

1 **A novel methodology for identifying environmental exposures using GPS**
2 **data**

3 Andreea Cetateanu¹, Bogdan Luca³, Andrei Alin Popescu³, Angie Page⁴, Ashley Cooper⁴, Andy
4 Jones²

5 ¹ School of Environmental Sciences, University of East Anglia, Norwich, Norfolk, NR4 7TJ,
6 UK.

7 ² Norwich Medical School, University of East Anglia, Norwich, Norfolk, NR4 7TJ, UK.

8 ³ School of Computing Sciences, University of East Anglia, Norwich, Norfolk, NR4 7TJ, UK.

9 ⁴ Centre for Exercise, Nutrition and Health Sciences, School for Policy Studies, University of
10 Bristol, 8 Priory Road, Bristol BS8 1TZ.

11
12 Correspondence to: Andreea Cetateanu, a.cetateanu@gmail.com

13 Tel: 00 44 (0)2075942637

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

16

17

18

19

20

21

22

23

24 **Abstract**

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

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

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

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

42

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

44

45

46

47

48

49 **A novel methodology for identifying environmental exposures using GPS** 50 **data**

51 **1. Introduction**

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

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

73 Whilst logging travel patterns using GPS measurements has become commonplace, managing
74 the considerable volume of GPS data collected and extracting meaningful outcome values is
75 difficult. GPS technologies are still developing, with associated different qualities of GPS
76 software and hardware, and even if the device is working at peak performance there will always

77 be some spatial error in the accuracy of location recording (Kerr et al., 2011), which differs
78 based on conditions and type of GPS receiver used. Location errors can emerge from factors
79 such as satellite propagation delays or precision of the device, and signal loss due to slow
80 location detection (initialization and start-up, whereby the GPS receiver needs some time to first
81 acquire signals from satellites) or ground cover such as trees.

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

95 To date a small number of researchers have attempted to produce more robust algorithms for
96 cleaning GPS data and extracting useful information such as travel mode from it (Auld et al.,
97 2009; Carlson et al., 2015; Chao et al., 2010; Feng and Timmermans, 2013; Lin et al., 2013;
98 Schuessler and Axhausen, 2009; Zheng et al., 2008). Whilst there is no uniform standard across
99 disciplines, most methods have several commonalities among them. They typically each attempt
100 to split the raw GPS data into smaller relevant segments (i.e. journeys or trips) on which further
101 analysis is carried out (e.g. determining transport mode for each trip). Usually some form of pre-
102 processing is carried out to remove outliers and de-noise the data, after which a main algorithm
103 is applied for analysis, and subsequently post-processing is used to further improve classification
104 accuracy. These main algorithms can be classified into machine learning approaches and criteria-
105 based approaches. In turn, machine learning approaches can be divided into supervised and
106 unsupervised methods.

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

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

120 Unsupervised methods overcome this disadvantage by not relying on training data for
121 predictions. They rather infer transportation modes based on the structure and the characteristics
122 of the input data, in some cases aided by expert-defined rules, e.g.(Schuessler and Axhausen,
123 2009). For example the work of Lin, et al. (2013) assumes that each transport mode generates
124 speeds from a certain distribution. They use raw GPS data to estimate the parameters of these
125 distributions and conduct statistical tests to determine the differences between these distributions
126 across different segments. Based on these inferred differences, they then use hierarchical
127 clustering to group trips into major groups which correspond to transport modes. Unreliable trips
128 are classified based on proximity to relevant locations such as bus stops. Most of these methods
129 are data intensive and require additional information, such as relevant landmark positions, and
130 would not work as well for studies that do not have such information available.

131 The method presented here (which will be called Trans-Mod) falls in the category of
132 unsupervised methods and is applied on the PEACH (Personal and Environmental Associations
133 with Children's Health) dataset containing the GPS locations of a sample of children in Bristol.
134 The development and testing of the methodology presented in this paper arose from the need to
135 extract only trips not in a motorised vehicle from the PEACH dataset in order to be able to

136 estimate exposure to the food environment and calculated associations with health outcomes
137 such as diet and weight status (results not presented here). The key requirement was to identify
138 times when children were inside a vehicle and those when they were not, as it is assumed that the
139 ability of children to access food outlets will be limited when they are in a vehicle. A model
140 known as a Hidden Markov Model (HMM) (Murphy, 2012) was used to model the differences in
141 speeds from raw GPS data generated by two travel modes: non-motorised (walking or slow
142 cycling) and in a motorised vehicle. HMMs have been previously used (Reddy et al., 2010) to
143 determine travel modes using the information provided by mobile phones (accelerometer and
144 GPS data). However, the method presented here has very low input data requirements, namely
145 just the registered timestamp of each GPS point and the distance between two consecutive
146 points, on the basis of which speed can be easily calculated. The present paper investigates how
147 accurately the method presented here differentiates between motorised and non-motorised travel
148 modes, and if the post-processing exposure estimates of exposure to the food environment differ
149 to those before processing.

150 **2. Methods**

151 ***2.1. Dataset***

152 The dataset used in developing the model presented here was obtained from PEACH, a study
153 undertaken in Bristol, UK, which investigates how the environment can influence physical
154 activity and dietary behaviours in children. Characteristics of the PEACH study sample have
155 been described in more detail elsewhere (Lachowycz et al., 2012; Wheeler et al., 2010). In brief,
156 this dataset provides up to 7 days of GPS data recorded in the morning (8am-9am), evening
157 (3pm-10pm), and at weekends (8am-10pm). In total, 688 children in their first year of secondary
158 school wore a Garmin Foretrex 201 GPS receiver recording data at 10-s intervals (epochs). The
159 GPS has limited battery life, and participants were asked to switch the GPS on at the end of
160 school, and off at bedtime. Research staff charged the units after the first two days of use.

161 GPS data from this study was used to measure personal exposure to the food environment.
162 Measures of the food environment exposure were computed in a Geographical Information
163 System (GIS) (ArcGIS 10.0 (ESRI Inc, Redlands, CA, USA)) using the UK Ordnance Survey
164 Points of Interest (PoI) dataset (OrdnanceSurvey, 2011), a dataset that includes the precise

165 location of 21 categories of food outlets. The location of all food outlets in the Points of Interest
166 data were mapped and grouped into three categories, based on evidence in the literature
167 (Cetateanu and Jones, 2014; Gustafson et al., 2012; Liese et al., 2007), as well as fieldwork visits
168 made by the authors to a sample of outlets falling within each category. The categories chosen
169 were ‘food outlets where people can purchase healthy food’ which was computed to include
170 markets, grocers, organic stores, supermarket chains and independent supermarkets; ‘food outlets
171 where people can purchase unhealthy food’ including bakeries, delicatessens, confectioners,
172 convenience stores and newsagents; and ‘food outlets where people can purchase fast food’ (fast
173 food outlets, takeaways, fast food delivery services that also have an eat in option, and fish and
174 chip shops).

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

188

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

190 *2.2. Trip and travel mode detection, data cleaning and smoothing*

191 *Stage 1: Pre-processing*

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

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

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

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

203 ***Stage 2: Processing***

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

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

215 To model this behaviour we created a HMM model with two hidden states corresponding to non-
216 motorised and motorised states respectively. Each state has its own Gaussian distribution of

217 speeds that represent the emission probabilities of the model. The transition probabilities
218 between the states reflect the likelihood of changing the travel mode.

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

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

231 ***Stage 3: Post-processing***

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

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

252 After processing, there were still some points over 15 kph classified by the model as *non-*
253 *motorised mode*. This was because the speeds were not high enough for the model to suggest
254 them as motorised vehicle points given their surrounding points were mostly non-vehicle. An
255 additional criterion was therefore imposed by marking all of these points as *motorised mode*.
256 This was based on previous practice in studies that have used the same dataset (Lachowycz et al.,
257 2012; Wheeler et al., 2010), where travel speeds above 15kph were judged to be journeys in
258 vehicles. Nevertheless, a limitation of this is that some instances of fast cycling may be classified
259 as motorised mode.

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

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

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

276 **3. Results**

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

281 The Baum-Welch algorithm converged to the parameters illustrated in Figure 2. It can be
282 observed that the emission distribution corresponding to a *non-vehicle* state is centred around
283 2.14 kph, while for the *vehicle* state it is centred around 26.86 kph. These values are consistent
284 with the initial assumption that the speeds should be able to differentiate well between the two
285 travel modes.

286 In terms of transition probabilities, the probability of moving from non-vehicle to vehicle was
287 0.0232 and the probability of moving from a vehicle to non-vehicle state was 0.1223. These low
288 values reflect the fact that the likelihood of two consecutive points corresponding to different
289 travel modes is much lower than that of them being the same. The probability of remaining in the
290 *non-vehicle* state is about 10% percent higher than the probability of remaining in the *vehicle*
291 state. This is explained by the fact that the data is highly right skewed (skewness= 3.401), thus
292 increasing the probability that if in a *non-vehicle* state, one remains in that state.

293 Out of the 366432 GPS points in the PEACH dataset used to train the HMM model, 64385 were
294 marked for removal during the pre-processing, processing and post-processing stages. This
295 meant that 17.57 % of the original GPS points were marked for removal, which represented:
296 0.37% (n= 1347) outliers, 0.08% (n= 282) aberrant speed, 0.006% (n= 21) participants with less
297 than 1 minute worth of GPS data, 15.94% (n= 58409) motorised vehicle points, 0.30% (n= 1087)
298 points representing trips below one minute total duration, and 0.88% (n= 3239) points registering
299 speeds over 15 kph. As a result, 302047 GPS points (82.43%) remained representing non-vehicle
300 points.

301

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

306

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

320

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

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

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

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

333 **Figure 6. Map showing a participant's trip in a day after classification** (© Crown

334 **Copyright/database right 2015. An Ordnance Survey/EDINA supplied service)**

335 [near here]

336 The level of agreement between the algorithm and the annotation by the first and last author was
337 tested with Cohen's kappa (k) on the sub-sample of 99 trips, and it was found that there was
338 strong agreement between the first and last author, as well as between both authors and the
339 algorithm ($k > 0.8$, $p < 0.001$). The first author and the algorithm agreed on the classification of
340 88% of the trips, the last author and the algorithm on 87%, and the first and last author on 89%.
341 Agreement was poorer when trips were classified as mixed by the algorithm, although this was
342 based on only 10 trips, while the first and last author classified differently to the algorithm on
343 just 5 motorised trips and 2 non-motorised trips.

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

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

353

354 **4. Discussion**

355 Complex methods for analysing GPS data exist (Byon et al., 2007; Gonzalez et al., 2008;
356 Moiseeva and Timmermans, 2010; Patterson et al., 2003; Reddy et al., 2010; Tsui and Shalaby,
357 2006; Zhang et al., 2011; Zheng et al., 2008). They have the potential to yield accurate results,
358 but have the disadvantage of relying on additional data (e.g. accelerometer readings, GIS maps
359 etc.) for their functioning. Also, besides the inherent biases and subjectivity, criteria based
360 methods also require additional data which sometimes is not available. For example Stopher et
361 al. (2008a) and Stopher et al. (2008b) need GPS quality and GIS information, whilst Bohte and
362 Maat (2009) and Chen et al. (2010) need GIS information. The method presented in this paper
363 aims to refine current understanding of measuring environmental exposures in studies using GPS
364 by employing a method that, unlike the above, does not require other information than the speed
365 and location of each GPS point. The model used is applied to a study that aims to investigate
366 associations between individual on foot (or slow cycling) exposure to the food environment and
367 dietary outcomes in children. It was found that for this particular application, there was a strong
368 agreement between the algorithm and two independent human experts, which suggests that,
369 although there is a degree of subjectivity in the human classification due to lack of objective
370 annotated data for the study, the model works as well as a time and resource consuming visual
371 classification method. Few papers report agreement between model and human classification
372 (Auld et al., 2009; Chao et al., 2010; Cho et al., 2011). As a result of application of the
373 algorithm, approximately 18% of the raw GPS data points were marked for removal, which
374 represented motorised vehicle journeys or GPS device inaccuracies. The exposures to the food
375 environment measured before and after processing were however strongly correlated.

376 One of the strengths of Trans-Mod is the fact that it is an unsupervised model, and hence it does
377 not require manually classified data for training, as supervised models do. Therefore, using
378 individual speed instances to judge the transportation mode is not limited by the fact that any
379 spurious changes in speeds could affect the inferred modes, a problem with supervised methods
380 (Lin et al., 2013). Furthermore, HMM is a mature statistical model that has been extensively and
381 successfully used in many fields. While there are various methods for identifying travel mode in
382 the literature, it was concluded that using a Gaussian-based model such as HMM and some
383 additional pre and post-processing criteria has rendered promising results for the experimental
384 data used. While other methods (Feng and Timmermans, 2013) have differentiated between

385 different modes (walk, car, bus, bike etc.), those researchers had access to more information than
386 available with the dataset used here and for the research purpose of this paper (i.e. identifying
387 exposure to the food environment), such as bus station location for finding bus trips. More
388 detailed information on the exact input variables that were required for the different methods in
389 the literature, can be found in the Gong et al. (2014) review. The method presented here works
390 only with just time-stamped GPS points (no additional data is needed) and it requires minimal
391 user interaction. For this method, the user interaction consisted of visually inspecting a sub-
392 sample of the data at the post-processing stage in order to test the robustness of the algorithm
393 classification.

394 The decision to choose a threshold of 5 minutes for differentiating between different trips was
395 based on evidence from the literature, as well as a sensitivity analysis that we performed with
396 different thresholds (ranging from 1 to 10 minutes), to see if changing the thresholds result in
397 significant differences between number of trips (Figure 7). We acknowledge that there is some
398 variation in number of trips when using different thresholds to separate trips. However it can be
399 seen in Figure 7 that the difference is more substantial between 1 and 2 minutes, after which it
400 levels out. For our study we have discounted 1 or 2 minutes as being a sensible threshold,
401 because this is the amount of time that could represent waiting in front of a traffic light (Stopher
402 et al., 2008b). We have also based this decision on evidence from the literature; when comparing
403 trip and identification thresholds, a review of methods available (Gong et al., 2014) identifies
404 300 seconds (which corresponds to 5 minutes) as being the maximum amount of time used in the
405 literature.

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

408 In terms of limitations, one consideration is that the PEACH dataset used to train the model is
409 applied to children living in a dense urban area and might not be generalizable to adults or
410 people living in rural areas. Furthermore, spatial accuracy of GPS might be lower in urban areas,
411 because of the density and height of buildings. For example, Schipperijn et al. (2014) ask for
412 caution when studying walking or cycling in dense urban environments, as walking and cycling
413 lanes are typically located closer to buildings and are narrower than vehicle lanes, which may

414 compromise spatial accuracy. Calculating on-foot exposures to the food environment might
415 make a bigger difference in adults after excluding motorised vehicle journeys, as they spend
416 more time in cars. Furthermore, the children in the PEACH study live in Bristol, which means
417 they are more likely to walk or cycle. This can indeed be observed by the fact that many of the
418 trips (62% excluding motorised and mixed mode and spurious points) represent non-motorised
419 journeys.

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

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

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

453 **5. Conclusion:**

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

462 **Disclaimer:**

463 Please note that the Python scripts that make up Trans-Mod have been made available for
464 download together with implementation instructions at:

465 https://www.dropbox.com/sh/0x4wdl6mnt5kvdv/AABJ_pIHbrxHo_kITSSjUlvQa?dl=0.

466 This software is supplied as-is, with no warranty of any kind expressed or implied. We have
467 made every effort to avoid errors in design and execution of this software, but we will not be
468 liable for its use or misuse. The user is solely responsible for the validity and consequences of

469 any results generated. Unfortunately the authors will not be able to provide individual support
470 with implementing the code on your own dataset.

471 **Acknowledgments:**

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

477 **References:**

- 478 Auld, J., Williams, C., Mohammadian, A., Nelson, P., 2009. An automated GPS-based prompted recall
479 survey with learning algorithms. *Transportation Letters: the International Journal of Transportation*
480 *Research* 1, 59-79.
- 481 Beekhuizen, J., Kromhout, H., Huss, A., Vermeulen, R., 2013. Performance of GPS-devices for
482 environmental exposure assessment. *Journal of Exposure Science and Environmental Epidemiology* 23,
483 498-505.
- 484 Bohte, W., Maat, K., 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-
485 based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C:*
486 *Emerging Technologies* 17, 285-297.
- 487 Bricka, S.G., Sen, S., Paleti, R., Bhat, C.R., 2012. An analysis of the factors influencing differences in
488 survey-reported and GPS-recorded trips. *Transportation Research Part C: Emerging Technologies* 21, 67-
489 88.
- 490 Byon, Y.-J., Abdulhai, B., Shalaby, A.S., 2007. Impact of sampling rate of GPS-enabled cell phones on
491 mode detection and GIS map matching performance, *Transportation Research Board 86th Annual*
492 *Meeting*.
- 493 Carlson, J.A., Jankowska, M.M., Meseck, K., Godbole, S., Natarajan, L., Raab, F., Demchak, B., Patrick, K.,
494 Kerr, J., 2015. Validity of PALMS GPS scoring of active and passive travel compared with SenseCam.
495 *Medicine and Science in Sports and Exercise* 47, 662-667.
- 496 Cetateanu, A., Jones, A.P., 2014. Understanding the relationship between food environments,
497 deprivation and childhood overweight and obesity: evidence from a cross sectional England-wide study.
498 *Health & Place* 27, 68-76.
- 499 Chaix, B., Kestens, Y., Perchoux, C., Karusisi, N., Merlo, J., Labadi, K., 2012. An Interactive Mapping Tool
500 to Assess Individual Mobility Patterns in Neighborhood Studies. *American Journal of Preventive*
501 *Medicine* 43, 440-450.
- 502 Chaix, B., Méline, J., Duncan, S., Merrien, C., Karusisi, N., Perchoux, C., Lewin, A., Labadi, K., Kestens, Y.,
503 2013. GPS tracking in neighborhood and health studies: A step forward for environmental exposure
504 assessment, a step backward for causal inference? *Health & Place* 21, 46-51.
- 505 Chao, X., Minhe, J., Wen, C., Zhihua, Z., 2010. Identifying travel mode from GPS trajectories through
506 fuzzy pattern recognition, *Seventh International Conference on on Fuzzy Systems and Knowledge*
507 *Discovery (FSKD 2010)*, pp. 889-893.

508 Chen, C., Gong, H., Lawson, C., Bialostozky, E., 2010. Evaluating the feasibility of a passive travel survey
509 collection in a complex urban environment: Lessons learned from the New York City case study.
510 Transportation Research Part A: Policy and Practice 44, 830-840.

511 Cho, G.H., Rodríguez, D.A., Evenson, K.R., 2011. Identifying walking trips using GPS data. Medicine and
512 Science in Sports and Exercise 43, 365-372.

513 Chung, E.H., Shalaby, A., 2005. A trip reconstruction tool for GPS-based personal travel surveys.
514 Transportation Planning and Technology 28, 381-401.

515 Dodge, S., Weibel, R., Forootan, E., 2009. Revealing the physics of movement: Comparing the similarity
516 of movement characteristics of different types of moving objects. Computers, Environment and Urban
517 Systems 33, 419-434.

518 Duncan, S., Stewart, T.I., Oliver, M., Mavoa, S., MacRae, D., Badland, H.M., Duncan, M.J., 2013. Portable
519 Global Positioning System Receivers: Static Validity and Environmental Conditions. American Journal of
520 Preventive Medicine 44, e19-e29.

521 Feng, T., Timmermans, H.J.P., 2013. Transportation mode recognition using GPS and accelerometer
522 data. Transportation Research Part C: Emerging Technologies 37, 118-130.

523 Gong, L., Morikawa, T., Yamamoto, T., Sato, H., 2014. Deriving Personal Trip Data from GPS Data: A
524 Literature Review on the Existing Methodologies. Procedia - Social and Behavioral Sciences 138, 557-
525 565.

526 Gonzalez, P., Weinstein, J., Barbeau, S., Labrador, M., Winters, P., Georggi, N.L., Perez, R., 2008.
527 Automating mode detection using neural networks and assisted GPS data collected using GPS-enabled
528 mobile phones, 15th World congress on intelligent transportation systems.

529 Gustafson, A.A., Lewis, S., Wilson, C., Jilcott-Pitts, S., 2012. Validation of food store environment
530 secondary data source and the role of neighborhood deprivation in Appalachia, Kentucky. BMC Public
531 Health 12.

532 Harrison, F., Burgoine, T., Corder, K., van Sluijs, E., Jones, A., 2014. How well do modelled routes to
533 school record the environments children are exposed to?: a cross-sectional comparison of GIS-modelled
534 and GPS-measured routes to school. International Journal of Health Geographics 13, 5.

535 Kerr, J., Duncan, S., Schipperjin, J., 2011. Using Global Positioning Systems in Health Research: A Practical
536 Approach to Data Collection and Processing. American Journal of Preventive Medicine 41, 532-540.

537 Kestens, Y., Lebel, A., Daniel, M., Thériault, M., Pampalon, R., 2010. Using experienced activity spaces to
538 measure foodscape exposure. Health and Place 16, 1094-1103.

539 Krenn, P.J., Titze, S., Oja, P., Jones, A., Ogilvie, D., 2011. Use of Global Positioning Systems to Study
540 Physical Activity and the Environment: A Systematic Review. American Journal of Preventive Medicine
541 41, 508-515.

542 Lachowycz, K., Jones, A.P., Page, A.S., Wheeler, B.W., Cooper, A.R., 2012. What can global positioning
543 systems tell us about the contribution of different types of urban greenspace to children's physical
544 activity? Health & Place 18, 586-594.

545 Liese, A.D., Weis, K.E., Pluto, D., Smith, E., Lawson, A., 2007. Food Store Types, Availability, and Cost of
546 Foods in a Rural Environment. Journal of the American Dietetic Association 107, 1916-1923.

547 Lin, M., Hsu, W.J., Lee, Z.Q., 2013. Detecting modes of transport from unlabelled positioning sensor
548 data. Journal of Location Based Services 7, 272-290.

549 Moiseeva, A., Timmermans, H., 2010. Imputing relevant information from multi-day GPS tracers for
550 retail planning and management using data fusion and context-sensitive learning. Journal of Retailing
551 and Consumer Services 17, 189-199.

552 Moon, T.K., 1996. The expectation-maximization algorithm. Signal processing magazine, IEEE 13, 47-60.

553 Murphy, K., 2012. Machine Learning: A Probabilistic Perspective. MIT Press.

554 OrdnanceSurvey, 2011. Available at <http://www.ordnancesurvey.co.uk/oswebsite/products/points-of-interest/index.html>
555

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

591