



OPEN

## Ensemble model for estimating continental-scale patterns of human movement: a case study of Australia

Karen McCulloch<sup>1,2✉</sup>, Nick Golding<sup>3</sup>, Jodie McVernon<sup>1,4,5</sup>, Sarah Goodwin<sup>6</sup> & Martin Tomko<sup>7✉</sup>

Understanding human movement patterns at local, national and international scales is critical in a range of fields, including transportation, logistics and epidemiology. Data on human movement is increasingly available, and when combined with statistical models, enables predictions of movement patterns across broad regions. Movement characteristics, however, strongly depend on the scale and type of movement captured for a given study. The models that have so far been proposed for human movement are best suited to specific spatial scales and types of movement. Selecting both the scale of data collection, and the appropriate model for the data remains a key challenge in predicting human movements. We used two different data sources on human movement in Australia, at different spatial scales, to train a range of statistical movement models and evaluate their ability to predict movement patterns for each data type and scale. Whilst the five commonly-used movement models we evaluated varied markedly between datasets in their predictive ability, we show that an ensemble modelling approach that combines the predictions of these models consistently outperformed all individual models against hold-out data.

Interaction models<sup>1</sup> informed by population-scale mobility data are critical enablers for transport planning and logistics<sup>2</sup>, inequality and economical activity analysis<sup>3</sup>, urban planning and the provision of critical services, including health care and infectious disease epidemiology<sup>4–7</sup>, and other areas where the assessment of the impact of population-scale migratory flows is of importance. Human movement patterns are modelled, extracted and analyzed to better understand the intensities of interactions within a population<sup>8</sup>. A salient application of nuanced mobility pattern mining is infectious disease epidemiology, where the spread of infectious diseases may be facilitated by highly connected individuals or groups of individuals<sup>9</sup>. The identification of (sub)populations at risk may enable better, more targeted public health response, or aid understanding of the contributions of such individuals to epidemic spread<sup>10–12</sup>. Recent work has focused on estimating the impact of restricting mobility as an intervention to slow the spread of SARS-CoV-2<sup>13</sup>. Research linking human movement patterns with the socio-economic status of sub-populations, however, also shows that these relationships are complex and may vary between cities or regions<sup>14</sup>. Yet, too often do population-scale models rely on simplistic mathematical assumptions and lack the understanding of the data that goes into the movement models to predict movement flows.

In practice, the problem of model choice is further confounded with the choice (or availability) of the data used to fit the model. Call Detail Records (CDR) from mobile phones<sup>6,15–18</sup>, population census data<sup>11,15,19</sup> and global positioning satellite (GPS)<sup>16,20–22</sup> data are currently the most widely available data applicable to studies of human movement (historically, other data have been explored, including bank notes<sup>23</sup>). The census is the most comprehensive and costly source of mobility data that strives to capture the population of a country, but has to sacrifice temporal resolution and nuance (data are collected on a single census day typically every five years). Mobility can be inferred from a single census question related to a person's journey to work which requires

<sup>1</sup>Victorian Infectious Diseases Reference Laboratory, Royal Melbourne Hospital at The Peter Doherty Institute for Infection and Immunity, Parkville, VIC, Australia. <sup>2</sup>Doherty Department, University of Melbourne, at The Peter Doherty Institute for Infection and Immunity, Parkville, VIC, Australia. <sup>3</sup>School of Biosciences, The University of Melbourne, Parkville, VIC, Australia. <sup>4</sup>Centre for Epidemiology and Statistics, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, Australia. <sup>5</sup>Infection Modelling, Murdoch Children's Research Institute, Parkville, Australia. <sup>6</sup>Faculty of Information Technology, Monash University, Caulfield, VIC, Australia. <sup>7</sup>Melbourne School of Engineering, The University of Melbourne, Parkville, VIC, Australia. ✉email: karen.mcculloch@unimelb.edu.au; tomkom@unimelb.edu.au

information on their location of usual residence and workplace on the day the census was conducted. CDR-based data are available only for a sub-population of phone users, typically those subscribed to one (out of multiple) mobile providers. The locations of placed or received calls are captured based on the nearest cell tower, at a relatively coarse and variable spatial resolution which is entirely dependent on the distribution of cell towers. Trips can then be inferred, but their purpose remains unknown. With the shift of mobile users to voice over IP and chat communication, these data are increasingly less representative of movement patterns. However, the use of GPS data collected from satellite navigation applications to infer human movement is becoming increasingly popular. Detailed GPS data tracking via mobile applications capture only small sub-populations of users, but come at a much finer spatial granularity and for all types of trips<sup>20,24</sup>. Yet, how well do distinct data sources capture population-scale mobility? The choice of dataset determines the success of the statistical model derived from it, yet all too often the choice is only driven by data availability. Determining what type of data is most appropriate to capture the mobility behaviour across a specific geographical region is treated as a secondary issue<sup>11,21</sup>.

Population mobility models can generally be divided into two groups, those that consider movement frequency (i.e. the number of trips) between two locations to be a decreasing function of distance (gravity models<sup>15,25</sup>) and those that assume the number of potential destinations between two locations determines movement frequency (intervening opportunities model<sup>26</sup>)<sup>21</sup>. However, more recently developed radiation models<sup>15,27</sup> fall in between these two groups of models and utilise a 2-step process to estimate movement frequency between two locations. In step 1, all locations are considered as possible destinations and are ranked depending on opportunities (which are proportional to the location population). In step 2, the location closest to the origin with the highest rank is chosen as the preferred destination<sup>15</sup>. Gravity models have been used to estimate large scale migration patterns across countries<sup>11</sup>. The radiation model has been shown to outperform the gravity model on occasion, including predicting movement patterns within a state and across country scale<sup>15</sup>. However, both model types have been shown to inadequately describe population movement in some settings, i.e. low income countries<sup>12</sup>.

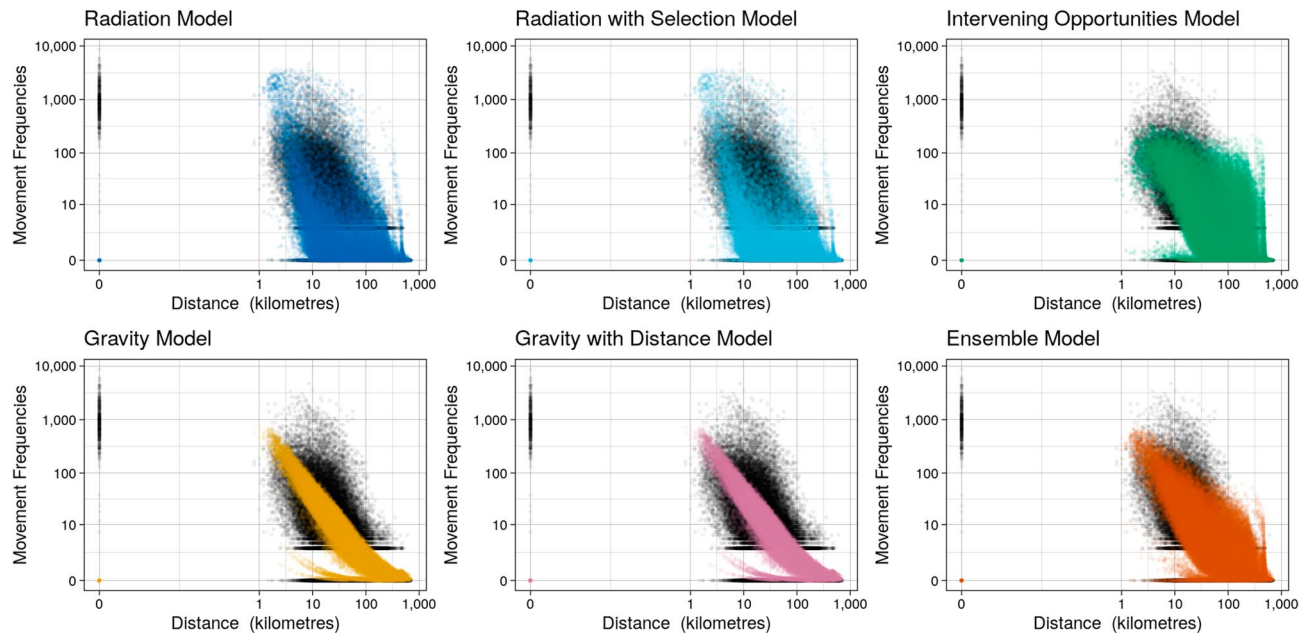
A form of ensemble model was first introduced by Wolpert<sup>28</sup> as stacked generalisations which are motivated by the goal of reducing the generalisation error (or the out-of-sample error). The idea was later formalised by Breiman<sup>29</sup> and the theoretical background was developed by Van der Laan et al.<sup>30</sup> Stacked generalisation involves training a new model to combine predictions of existing models in order to improve predictive performance<sup>31</sup>. During this process the new model learns weights to assign to each existing model based on expected performance. This improves on model averaging processes in which each existing model would be weighted equally. The use of ensemble model approaches aims to shift thinking from selecting the single best available model to deriving an improved model which incorporates information from all available models.

This paper considers the combined challenges of data scales and model types to infer mobility patterns across Australia. Our work is motivated by the need to understand population mobility as a necessary step in the assessment of access to health care services in Australia which varies widely across the continent. We address the need for movement models capturing mobility flows across large geographic areas but representative across spatial scales. The models should capture short distance and long distance travel, adapt to highly heterogeneous space use and diverse geographies. As currently no single model performs consistently best across such varied conditions, we hypothesize that a weighted, locally adaptive combination of each of these models in an ensemble model will better capture the observed movement patterns. Models are trained on data obtained from the Australian Bureau of Statistics (ABS) Census (a snapshot view of the population mobility through their journeys to work between statistical areas)<sup>32</sup> and a dataset collected by the provider of a smartphone-based GPS satellite navigation application. This dataset captures a smaller sample of the population, with a richer set of journeys capturing diverse, yet unspecified, travel needs. We compare model predictions across all models, both datasets, and both spatial scales. In doing so we highlight the strengths and weaknesses in predictive capacity of currently available movement models. Finally, we implement an ensemble model that leads to improved, more accurate predictions of movement flows across different geographical scales.

## Results

The ABS journey to work census data contains counts of individuals travelling between their region of home residence and regions of place of work on the census day, where regions were defined at Statistical Area 2 (SA2) granularity. In Fig. 1 we illustrate the differences in predictive capacity of the five models described in the literature (herein referred to collectively as the base models), using journey to work census data for the Australian state of Victoria. From Fig. 1, it appears that the intervening opportunities model captures more of the observed movement patterns at the state scale than the other base models. The radiation models appear to capture the breadth of the observed data but the intensity of movement predictions is misplaced, i.e., radiation models over-predict the lower frequencies of movement over long distances. In contrast, the gravity models predictions cover a narrow band of the observed data but with higher intensity of predictions. By assigning weights to the predictions generated by each of the base models, the ensemble model is able to capture a wider range of these features. The differences in predictive performance for each of these models is further reflected in Fig. 2C where it is clear that the ensemble model outperforms all base models. The Poisson deviance reported in Fig. 2 is derived by comparing model predictions made on out-of-sample data to the corresponding observed data and can be interpreted as an estimate of model performance relative to other models trained on the same data. A lower deviance indicates a better model fit. Similar patterns were seen for predictions generated from the whole of Australian census data (see Figure S1, Supplementary Material). To aid comparison of base model predictions we provide an alternative visualisation of Fig. 1 in Figure S2 of the Supplementary Material.

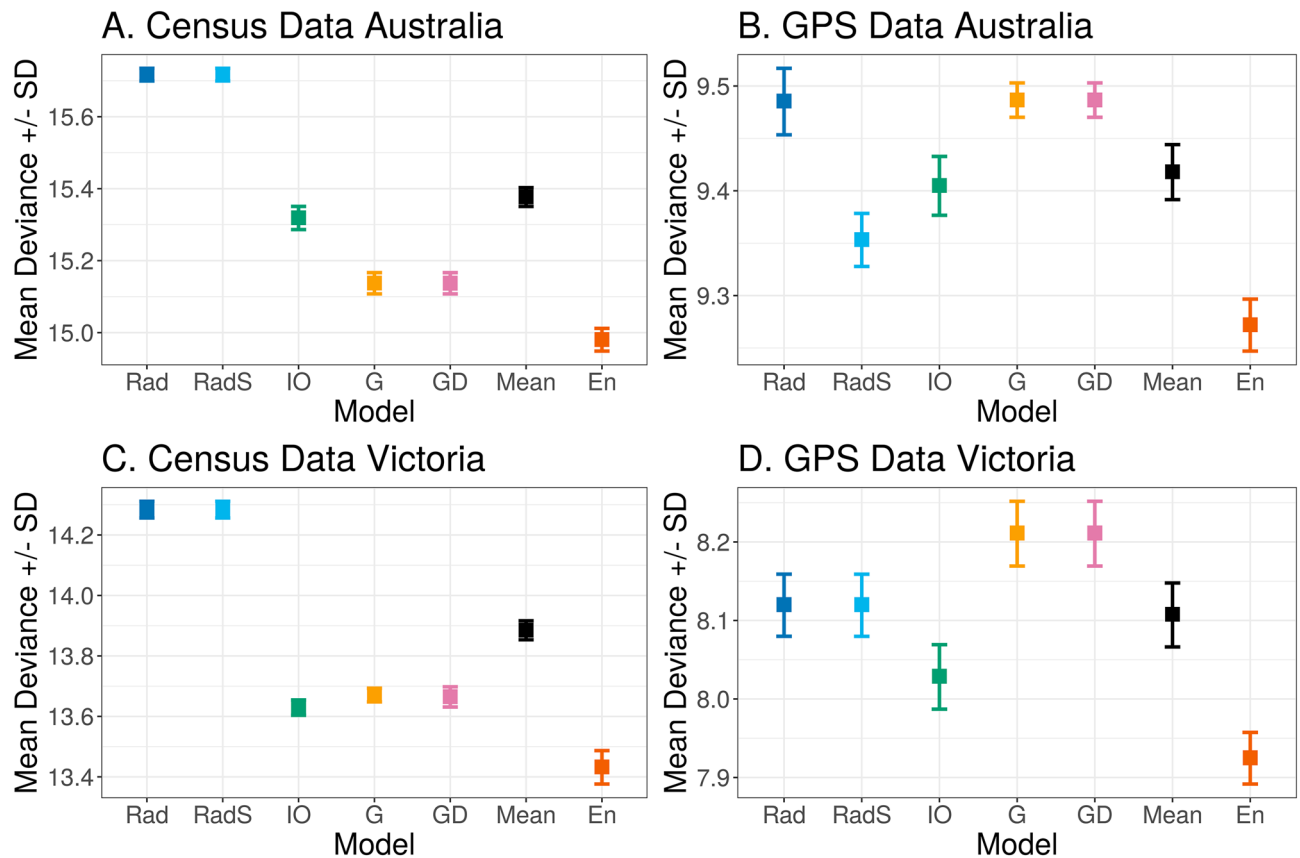
Anonymised GPS data from a smartphone based satellite navigation application were aggregated to provide population level travel flow frequencies between SA2 regions. The observed movement patterns for the GPS data



**Figure 1.** Comparison of predicted frequency of movements by distance (kilometres) between pairs of locations for each model type, based on journey to work Census data. Black points show the raw movement frequencies that were used to train each model. Coloured points show the resulting model predictions. Movement frequency and distance are shown on a log-scale to better highlight differences between model predictions. This figure was produced using the R<sup>33</sup> package ggplot2<sup>34</sup>.

differ the census data (Figures S3 to S6 of Supplementary Material). This is not surprising given the inherently different nature of the data sources, for example the GPS data sample size was much smaller and the maximum observed movement frequency is about half of that seen in the census data. There is more variation in the Poisson deviance estimates for the GPS data (Fig. 2B and D) than the census data. These data encompass a variety of movement flows as the reasons for individual trips is unknown, this is a likely source of some of the variation seen. Therefore, the trained models struggle to capture a wider variety of movement patterns to reflect the input data. Interestingly, out of the five base models the intervening opportunities model performs best at the smaller geographic state scale (Victoria) across both the ABS census journey to work data and the GPS data. At the continental scale, out of the five base models, the gravity models perform best for the ABS census journey to work data, whereas the radiation with selection model performs best for the GPS data. There are other similarities across geographic scales for the datasets, in general the gravity and intervening opportunities models outperform the radiation models for the census data across scales, while the radiation models and intervening opportunities models generally perform better for the GPS data. These results suggest that each dataset is capturing different movement patterns (i.e. purpose of trip) and support the idea that it is necessary to consider multiple types of movement data in order to best capture population movements. These results also highlight that when it comes to population movements there is no single model that optimally fits different data sources. By combining information from each of the base models into an ensemble model we are able to capture more of the observed movement patterns which would have otherwise been missed. The derived ensemble model consistently out-performs the widely used and reported models in the literature for each dataset and both geographic scales (state of Victoria and continental Australia). This can be seen clearly in Figs. 1 and 2 (and Figures S1 to S4 of Supplementary Material). The parameters for the ensemble models trained on Victorian and Australian census and GPS data are provided in the Supplementary Material. In addition to this we provide the Poisson deviance results for each model and dataset from the 5-fold cross validation (see Methods Section for further details) process.

Differences in movement patterns and intensities predicted by the base models at the state scale for each dataset are illustrated in Fig. 3. These graphs further highlight the importance of selecting the best fitting model, as the patterns of predicted movement vary significantly between models trained on the same dataset. We can also compare patterns across the two datasets for the same model type. Generally, across the model types the intensities and patterns that have been predicted for the census data are more defined than those for the GPS data. The Melbourne Metropolitan region is clearly visible in all results. The most notable difference in predictions is by the intervening opportunities model on the census (Fig. 3c) and GPS (Fig. 3d) datasets. The census data predicts higher intensities of movement across long distances, while the GPS data better distinguishes between short to medium distance movement patterns. This is also reflected in Fig. 1 and Figure S3. The movement pattern in Fig. 3g and h illustrate the ensemble model predictions for ABS census data and GPS data respectively. The resulting weights of the ensemble model for the state scale census data indicate that the radiation and radiation with selection model predictions are not very informative. The radiation models predictions are very similar for the census data and they are weighted in such a way that the information provided will almost cancel out (Table 1) when making predictions utilising the ensemble model. Therefore, the ensemble model for census data



**Figure 2.** Comparison of mean deviance between models for each dataset and geographic scale. Legend: G = gravity model; GD = gravity with distance model; Rad = radiation model; RadS = radiation with selection model; IO = intervening opportunities model; En = ensemble model; Mean = mean deviance across all model types. Note the y-axis is on a log scale and the absolute value of the deviance is not comparable between datasets. The deviance is a measure of how well a given model predicts the test data (which was not used to train the models) relative to the other models for a given dataset. The error bars show how the deviance varied for each model type across the 5-folds. This figure was produced using the R<sup>33</sup> package ggplot2<sup>34</sup>.

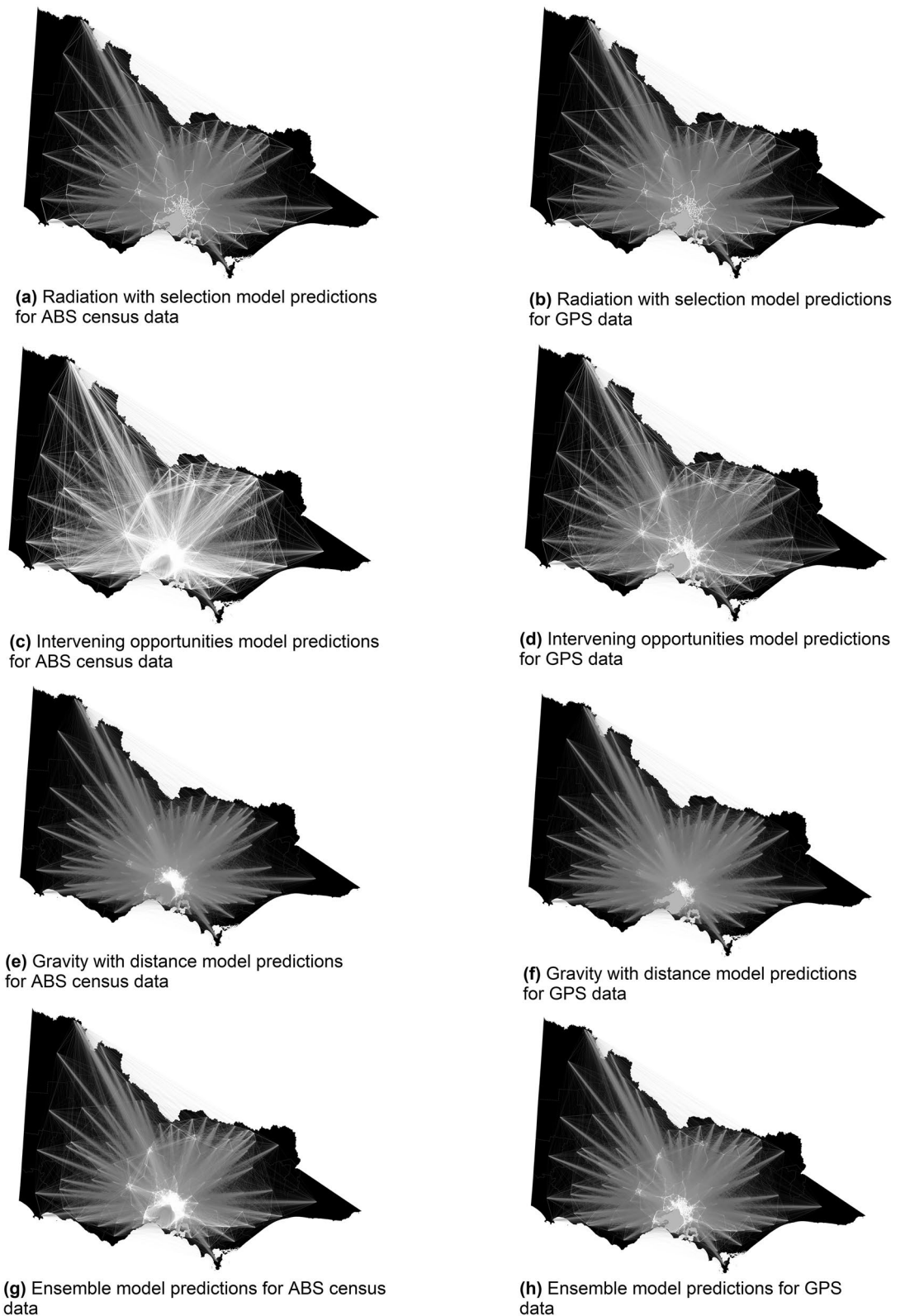
is influenced more by the intervening opportunities and gravity models. Similarly, the ensemble model for the state scale GPS data is influenced more by the intervening opportunities and radiation model predictions. These differences further reflect the different types of movement patterns being captured within each dataset. Figure 4 illustrates the differences in predictive capacity of the base models compared to the ensemble model trained on the ABS census data at the continental scale. The gravity models (Fig. 4B and C) highlight major population hubs such as capital cities across the eastern jurisdictions in Australia with minimal movement flows outside of these hubs. In contrast to the gravity model predictions, the intervening opportunities model (Fig. 4D) predicts long range movement flows at high intensities. The ensemble model predictions clearly highlight the population hubs of the capital cities in addition to smaller regional hubs. Figures 3 and 4 highlight the importance of selecting the most appropriate dataset and geographic scale to predict movement patterns and intensities relevant to the questions being asked.

## Discussion

We have illustrated that an ensemble model with learned weights from currently available movement models outperforms any individual base model in terms of predictive capacity. We hypothesized that different interaction models are suitable for modelling different parts of the specified geographic region. This cannot be simply achieved by averaging across the models (Fig. 2, Mean results) but it can be achieved by utilising an ensemble model. The ensemble model learns from the predictions made by the base models and assigns weights to each set of predictions to reflect which models were the most accurate. The resulting ensemble model predictions better reflect the underlying patterns of movement displayed by the data on which the base models were trained. Furthermore, interactions across large, heterogeneous regions are captured better by the ensemble model. In this study we used Australia as our case study due to the diverse landscape, however this methodology can easily be applied to any other region and landscape.

We have shown that no single model can comprehensively capture the interactions across large, heterogeneous regions. This highlights the importance of considering different types of interactions (predicted by currently available movement models) through the use of ensemble models. In addition, we have also shown that the same





**Figure 3.** Range and intensity of interaction in Victoria revealed through different model types using the ABS census journey to work data (left column) and GPS data (right column). Movement predictions from each model were scaled to be between 0 and 1 to aid visual comparison. This range is mapped to line thickness, opacity and brightness (from black to white), where vivid thick white lines represent high intensity movement and low intensity movement is shown via faint thin darker lines. The Gravity and radiation model prediction maps are omitted as they are visually similar to the gravity with distance and radiation with selection model predictions respectively. The maps in this figure were produced using the R<sup>33</sup> package ggplot2<sup>34</sup> (see Visualisation of predictions section in Methods for further detail). Preview<sup>35</sup> and Paint 3D<sup>36</sup> were used to combine the maps into one image and add the text.



**Figure 4.** Ensemble model predictions for Australia using ABS census data (A). Inset maps for gravity (B), gravity with distance (C), intervening opportunities (D), radiation (E) and radiation with selection (F) models respectively all trained using ABS data. Movement predictions from each model were scaled to be between 0 and 1 to aid visual comparison. Vivid thick white lines represent high intensity movement, whilst faint thin darker lines represent low intensity movement. The maps in this figure were produced using the R<sup>33</sup> package ggplot2<sup>34</sup> (see Visualisation of predictions section in Methods for further detail). Preview<sup>35</sup> and Paint 3D<sup>36</sup> were used to combine the maps into one image, overlay the squares and add the text.

type of model trained on different data will lead to different models and as such it is important to consider the type of data required for estimating movement flows in different contexts.

For some applications and geographical scales of inquiry, the accuracy of the flows predicted by the gravity and intervening opportunities models may be sufficient (e.g., freight transportation, job seeking behaviour). Recent research has, however, focused on identifying gaps in the formulation and application of quantitative models of movement to improve the accuracy and predictive power beyond levels achieved thus far. The radiation model was proposed by Simini et al.<sup>15</sup> as an alternative to the gravity model which may under certain conditions lead to more accurate predictions and has been shown to outperform the gravity model in predicting movement flows across multiple temporal scales. These mathematically elegant and computationally accessible movement models have enabled a widespread proliferation of mobility studies providing a coarse understanding of the population dynamics from national to metropolitan scales. Yet, we argue that the majority of these studies applied the models without attempting to understand their nuanced assumptions, their applicability across geographies and scales, and the errors in the predictions. Population-level movements differ by the motivation and need travel satisfies (e.g., journey to work or for leisure, or flight from a natural or human hazard), by mode of transport, the

geographies covered (cities, coastal plains or mountainous areas), and the temporal dependencies of consecutive travel, each presenting complications for these fundamentally elegant, yet simplistic statistical models.

More recent research in this field<sup>4,12,27</sup> has emphasized the need to validate competing models and proposed a variety of model extensions, each with uncertain transfer-ability to different study areas with distinct space-use patterns or to other applications (transport logistics, population migration and commuting populations). A rigorous comparison of movement flows predicted using different models by Lenormand et al.<sup>19</sup> highlights that the same input data must be used across models for meaningful comparison. In their study, the gravity models with an exponential decay distance function outperformed the intervening opportunities model, highlighting how a person's home location presents a more important parameter than the number of intervening opportunities (work locations). Yet, the ability to tailor and fit nuanced, sophisticated, and locally adapted models<sup>37</sup> is usually either not possible (due to lack of data, and the urgency to model a dynamically evolving phenomenon) or not easily transferable to a different geography or context. A possible departure is the approach of Kraemer et al.<sup>7</sup>, researching generic parameterisations of models of human mobility. More recent research has focused on developing models to improve the predictability of large-scale mobility by utilising a data-driven preferential attachment model<sup>18</sup>.

Ensemble models are valuable tools that can be used to improve predictive accuracy based on existing models, however a key limitation is the required computational power to train base models and carry out cross validation. Given the improvements in computational power since these methods were developed, they are likely to become increasingly popular in future. The Poisson deviance is a measure of how well the model predicts to the out-of-sample data, for this reason it cannot be used as a metric to compare across data sets and can only be used to compare relative model fits on the same data. The ABS census and GPS data both represent different inherently biased samples of population movements. A clear limitation of the census data is that it excludes population movements for those who are not currently employed, in addition the census is a single day sampling event and thus it is only reflective of the population that was employed on the day the census was conducted. The GPS data was sourced from a navigation application and as a result would capture different types of movement, restricted to those who used the navigation application. Therefore, the ABS census data may be reflective of the general working population movements whereas the GPS data will likely only reflect a portion of the population movements. Future work should consider expanding the GPS data set to see if patterns observed in this study hold true for a larger sample.

Our findings can be used to inform epidemiological research across country scale. This research was largely motivated by the need to understand how populations travel in order to access relevant health care. As such, firstly deriving models from appropriate data that can be utilised to predict movement flows between regions more accurately is of key importance. Secondly, combined with appropriate data these movement models can be further utilised to determine a measure of accessibility to relevant health care. Measuring accessibility to health care services also requires consideration of different sub-populations affected, which will again require appropriate use of data.

Standard movement model frameworks (such as the well known gravity and radiation models) take into account the Euclidean distance between the origin region and the destination region, while this can provide some level of insight into overall movement patterns it is not a realistic assumption for longer range movement within Australia. Due to the diverse landscape of Australia, it is necessary to consider the road distance or the road travel time between two regions and not simply the Euclidean distance which does not take into account the diversion of roads around mountain ranges, national parks or deserts. An extension to this work would be to take into account different measures of time and distance, as well as to look at changing patterns of movement over time to see if the same results hold as presented here. Furthermore, finding ways to integrate large spatial scale movement data (i.e. census) with more granular movement data (i.e. GPS) to derive a single model utilising both data types is another crucial way the analyses presented here could be extended upon.

## Methods

**Data sources.** *Census data.* The Australian Bureau of Statistics (ABS) Census data are collected every 5 years and are publicly available online. Within the Employment, Income and Unpaid work section, the journey to work dataset<sup>32</sup> contains counts of individuals travelling between their region of home residence and regions of place of work on the census day. We use these data for regions at Statistical Area 2 (SA2) granularity a) within Victoria and b) across the whole of Australia, based on findings of the 2011 census. The ABS define SA2 regions to have a population size between 3,000 and 25,000, with the mean population size across all SA2 regions equal to 10,000. This results in varied spatial scales of SA2 regions across Australia, with urban regions being more populated and smaller than those less populated regions in remote areas.

*Global positioning satellite data.* Anonymised GPS data from approximately 150,000 unique Australian users of a smartphone-based satellite navigation application capturing individual trips were processed and aggregated to provide population level travel flow frequencies between Statistical Area 2 (SA2) regions from January to October 2016 for a) Victoria and b) the whole of Australia. Aggregated trip frequencies over the entire dataset coverage period were processed to obtain an origin-destination matrix containing the average daily movement between SA2 regions. Each matrix was then normalised to produce a matrix of intensity of contact between origins and destinations. The raw data are not publicly available, however the processed origin - destination matrices are available upon request.

A visual comparison of the observed ABS census and GPS data can be found in Figures S5 and S6 of the Supplementary Material.

**Movement models.** Here, we describe the parameterised movement models used to fit the two described data sources and to predict patterns of population movement across Australia.

1. Gravity Model: The gravity model assumes that the number of individuals moving between locations (origins and destinations) is proportional to a product of the population size of each location, and is inversely proportional to distance. The gravity model assumes that the magnitude of the number of trips,  $T_{ij}$ , between locations  $i$  and  $j$  is defined by:

$$T_{ij} \propto \frac{m_i n_j}{r_{ij}}$$

where  $m_i$  is the population size of the origin,  $n_j$  is the population size at the destination and  $r_{ij}$  is the distance between  $i$  and  $j$ . We use a common functional form of the gravity model which is defined as follows:

$$T_{ij} = \theta \frac{m_i^\alpha n_j^\beta}{r_{ij}^\gamma}$$

where  $T_{ij}$  is the number of individuals moving between location  $i$  and  $j$ ,  $m_i$  represents the population at the origin location,  $n_j$  is the population at the destination location and  $r_{ij}$  is the distance between the two locations,  $\theta$  is a constant.

The parameters  $\alpha$  and  $\beta$  tune the dependence of dispersal on the origin and destination populations respectively<sup>15,38</sup>.

2. Gravity with Distance cut off: This model is a modification of the gravity model which allows short and long trips to be fitted separately, resulting in more accurate predictions<sup>10</sup>. The probability of travelling between two locations within a given distance,  $\delta$ , is different to the probability of travelling between two locations which are separated by a distance greater than  $\delta$ . Therefore, if  $r_{ij} < \delta$  we have

$$T_{ij} = \theta_1 \frac{m_i^{\alpha_1} n_j^{\beta_1}}{r_{ij}^{\gamma_1}}$$

and if  $r_{ij} > \delta$  we have

$$T_{ij} = \theta_2 \frac{m_i^{\alpha_2} n_j^{\beta_2}}{r_{ij}^{\gamma_2}}$$

where  $\theta_1$  and  $\theta_2$  are proportionality constants. The exponents  $\alpha$  and  $\beta$  tune the dependence of dispersal on the origin and destination population sizes ( $m_i$  and  $n_j$  respectively) and the distance between two regions,  $r_{ij}$ <sup>38</sup>.

3. Original Radiation: The (original) radiation model generally assumes the rational of job selection. It follows the general rule that the number of employment opportunities in each region is proportional to its resident population, assuming full employment (that is the number of people in a region is equal to the number of jobs in that region). Moreover, the individuals in each region choose the closest job to their home. The average movement flow from location  $i$  to location  $j$  predicted by the radiation model is defined by:

$$T_{ij} = \theta \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})}$$

where  $m_i$  is the population at the origin and  $n_j$  at the destination,  $s_{ij}$  denotes the total population in a specified radius around population centres  $m_i$  and  $n_j$ , and  $\theta$  is the proportion of the population in region  $i$  that commute<sup>15,38</sup>. In this study, the radiation model parameters are estimated based on information about opportunities in each region.

4. Radiation with Selection: This model is a combination of the original radiation model (above) and intervening opportunities model (see below). The average movement from region  $i$  to region  $j$  according to the radiation model with selection is defined by:

$$T_{ij} = \theta \frac{(1 - \frac{\lambda^P}{P})(1 - \frac{\lambda^Q}{Q})}{(1 - \frac{\lambda^R}{R})}$$

where  $\theta$  is the proportion of population  $i$  that commute,  $P$  is the population at the origin location  $i$  and  $Q$  at the destination location  $j$ ,  $R$  denotes the total population in a radius  $\lambda$  around population centres  $P_i$  and  $Q_j$ <sup>5,15,27,38</sup>.

5. Intervening Opportunities: Stouffer<sup>26</sup> formulated the intervening opportunities model based on the idea that distance and mobility are not directly related as suggested by the gravity model. The reasoning behind this is formed by the idea that the decision to make a trip is related to the relative accessibility of opportunities that satisfy the objective of the trip<sup>21</sup>. That is, the higher the number of intervening opportunities between two points, the lower the number of people travelling the whole distance between those two points. This results in the following definition:



Parameter	Ensemble model			
	Census Victoria data	Census Australia data	GPS Victoria data	GPS Australia data
$\beta_0$	0.9096	0.9442	-0.0021	0.3163
$\beta_1$	0.3215	0.5347	2897.8930	1.1379
$\beta_2$	-0.0036	-0.0687	-2897.9680	-0.8193
$\beta_3$	821.6171	25.7575	-93.8892	-0.0894
$\beta_4$	-821.4218	-25.5504	94.1929	0.6769
$\beta_5$	0.3701	0.1945	0.3829	0.1419

**Table 1.** Ensemble model weights. Parameter description:  $\beta_0$  model intercept,  $\beta_1$  gravity model weight,  $\beta_2$  gravity with distance model weight,  $\beta_3$  radiation model weight,  $\beta_4$  radiation with selection model weight and  $\beta_5$  intervening opportunities model weight.

$$T_{ij} \propto \frac{n_j}{(s_{ij} + m_i)}$$

where  $n_j$  is the total population in location  $j$  and  $(s_{ij} + m_i)$  is the total population in locations between  $i$  and  $j$ <sup>26</sup>. This formulation of the model was redefined as a stochastic approach by Schneider in 1959 as explained in Simini et al.<sup>15</sup>. The probability that an individual in location  $i$  travels and ends their trip in location  $j$  is equal to the probability that location  $j$  contains an acceptable destination and that an acceptable destination closer to the origin location  $i$  has not been found. Therefore, the number of trips between location  $i$  and  $j$  is defined by:

$$T_{ij} \propto e^{-\lambda(s_{ij}+n_i)^\theta} - e^{-\lambda(s_{ij}+n_i+n_j)^\theta}$$

where  $e^{-\lambda}$  is the probability that a single opportunity is not sufficiently attractive as a destination<sup>15</sup>. We make use of this intervening opportunities model defined by<sup>15</sup> for the purposes of this study. As the above movement models do not allow for movements within the same region (i.e. self loops), all predicted values for which the origin and destination location are the same were set to zero.

**Ensemble movement model.** An ensemble movement model is derived based on the hypothesis that combining information from each type of movement model outlined above will produce more robust predictions of movement flows. We derive a type of ensemble model referred to as a stacked model, an idea first introduced by Wolpert<sup>28</sup> as stacked generalizations. Breiman<sup>29</sup> later introduced the concept of stacked regressions, which developed on Wolpert's ideas and formalized a way to generate improved prediction accuracy by forming linear combinations of different predictors which were trained on the same dataset. Below we describe our ensemble model which utilises a Poisson generalized linear model (GLM)<sup>39</sup> as the 'stacker' to derive more robust predictions of movement flows between locations using the predictions generated from each of the above models as our input data.

Let  $\Lambda_a(i, j)$  be the predicted movement flow matrix derived from model  $a$ , where  $(i, j) \in \{1 \dots z\}$  and  $z$  is the number of locations of interest. Then  $\lambda_{ak}$  is the corresponding vectorized form of  $\Lambda_a$  with  $k = 1 \dots z^2$ . In the following  $a = 1$  denotes predictions generated from the gravity model;  $a = 2$  denotes gravity with distance model;  $a = 3$  denotes radiation model;  $a = 4$  denotes radiation with selection model and  $a = 5$  denotes intervening opportunities model. Let  $y_k$  be the ensemble predicted movement flow and  $\mu_k$  be the expected value of  $y_k$ . We estimate  $\mu_k$  by deriving a Poisson GLM which takes the following form

$$\begin{aligned} \log(\mu_k) &= \beta_0 + \log(\lambda_{1k}^{\beta_1}) + \log(\lambda_{2k}^{\beta_2}) + \log(\lambda_{3k}^{\beta_3}) + \log(\lambda_{4k}^{\beta_4}) + \log(\lambda_{5k}^{\beta_5}) \\ &= \beta_0 + \beta_1 \log(\lambda_{1k}) + \beta_2 \log(\lambda_{2k}) + \beta_3 \log(\lambda_{3k}) + \beta_4 \log(\lambda_{4k}) + \beta_5 \log(\lambda_{5k}) \end{aligned}$$

where each of the  $\beta_i$  coefficients (also referred to as weights) are fitted to the input data. Then,  $E[y_k] = \mu_k$ . The estimated weights for the ensemble models for each dataset are provided in Table 1. After transforming the vector  $\mu_k$  back into a matrix of size  $z \times z$ , we obtain the matrix for the predicted movement flows between locations  $i$  and  $j$

$$E[y(i, j)] = \mu(i, j).$$

**Model parameterisation and prediction.** We parameterise the five movement models described above by fitting each to the two observed population movement flow datasets (ABS and GPS) separately. Model fitting was performed using the movement package<sup>38</sup> in R<sup>40</sup>. The fitting process results in estimates for each of the tunable parameters specified in models (1) to (5) above. Parameters for each model are optimised based on log likelihoods against the observed data assuming a poisson sampling distribution, as in the ensemble model. The Broyden-Fletcher-Goldfarb-Shanno (BFGS)<sup>41</sup> method is utilised to carry out the parameter optimisation process.

**Cross validation.** A  $k$ -fold cross validation with  $k = 5$  folds has been used to assess how well the above models predict movement. First, each dataset (ABS and GPS) is divided up into 5 folds (or sections). We then parameterised the movement models (described by (1) to (5) above) by training them using 4 folds (or 80% of the data). Then the parameterised models were used to predict the withheld fold (or remaining 20% of the data) in order to evaluate prediction accuracies of the models. This process was repeated until each of the 5 folds had been used as the withheld data. The 5-fold cross validation was also carried out for the ensemble model. This approach was implemented using the Caret<sup>42</sup> R package.

After training the model, the accuracy of predictions was evaluated by computing the Poisson deviance. Poisson deviance allows measurement of the error between the observed and predicted values and thus enables assessment of how well each individual model predicts observed movement intensities between locations. It is important to note that the Poisson deviance provides a measure of relative fit. This means that values of the Poisson deviance can be compared across each fold for the same model but values cannot be compared between models. Tables S1 to S4 in the Supplementary Material give the Poisson deviance for the comparison of predicted and observed movement flows for each fold in the cross validation for each combination of model type and data source. Tables S5 to S8 in the Supplementary Material provide parameter estimates generated from 5-fold cross validation for each of the model types.

Once each model was parameterised, we used it to predict the movement flows and evaluate the accuracy of the predictions against the observed data. Even though we fitted the models using frequency travel data between SA2 regions in Australia (which can be large geographical areas), the final parameterised models can be utilised to predict the likelihood of movement flows between smaller geographical areas of any size, with the caveat that the location and population size of the origin and destination areas is still required.

**Visualisation of predictions.** To aid geographical visualisation, the predicted movements for each model were put into an origin - destination matrices and scaled to ensure all predictions were between 0 and 1. This allowed better comparison of different model predictions within and between datasets. Figures 1 to 4 were produced using the R<sup>35</sup> package ggplot2<sup>34</sup>. Preview<sup>35</sup> and Paint 3D<sup>36</sup> were used to combine the maps into one image, overlay the squares and add the text on Figs. 3 and 4. Data used in Figs. 3 and 4 consist of predictions generated from the individual models and the ensemble model, which utilised ABS census and GPS data (see Methods section for details) as inputs into the models. The geographic boundaries of Australia and the state of Victoria for the maps in Figs. 3 and 4 were sourced from the ABS and mapped using the Geocentric Datum of Australia 1994 (GDA94) coordinate reference system.

## Data availability

The Australian Bureau of Statistics census journey to work data is freely available to download from their website. The raw GPS data utilised are not publicly available, however the processed origin - destination matrices are available upon request. All model predicted origin-destination matrices are available upon request.

Received: 30 September 2020; Accepted: 11 February 2021

Published online: 26 February 2021

## References

- Oshan, T. M. The spatial structure debate in spatial interaction modeling: 50 years on. *Progress Hum. Geogr.* <https://doi.org/10.1177/0309132520968134> (2020).
- Zhang, X., Xu, Y., Tu, W. & Ratti, C. Do different datasets tell the same story about urban mobility—a comparative study of public transit and taxi usage. *J. Transp. Geogr.* **70**, 78–90. <https://doi.org/10.1016/j.jtrangeo.2018.05.002> (2018).
- Weiss, D. J. *et al.* A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* **553**, 333–336. <https://doi.org/10.1038/nature25181> (2018).
- Piovani, D., Arcaute, E., Uchoa, G., Wilson, A. & Batty, M. Measuring Accessibility using Gravity and Radiation Models. *R. Soc. Open Sci.* **5** (2018). <https://doi.org/10.1098/rsos.171668>. arXiv:1802.06421.
- Tizzoni, M. *et al.* On the use of human mobility proxies for modeling epidemics. *PLoS Comput. Biol.* **10**, <https://doi.org/10.1371/journal.pcbi.1003716> (2014).
- Bengtsson, L. *et al.* Using mobile phone data to predict the spatial spread of cholera. *Sci. Rep.* **5**, 1–5. <https://doi.org/10.1038/srep08923> (2015).
- Kraemer, M. U. G. *et al.* Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings. *Sci. Rep.* **9**, 5151. <https://doi.org/10.1038/s41598-019-41192-3> (2019).
- Meekan, M. G. *et al.* The ecology of human mobility. *Trends Ecol. Evol.* **32**, 198–210. <https://doi.org/10.1016/j.tree.2016.12.006> (2017).
- Brdar, S., Gavrić, K., Čulibrk, D. & Crnojević, V. Unveiling spatial epidemiology of HIV with mobile phone data. *Sci. Rep.* **6**, 19342. <https://doi.org/10.1038/srep19342> (2016).
- Viiboud, C. *et al.* Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312**, 447–451 (2006).
- Wesolowski, A. *et al.* The use of census migration data to approximate human movement patterns across temporal scales. *PLoS ONE*, <https://doi.org/10.1371/journal.pone.0052971> (2013).
- Wesolowski, A., O'Meara, W. P., Eagle, N., Tatem, A. J. & Buckee, C. O. Evaluating spatial interaction models for regional mobility in Sub-Saharan Africa. *PLoS Comput. Biol.* **11**, 1–16. <https://doi.org/10.1371/journal.pcbi.1004267> (2015).
- Scala, A. *et al.* Time, space and social interactions: exit mechanisms for the Covid-19 epidemics. *Sci. Rep.* **10**, <https://doi.org/10.1038/s41598-020-70631-9> (2020).
- Xu, Y., Belyi, A., Bojic, I. & Ratti, C. Human mobility and socioeconomic status: analysis of Singapore and Boston. *Comput. Environ. Urban Syst.* **72**, 51–67. <https://doi.org/10.1016/j.compenvurbsys.2018.04.001> (2018).
- Simini, F., González, M. C., Maritan, A. & Barabási, A. L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100. <https://doi.org/10.1038/nature10856> (2012).
- Mucelli Rezende Oliveira, E., Carneiro Viana, A., Sarraute, C., Brea, J. & Alvarez-Hamelin, I. On the regularity of human mobility. *Pervasive Mobile Comput.* **33**, 73–90. <https://doi.org/10.1016/j.pmcj.2016.04.005> (2016).

17. Silveira, L. M., de Almeida, J. M., Marques-Neto, H. T., Sarraute, C. & Ziviani, A. MobHet: predicting human mobility using heterogeneous data sources. *Comput. Commun.* **95**, 54–68. <https://doi.org/10.1016/j.comcom.2016.04.013> (2016).
18. Zhang, H. T. *et al.* Spatiotemporal property and predictability of large-scale human mobility. *Physica A* **495**, 40–48. <https://doi.org/10.1016/j.physa.2017.12.024> (2018).
19. Lenormand, M., Bassolas, A. & Ramasco, J. J. Systematic comparison of trip distribution laws and models. *J. Transp. Geogr.* **51**, 158–169. <https://doi.org/10.1016/j.jtrangeo.2015.12.008> (2016).
20. Paz-Soldan, V. A. *et al.* Strengths and Weaknesses of Global Positioning System (GPS) Data-loggers and semi-structured interviews for capturing fine-scale human mobility: findings from Iquitos. *Peru. PLoS Negl. Trop. Dis.* **8**, <https://doi.org/10.1371/journal.pntd.0002888> (2014).
21. Barbosa, H. *et al.* Human mobility: models and applications. *Phys. Rep.* **734**, 1–74. <https://doi.org/10.1016/j.physrep.2018.01.001> (2018).
22. Noulas, A., Scellato, S., Lambiotte, R., Pontil, M. & Mascolo, C. A tale of many cities: universal patterns in human urban mobility. *PLoS One* **7**, <https://doi.org/10.1371/journal.pone.0037027> (2012).
23. Brockman, D., Hufnagel, L. & Geisel, T. The scaling laws of human travel. *Nature* **439**, 363–368. <https://doi.org/10.1038/nature04292> (2006).
24. Scherrer, L., Tomko, M., Ranacher, P. & Weibel, R. Travelers or locals? Identifying meaningful sub-populations from human movement data in the absence of ground truth. *EPJ Data Sci.* **7**, 19. <https://doi.org/10.1140/epjds/s13688-018-0147-7> (2018).
25. Zipf, G. The P1.P2/D hypothesis: on the intercity movement of persons. *Am. Soc. Rev.* **11**, 677–686 (1946).
26. Stouffer, S. A. Intervening opportunities: a theory relating mobility and distance. *Am. Soc. Rev.* **5**, 39–62 (1940).
27. Simini, F., Maritan, A. & Nédá, Z. Human mobility in a continuum approach. *PLoS ONE* **8**, <https://doi.org/10.1371/journal.pone.0060069> (2013).
28. Wolpert, D. Stacked generalization (stacking). *Neural Netw.* **5**, 241–259 (1992).
29. Breiman, L. Stacked regressions. *Mach. Learn.* **24**, 49–64 (1996).
30. van der Laan, M. J., Polley, E. C. & Hubbard, A. E. Super learner. *Statistical applications in genetics and molecular biology* **6** (2007). <https://www.degruyter.com/view/journals/sagmb/6/1/article-sagmb.2007.6.1.1309.xml.xml>. <https://doi.org/10.2202/1544-6115.1309>.
31. Boehmke, B. & Greenwell, B. *Hands-on machine learning with R* (Chapman and Hall, 2019). <https://bradleyboehmke.github.io/HOML/>.
32. Australian Bureau of Statistics. Employment, Income and Unpaid Work (2011).
33. RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA (2015). <http://www.rstudio.com/>.
34. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, New York, 2009). <http://ggplot2.org>.
35. Apple Inc. *Preview*. Apple Inc (2020). <https://support.apple.com/en-au/guide/preview/apd2185d6ee5/mac>.
36. Corporation, M. *Paint 3D* (Microsoft Corporation, New York, 2018).
37. Yan, X.-Y., Wang, W.-X., Gao, Z.-Y. & Lai, Y.-C. No title. *Nat. Commun.* **8**, 1639 (2017).
38. Golding, N. *et al.* movement: modelling and analysing human movement data. R package version 0.4. (2015). <https://github.com/SEEG-Oxford/movement>.
39. Dobson, A. J. *An Introduction to Generalized Linear Models* 2nd edn. (Chapman & Hall/CRC, Boca Raton, 2001).
40. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna (2017). <https://www.R-project.org/>.
41. Fletcher, R. *Practical Methods of Optimization* 2nd edn. (Wiley, New York, 1987).
42. Kuhn, M. C. f. *et al.* Classification and Regression Training (2018). <https://github.com/topepo/caret/>.

## Acknowledgements

This work has been supported by FrontierSI (formerly CRC for Spatial Information), a not-for-profit company that exists to deliver major benefits to governments, industry and the community in Australia and New Zealand through the application of spatial information. The authors would also like to thank Sygic for allowing access to anonymized data for the purposes of this research.

## Author contributions

K.M. carried out the modelling, analysed the results, drafted the manuscript. N.G. conceived the experiments, developed the software that was used to derive the movement models, analysed results. J.M. conceived the experiments, analysed results. S.G. produced the visualizations for Figs. 3 and 4 and provided expertise on data visualization. M.T. conceived the experiments, analysed results, sourced data (GPS). All authors reviewed the manuscript.

## Competing interest

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-84198-6>.

**Correspondence** and requests for materials should be addressed to K.M. or M.T.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021