather than the period for which a location was recorded in the GPS as the devices used did not operate within a building. In order to better measure environmental exposures to food, the aim of this paper was to identify for later removal any points that might represent time spent in a motorised vehicle, or spurious GPS points due to influences like poor satellite signal. The model used to do this is graphically represented below in Figure 1.

Figure 1. Flow diagram of steps **[near here]**

2.2.Trip and travel mode detection, data cleaning and smoothing

Stage 1: Pre-processing

In the first instance several criteria were developed to mark points for later removal that would not represent true exposures. These included GPS drift (i.e. GPS records which suggest that a child has moved an implausible amount in a short space of time, meaning there has been some inaccuracy in the GPS locations, often as the signal was obstructed by buildings or trees), as well as short participant reads (i.e. participants registering a very low number of GPS points overall, which typically represented poor device wear compliance or problems with the GPS signal). The criteria developed are as follows:

1. Marking outliers: for each participant, select the list of points that are further than 500m from any other GPS points belonging to them.

2. Marking aberrant speed: all points having more than 100 kph.

3. Marking short participant reads: all participants with less than 1 minute total GPS wear time.

Stage 2: Processing

For each participant, the points were ordered according to their timestamp and the obtained series of GPS points were subsequently divided into trips. A trip was considered to be a number of consecutive points for which the time difference between every two consecutive points was less than 5 minutes. If the time difference between two consecutive points in time was greater than 5 minutes, this was set to mark the beginning of a new trip. The rationale behind this is considered in the Discussion section.

We represent a trip as a sequence of speeds and we want to infer the travel modes that generated those speeds. We expect the non-motorised travel mode to give rise to speeds that are on average lower than the motorised mode. It is of course possible that several transportation modes have been used during one trip. Such a trip will be referred to a as a *mixed* trip (i.e., it includes both motorised and non-motorised modes).

To model this behaviour we created a HMM model with two hidden states corresponding to non-motorised and motorised states respectively. Each state has its own Gaussian distribution of speeds that represent the emission probabilities of the model. The transition probabilities between the states reflect the likelihood of changing the travel mode.

The model was tuned on 50 randomly chosen trips using a version of the Expectation-Maximisation algorithm (Moon, 1996), known as the Baum-Welch algorithm (Welch, 2003). This algorithm starts with some random values for the model parameters (transition, emission and initial probabilities) and gradually updates them until they converge, without using any other piece of information than the input sequence of speeds. Full details of this algorithm are given elsewhere (Murphy, 2012).

For each trip, using the tuned model, a Viterbi algorithm (Viterbi, 1967) is able to identify the most likely combination of travel modes that generated the observed sequence of speeds. Unlike fixed threshold-based approaches, the classification of points into motorised/non-motorised travel modes is dynamic. The algorithm makes the decision by computing the likelihood of the speed being generated from either of the two modes, taking into account also the most likely modes of the points around it.

Stage 3: Post-processing

Some post-processing steps were employed in order to correct some issues which can appear on a small subset of the data. Such methods are readily integrated in the program and do not require additional user interaction. In the first step, short segments (for which the overall duration is less than 1 minute in total GPS time) were marked separately with the purpose of later being eliminated from the raw GPS data. This was based on the assumption that it is very unlikely that such short segments would represent actual *non-motorised* trips. A limitation could be that some very short trips which may actually be access and egress trips are eliminated, although for this analysis we visually checked all these short segments and identified them as spurious. Furthermore, instances can be observed whereby there is an outlier (isolated point) adjacent to two points that have been classified of a different state in a trip. It was considered that a change of transportation mode that spans only one point is very unlikely. This was thus corrected by changing the state of the outlier to the state of its neighbours.

To address situations where the wearer was in a vehicle that was slowing down, an additional criteria was developed whereby if *non-motorised* trips spanned less than 2 minutes and were surrounded by *vehicle* points, these were marked as *motorised vehicle* points. Furthermore, there were instances where within a trip some points were classified as *motorised* and some as *non-motorised*, but the *motorised* points represented a very small proportion of the whole trip, which was mostly dominated by *non-motorised* points. An additional criterion was therefore imposed whereby if less than 5% or less than 5 of the points in a trip were classified as *motorised* and the rest were *non-motorised*, all the points in that trip were considered as *non-motorised mode*.

After processing, there were still some points over 15 kph classified by the model as *non-motorised mode*. This was because the speeds were not high enough for the model to suggest them as motorised vehicle points given their surrounding points were mostly non-vehicle. An additional criterion was therefore imposed by marking all of these points as *motorised mode*. This was based on previous practice in studies that have used the same dataset (Lachowycz et al., 2012; Wheeler et al., 2010), where travel speeds above 15kph were judged to be journeys in vehicles. Nevertheless, a limitation of this is that some instances of fast cycling may be classified as motorised mode.

The PEACH dataset does not contain any annotation data regarding the travel modes of the participants. Thus, in order to estimate the accuracy of our method, a sub-sample of 99 randomly selected trips (33 motorised mode, 33 non-motorised mode and 33 trips containing both motorised and non-motorised mode, termed here as mixed) were labelled by researchers (the first and the last authors) by overlaying the trips on a base map in ArcGIS and taking into account the several criteria such as the size of the roads the participant used, and the speed of GPS points. Cohen's kappa test for 2-way inter-rater agreement (k) was run to determine the level of agreement between the first and last author on the classification of trips as 'motorised mode', 'non-motorised mode' or 'mixed mode', as well as between the algorithm and the first, and last author respectively..

In order to determine the potential impact of trip classification on measures of environmental exposure, the similarity of the exposure measures to the food environment calculated on the raw GPS data versus the cleaned GPS data was investigated using Pearson's correlation coefficients.

The algorithm was implemented in Python 2.7. For the Hidden Markov Model the implementation from the Sklearn 0.31.1 package was used. All other statistical analysis was undertaken in SPSS (version 21, IBM Corp, Armonk, NY, USA).

3. Results

Before any processing there were 366432 GPS points in the PEACH dataset that was used to train the HMM model, which represented a total of 4018 trips (or segments). Out of these, 2488 were classified as non-motorised only trips, 443 were motorised and the rest were mixed trips (including both motorised vehicle and non-motorised points).

The Baum-Welch algorithm converged to the parameters illustrated in Figure 2. It can be observed that the emission distribution corresponding to a *non-vehicle* state is centred around 2.14 kph, while for the *vehicle* state it is centred around 26.86 kph. These values are consistent with the initial assumption that the speeds should be able to differentiate well between the two travel modes.

In terms of transition probabilities, the probability of moving from non-vehicle to vehicle was 0.0232 and the probability of moving from a vehicle to non-vehicle state was 0.1223. These low values reflect the fact that the likelihood of two consecutive points corresponding to different travel modes is much lower than that of them being the same. The probability of remaining in the *non-vehicle* state is about 10% percent higher than the probability of remaining in the *vehicle* state. This is explained by the fact that the data is highly right skewed (skewness= 3.401), thus increasing the probability that if in a *non-vehicle* state, one remains in that state.

Out of the 366432 GPS points in the PEACH dataset used to train the HMM model, 64385 were marked for removal during the pre-processing, processing and post-processing stages. This meant that 17.57 % of the original GPS points were marked for removal, which represented: 0.37% (n= 1347) outliers, 0.08% (n= 282) aberrant speed, 0.006% (n= 21) participants with less than 1 minute worth of GPS data, 15.94% (n= 58409) motorised vehicle points, 0.30% (n= 1087) points representing trips below one minute total duration, and 0.88% (n= 3239) points registering speeds over 15 kph. As a result, 302047 GPS points (82.43%) remained representing non-vehicle points.

Figure 2. The HMM model after training. The purple vertices represent the states of the model, the numbers on arrow from state *u* to state *v* represent the transition probability from the state *u* to the state *v* and the distributions in the yellow rectangles represent the emission probabilities. **[near here]**

In order to visually represent results from the model, plots were generated to represent all 4018 pairs of trips before and after post-processing. Figures 3, 4 and 5 represent three such examples, whereby the left-hand side graph represents the classification of GPS points during the processing stage, and the right hand side graph represents the classification of points at the post-processing stage. In Figure 3, which represents one trip, the algorithm classifies some points as *non-motorised*, and others as *motorised* at the processing stage. Some points are considered as *non-motorised* because when a car slows down, the speeds are considered by the model as too low to be *motorised vehicle* points. However, the number of consecutive points marked as *non-motorised* spanned less than 2 minutes and were surrounded by *motorised vehicle* points. Therefore, these were changed to *motorised vehicle* points in the post-processing stage of the model. Therefore, we built our model such that it's inherent statistical framework determines that it is more likely for a motorised vehicle (for example a car) to have slowed down for a few seconds than for a person to get out while being in the car for such a short time.

In the example of Figure 4 the *motorised vehicle* points represented only 5 points of the whole trip, which was mostly dominated by *non-motorised* points. These points are therefore marked as *non-motorised vehicle* at the post-processing stage. In Figure 5, less than 5% of GPS points in the trip are *motorised vehicl*e, and therefore at post-processing these are marked as *non-motorised vehicle*; however, some of these points register speeds of over 15 kph, because the speeds were not high enough for the model to suggest them as *motorised* points given their surrounding points were mostly *non-motorised* . Therefore, these are marked for later removal (i.e.: non-motorised mode> 15 kph). Figure 6 illustrates an example of the total GPS trips (synthesised to preserve anonymity) of one hypothetical participant in one day, after processing.

Figure 3. Example of a trip during and after processing **[*near here]***

Figure 4. Example of a trip during and after processing ***[near here]***

Figure 5. Example of a trip during and after processing ***[near here]***

Figure 6. Map showing a participant's trip in a day after classification (© Crown Copyright/database right 2015. An Ordnance Survey/EDINA supplied service) ***[near here]***

The level of agreement between the algorithm and the annotation by the first and last author was tested with Cohen's kappa (k) on the sub-sample of 99 trips, and it was found that there was strong agreement between the first and last author, as well as between both authors and the algorithm ($k>0.8$, $p<0.001$). The first author and the algorithm agreed on the classification of 88% of the trips, the last author and the algorithm on 87%, and the first and last author on 89%. Agreement was poorer when trips were classified as mixed by the algorithm, although this was based on only 10 trips, while the first and last author classified differently to the algorithm on just 5 motorised trips and 2 non-motorised trips.

When comparing the absolute differences in measures of exposure to the food environment before and after processing (Table 1), it can be observed that the exposure measures calculated on the raw GPS data were unsurprisingly statistically significantly higher than the post-processing values. However, the correlation coefficient of the pre and post processing exposure measures was of 0.98 or above for each of the three food outlet types examined ($p<0.001$). This shows that children who had high levels of exposure before processing also had high levels of exposure after processing. Therefore the processing led to lower levels of estimated absolute exposure but did not substantially modify the ordering of children.

Table 1. Comparison of before with after processing exposures **[near here]**

4. Discussion

Complex methods for analysing GPS data exist (Byon et al., 2007; Gonzalez et al., 2008; Moiseeva and Timmermans, 2010; Patterson et al., 2003; Reddy et al., 2010; Tsui and Shalaby, 2006; Zhang et al., 2011; Zheng et al., 2008). They have the potential to yield accurate results, but have the disadvantage of relying on additional data (e.g. accelerometer readings, GIS maps etc.) for their functioning. Also, besides the inherent biases and subjectivity, criteria based methods also require additional data which sometimes is not available. For example Stopher et al. (2008a) and Stopher et al. (2008b) need GPS quality and GIS information, whilst Bohte and Maat (2009) and Chen et al. (2010) need GIS information. The method presented in this paper aims to refine current understanding of measuring environmental exposures in studies using GPS by employing a method that, unlike the above, does not require other information than the speed and location of each GPS point. The model used is applied to a study that aims to investigate associations between individual on foot (or slow cycling) exposure to the food environment and dietary outcomes in children. It was found that for this particular application, there was a strong agreement between the algorithm and two independent human experts, which suggests that, although there is a degree of subjectivity in the human classification due to lack of objective annotated data for the study, the model works as well as a time and resource consuming visual classification method. Few papers report agreement between model and human classification (Auld et al., 2009; Chao et al., 2010; Cho et al., 2011). As a result of application of the algorithm, approximately 18% of the raw GPS data points were marked for removal, which represented motorised vehicle journeys or GPS device inaccuracies. The exposures to the food environment measured before and after processing were however strongly correlated.

One of the strengths of Trans-Mod is the fact that it is an unsupervised model, and hence it does not require manually classified data for training, as supervised models do. Therefore, using individual speed instances to judge the transportation mode is not limited by the fact that any spurious changes in speeds could affect the inferred modes, a problem with supervised methods (Lin et al., 2013). Furthermore, HMM is a mature statistical model that has been extensively and successfully used in many fields. While there are various methods for identifying travel mode in the literature, it was concluded that using a Gaussian-based model such as HMM and some additional pre and post-processing criteria has rendered promising results for the experimental data used. While other methods (Feng and Timmermans, 2013) have differentiated between different modes (walk, car, bus, bike etc.), those researchers had access to more information than available with the dataset used here and for the research purpose of this paper (i.e. identifying exposure to the food environment), such as bus station location for finding bus trips. More detailed information on the exact input variables that were required for the different methods in the literature, can be found in the Gong et al. (2014) review. The method presented here works only with just time-stamped GPS points (no additional data is needed) and it requires minimal user interaction. For this method, the user interaction consisted of visually inspecting a sub-sample of the data at the post-processing stage in order to test the robustness of the algorithm classification.

The decision to choose a threshold of 5 minutes for differentiating between different trips was based on evidence from the literature, as well as a sensitivity analysis that we performed with different thresholds (ranging from 1 to 10 minutes), to see if changing the thresholds result in significant differences between number of trips (Figure 7). We acknowledge that there is some variation in number of trips when using different thresholds to separate trips. However it can be seen in Figure 7 that the difference is more substantial between 1 and 2 minutes, after which it levels out. For our study we have discounted 1 or 2 minutes as being a sensible threshold, because this is the amount of time that could represent waiting in front of a traffic light (Stopher et al., 2008b). We have also based this decision on evidence from the literature; when comparing trip and identification thresholds, a review of methods available (Gong et al., 2014) identifies 300 seconds (which corresponds to 5 minutes) as being the maximum amount of time used in the literature.

Figure 7. Number of trips according to different thresholds (in minutes) to separate trips **[near here]**

In terms of limitations, one consideration is that the PEACH dataset used to train the model is applied to children living in a dense urban area and might not be generalizable to adults or people living in rural areas. Furthermore, spatial accuracy of GPS might be lower in urban areas, because of the density and height of buildings. For example, Schipperijn et al. (2014) ask for caution when studying walking or cycling in dense urban environments, as walking and cycling lanes are typically located closer to buildings and are narrower than vehicle lanes, which may compromise spatial accuracy. Calculating on-foot exposures to the food environment might make a bigger difference in adults after excluding motorised vehicle journeys, as they spend more time in cars. Furthermore, the children in the PEACH study live in Bristol, which means they are more likely to walk or cycle. This can indeed be observed by the fact that many of the trips (62% excluding motorised and mixed mode and spurious points) represent non-motorised journeys.

The GPS model used in this instance was a Garmin Foretex 201, which records location every 10 seconds, a lower frequency than some studies, and this particular device does not use Doppler measures or Horizontal Dilution of Precision which can be used to identify spurious locations due to a poor satellite signal. It could be that applying the algorithm on newer higher performing devices with longer battery life might render higher accuracy of the algorithm. It has indeed been noted in the literature (Beekhuizen et al., 2013; Duncan et al., 2013) that there can be substantial variation in positional error of different GPS models. An additional limitation is that we did not have travel diary data against which to compare classification outcomes, although studies that have done that have shown that classification of algorithm and diary reported trips are similar (Chao et al., 2010; Cho et al., 2011). Nevertheless it is common that trips are reported in travel survey data but are not identified in the GPS data, and reasons for this may include delayed GPS wear at the start of the day, unplanned trips at the end of the day after GPS has been removed, or loss of signal (Wolf et al., 2003a; Wolf et al., 2003b).

Historically studies in the field of public health have typically not attempted to decompose GPS tracks by systematically assessing the nature of activities practiced at the different places and the transportation modes for each trip (Chaix et al., 2013), yet there is now increasing interest in doing so. In the transportation field some studies have combined GPS tracking with precise mobility surveys that collect information on activities and transportation modes. While the method presented here differentiates between motorised and non-motorised exposures based on GPS data collected over 7 days, a survey was not conducted on the nature of activities at specific locations. Therefore, there was no way of knowing if non-motorised exposures to the retail food environment meant that participants actually made use of those particular food outlets.

In this sample, it was observed that likely exposure to the food environment was somewhat over-estimated when not considering time spent in a vehicle, although the correlation between the pre- and post-processing exposure estimates was high. If the requirement of a study is to estimate some form of dose-response relationship between exposure and outcomes, we recommend identification of in-motorised vehicle datapoints in order to refine exposure assessment. Understanding how exposures differ between times spent in vehicles and times spent on foot might be important, for example, in studies attempting to inform planning regulations for fast food outlet density. However, based on our findings at least, applying the algorithm on the sample presented here would not make a significant difference to the statistical strength of association between exposure and outcomes because the pre and post exposure measures to the food environment were strongly correlated.

5. Conclusion:

This paper presents an algorithm, Trans-Mod, to clean GPS data that can be specifically applied to health studies making use of GPS in order to better assess exposure to facilities in the environment by identifying times spent inside and outside vehicles. When applied to an example dataset of food environment exposures amongst children in southwest England, the algorithm suggested that actual opportunities for a sample of children to purchase food might be somewhat over-estimated if time spent in vehicles was not identified, although estimate of exposure prior to processing were strongly correlated with those after processing. The utility of the application of such methods is therefore dependent on the motivation of the research.

Disclaimer:

Please note that the Python scripts that make up Trans-Mod have been made available for download together with implementation instructions at:
https://www.dropbox.com/sh/0x4wdl6mnt5kvdv/AABJ_pIHbrxHo_klTSSjUIvQa?dl=0.

This software is supplied as-is, with no warranty of any kind expressed or implied. We have made every effort to avoid errors in design and execution of this software, but we will not be liable for its use or misuse. The user is solely responsible for the validity and consequences of any results generated. Unfortunately the authors will not be able to provide individual support with implementing the code on your own dataset.

Acknowledgments:

APJ was partially supported by the Centre for Diet and Activity Research (CEDAR), a UK Clinical Research Collaboration Public Health Research Centre of Excellence. Funding from the British Heart Foundation, Economic and Social Research Council, Medical Research Council, National Institute for Health Research and the Wellcome Trust, under the auspices of the UK Clinical Research Collaboration, is gratefully acknowledged.

References:

Auld, J., Williams, C., Mohammadian, A., Nelson, P., 2009. An automated GPS-based prompted recall survey with learning algorithms. Transportation Letters: the International Journal of Transportation Research 1, 59-79.

Beekhuizen, J., Kromhout, H., Huss, A., Vermeulen, R., 2013. Performance of GPS-devices for environmental exposure assessment. Journal of Exposure Science and Environmental Epidemiology 23, 498-505.

Bohte, W., Maat, K., 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. Transportation Research Part C: Emerging Technologies 17, 285-297.

Bricka, S.G., Sen, S., Paleti, R., Bhat, C.R., 2012. An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. Transportation Research Part C: Emerging Technologies 21, 67-88.

Byon, Y.-J., Abdulhai, B., Shalaby, A.S., 2007. Impact of sampling rate of GPS-enabled cell phones on mode detection and GIS map matching performance, Transportation Research Board 86th Annual Meeting.

Carlson, J.A., Jankowska, M.M., Meseck, K., Godbole, S., Natarajan, L., Raab, F., Demchak, B., Patrick, K., Kerr, J., 2015. Validity of PALMS GPS scoring of active and passive travel compared with SenseCam. Medicine and Science in Sports and Exercise 47, 662-667.

Cetateanu, A., Jones, A.P., 2014. Understanding the relationship between food environments, deprivation and childhood overweight and obesity: evidence from a cross sectional England-wide study. Health & Place 27, 68-76.

Chaix, B., Kestens, Y., Perchoux, C., Karusisi, N., Merlo, J., Labadi, K., 2012. An Interactive Mapping Tool to Assess Individual Mobility Patterns in Neighborhood Studies. American Journal of Preventive Medicine 43, 440-450.

Chaix, B., Méline, J., Duncan, S., Merrien, C., Karusisi, N., Perchoux, C., Lewin, A., Labadi, K., Kestens, Y., 2013. GPS tracking in neighborhood and health studies: A step forward for environmental exposure assessment, a step backward for causal inference? Health & Place 21, 46-51.

Chao, X., Minhe, J., Wen, C., Zhihua, Z., 2010. Identifying travel mode from GPS trajectories through fuzzy pattern recognition, Seventh International Conference on on Fuzzy Systems and Knowledge Discovery (FSKD 2010), pp. 889-893.

Chen, C., Gong, H., Lawson, C., Bialostozky, E., 2010. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. Transportation Research Part A: Policy and Practice 44, 830-840.
Cho, G.H., Rodríguez, D.A., Evenson, K.R., 2011. Identifying walking trips using GPS data. Medicine and Science in Sports and Exercise 43, 365-372.
Chung, E.H., Shalaby, A., 2005. A trip reconstruction tool for GPS-based personal travel surveys. Transportation Planning and Technology 28, 381-401.
Dodge, S., Weibel, R., Forootan, E., 2009. Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. Computers, Environment and Urban Systems 33, 419-434.
Duncan, S., Stewart, T.I., Oliver, M., Mavoa, S., MacRae, D., Badland, H.M., Duncan, M.J., 2013. Portable Global Positioning System Receivers: Static Validity and Environmental Conditions. American Journal of Preventive Medicine 44, e19-e29.
Feng, T., Timmermans, H.J.P., 2013. Transportation mode recognition using GPS and accelerometer data. Transportation Research Part C: Emerging Technologies 37, 118-130.
Gong, L., Morikawa, T., Yamamoto, T., Sato, H., 2014. Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies. Procedia - Social and Behavioral Sciences 138, 557-565.
Gonzalez, P., Weinstein, J., Barbeau, S., Labrador, M., Winters, P., Georggi, N.L., Perez, R., 2008. Automating mode detection using neural networks and assisted GPS data collected using GPS-enabled mobile phones, 15th World congress on intelligent transportation systems.
Gustafson, A.A., Lewis, S., Wilson, C., Jilcott-Pitts, S., 2012. Validation of food store environment secondary data source and the role of neighborhood deprivation in Appalachia, Kentucky. BMC Public Health 12.
Harrison, F., Burgoine, T., Corder, K., van Sluijs, E., Jones, A., 2014. How well do modelled routes to school record the environments children are exposed to?: a cross-sectional comparison of GIS-modelled and GPS-measured routes to school. International Journal of Health Geographics 13, 5.
Kerr, J., Duncan, S., Schipperjin, J., 2011. Using Global Positioning Systems in Health Research: A Practical Approach to Data Collection and Processing. American Journal of Preventive Medicine 41, 532-540.
Kestens, Y., Lebel, A., Daniel, M., Thériault, M., Pampalon, R., 2010. Using experienced activity spaces to measure foodscape exposure. Health and Place 16, 1094-1103.
Krenn, P.J., Titze, S., Oja, P., Jones, A., Ogilvie, D., 2011. Use of Global Positioning Systems to Study Physical Activity and the Environment: A Systematic Review. American Journal of Preventive Medicine 41, 508-515.
Lachowycz, K., Jones, A.P., Page, A.S., Wheeler, B.W., Cooper, A.R., 2012. What can global positioning systems tell us about the contribution of different types of urban greenspace to children's physical activity? Health & Place 18, 586-594.
Liese, A.D., Weis, K.E., Pluto, D., Smith, E., Lawson, A., 2007. Food Store Types, Availability, and Cost of Foods in a Rural Environment. Journal of the American Dietetic Association 107, 1916-1923.
Lin, M., Hsu, W.J., Lee, Z.Q., 2013. Detecting modes of transport from unlabelled positioning sensor data. Journal of Location Based Services 7, 272-290.
Moiseeva, A., Timmermans, H., 2010. Imputing relevant information from multi-day GPS tracers for retail planning and management using data fusion and context-sensitive learning. Journal of Retailing and Consumer Services 17, 189-199.
Moon, T.K., 1996. The expectation-maximization algorithm. Signal processing magazine, IEEE 13, 47-60.
Murphy, K., 2012. Machine Learning: A Probabilistic Perspective. MIT Press.
OrdnanceSurvey, 2011. Available at http://www.ordnancesurvey.co.uk/oswebsite/products/points-of-interest/index.html

Patterson, D.J., Liao, L., Fox, D., Kautz, H., 2003. Inferring high-level behavior from low-level sensors, UbiComp 2003: Ubiquitous Computing. Springer, pp. 73-89.

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., Srivastava, M., 2010. Using mobile phones to determine transportation modes. ACM Transactions on Sensor Networks (TOSN) 6, 13.

Schipperijn, J., Kerr, J., Duncan, S., Madsen, T., Klinker, C.D., Troelsen, J., 2014. Dynamic Accuracy of GPS Receivers for Use in Health Research: A Novel Method to Assess GPS Accuracy in Real-World Settings. Front Public Health 2, 21.

Schuessler, N., Axhausen, K., 2009. Processing raw data from global positioning systems without additional information, Transportation Research Record, pp. 28-36.

Stopher, P., Clifford, E., Zhang, J., FitzGerald, C., 2008a. Deducing mode and purpose from GPS data. Institute of Transport and Logistics Studies.

Stopher, P., FitzGerald, C., Xu, M., 2007. Assessing the accuracy of the Sydney Household Travel Survey with GPS. Transportation 34, 723-741.

Stopher, P., FitzGerald, C., Zhang, J., 2008b. Search for a global positioning system device to measure person travel. Transportation Research Part C: Emerging Technologies 16, 350-369.

Tsui, S., Shalaby, A., 2006. Enhanced system for link and mode identification for personal travel surveys based on global positioning systems. Transportation Research Record: Journal of the Transportation Research Board 1972, 38-45.

Viterbi, A.J., 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. Information Theory, IEEE Transactions on 13, 260-269.

Welch, L., 2003. Hidden Markov Models and the Baum-Welch Algorithm. IEEE Information Theory Society Newsletter 53.

Wheeler, B.W., Cooper, A.R., Page, A.S., Jago, R., 2010. Greenspace and children's physical activity: A GPS/GIS analysis of the PEACH project. Preventive Medicine 51, 148-152.

Wolf, J., Loechl, M., Thompson, M., Arce, C., 2003a. Trip rate analysis in GPS-enhanced personal travel surveys in: Stopher, P., Jones, P. (Eds.), Transport Survey Quality and Innovation, pp. 483-498.

Wolf, J., Oliveira, M., Thompson, M., 2003b. Impact of Underreporting on Mileage and Travel Time Estimates: Results from Global Positioning System-Enhanced Household Travel Survey. Transportation Research Record: Journal of the Transportation Research Board 1854, 189-198.

Zhang, L., Dalyot, S., Eggert, D., Sester, M., 2011. Multi-stage approach to travel-mode segmentation and classification of gps traces, ISPRS Workshop on Geospatial Data Infrastructure: from data acquisition and updating to smarter services.

Zheng, Y., Liu, L., Wang, L., Xie, X., 2008. Learning transportation mode from raw gps data for geographic applications on the web, Proceedings of the 17th international conference on World Wide Web. ACM, Beijing, China, pp. 247-256.