





Predicting Pedestrian Counts using Machine Learning

Molly Asher, Yannick Oswald , and Nick Malleison 

School of Geography, University of Leeds, UK

Correspondence: Nick Malleison (n.s.malleison@leeds.ac.uk)

Abstract. The study of urban population dynamics has long been an important area of research. In particular, the ability to accurately predict the number of pedestrians in a place and time is critical for urban management, population health, crime, and for quantifying the impacts of public events. However, it can be extremely difficult to analyse the size and characteristics of the ambient population due to limited data availability and difficulties in capturing non-linear relationships between pedestrian counts and external factors. This paper reports on an ongoing project that is using machine learning techniques to: (i) better understand the impact that the built environment and other contextual factors, such as weather conditions, will have on the size of the pedestrian population during the day and; (ii) predict the number of pedestrians under different conditions. The case study area is the city of Melbourne, Australia, where abundant pedestrian count data exist. Early results demonstrate that, broadly, the model appears to perform sufficiently well to be useful, and that modelling errors are not consistent across space or time (some times/places are easier to predict than others).

Keywords. ambient population, temporary population, random forest, machine learning, footfall, pedestrian counts, open data

1 Introduction

The study of population dynamics has long been an important area of research, with implications for fields ranging from urban planning and transportation to public health and safety. In particular, the ability to accurately predict the size of the pedestrian or *ambient population* – defined as the population “within a given geographical area at a specific point in time, excluding individuals at their place of residence and those utilising modes of transport” (Whipp et al., 2021) – is critical for: urban management; health, i.e. understanding the relationship between pollution and individual exposure (Park and Kwan, 2017); crime, i.e. quantifying the impacts of visitors and residents on crime rates (Boivin and Felson, 2017); and for estimat-

ing the success of public events, i.e. measuring the ability of particular events to draw crowds to cities.

Despite its importance for understanding urban dynamics, it can be extremely difficult to analyse the size and characteristics of the ambient, or “temporary” (Panczak et al., 2020) population (Malleison and Andresen, 2016). There are two main reasons for this. Firstly, whilst there are usually abundant data with information about the number and characteristics of residents in an area from sources such as household surveys and population censuses, similar data do not usually exist for temporary/ambient populations. Secondly, the relationship between the urban environment and the behaviour of the ambient population is difficult to model as many (inter)relationships between variables may be non-linear. For example, a narrow, historical street may *deter* pedestrians during weekdays as people use main thoroughfares to move quickly between activities, but may *attract* people at weekends when they might like to spend time exploring the more unusual/interesting parts of a city.

This paper reports on an ongoing project that is using machine learning techniques to:

- better understand the impact that the built environment and other contextual factors, such as weather conditions, will have on the number of pedestrians;
- predict the size of the pedestrian population under different conditions.

The case study area is the city of Melbourne, Australia, where abundant pedestrian data exist thanks to a large number of sensors that have been installed by the local government and reported on a publicly available open data portal.

We will discuss the data and methods employed in Sections 2 and 3 respectively, with reproducibility addressed in Section 4. We then present preliminary results in Section 5 and draw conclusions in Section 6.

2 Data

Our case study area is the city of Melbourne, Australia. The city has abundant high-resolution data openly avail-

able at the City of Melbourne Open Data Portal (<https://data.melbourne.vic.gov.au/pages/home/>). This includes hourly counts of pedestrians and a wealth of useful information about the built environment that can be used to attempt to predict footfall. Separately the Melbourne weather service also provides historic weather data (hourly temperature, humidity, pressure, wind speed, and a binary measure of the presence of rainfall).

The footfall data are from 82 sensors that detect the movement of people and can be used to measure the ambient population (City of Melbourne Open Data Portal, 2023). The sensors are “typically installed under an awning or on a street pole to form a counting zone on the footpath below” and record “movements”, although the precise mechanism for detecting pedestrians is not stated (City of Melbourne Open Data Portal, 2023). Although some sensors were reporting as far back as 2009, many of the sensors were installed more recently. In addition, some sensors do not have full count information throughout their entire period of operation, as illustrated by Figure 1. Figure 2 illustrates the locations of all sensors at the time of writing.

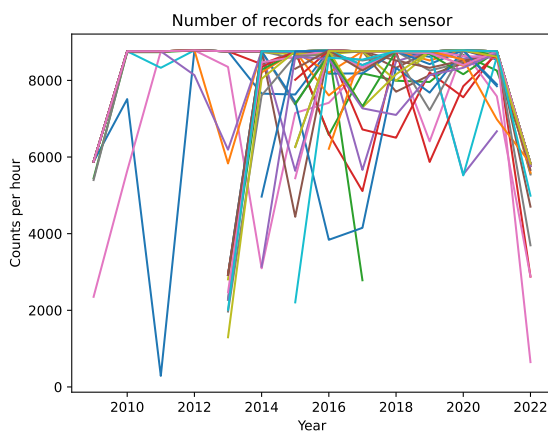


Figure 1. The number of records for each sensor over time. Each sensor is represented with a different colour. There are 8,760 hours per year (excluding leap years), so sensors with fewer than this number of records in a year are missing some counts. Note that the apparent decrease in counts in 2022 is an artefact of the date that the data were downloaded – August 2022 – that meant 2022 is incomplete.

To capture additional factors that will influence footfall, we also extracted the day of week, week number, month, season and year from the raw time of each pedestrian count record.

Relevant data on the built environment that were used to create explanatory variables include:

Pedestrian Footpath Network: we calculate the *betweenness* – a measure that is typically used in space syntax (Bafna, 2003) to quantify how well a link is connected to the wider network – for the network of roads and pedestrian footpaths to capture the level of connectivity for each road, hypothesising that better

connected roads are likely to exhibit greater pedestrian traffic (Leccese et al., 2020);

Street Furniture: locations of benches, information pillars, litter bins, street lights, etc;

Buildings: locations, types and sizes of different buildings (residences, shops, hospitals, leisure establishments, etc.);

Landmarks: including places of worship, community centres, etc.

A 100m radius was drawn around each sensor in order to associate the sensors with the built environment data. For the following features, a count of the number of objects within each sensor radius was calculated: street furniture items; lights; buildings; and landmarks. For betweenness, the value of the footpath edge that was closest to the sensor was taken. The average number of floors within the sensor’s radius was also calculated and included as a variable.

3 Methods

3.1 Model evaluation

As discussed in Section 1, a linear model is unlikely to correctly capture the relationships between the pedestrian count and the explanatory variables. Therefore we additionally considered a number of machine learning algorithms which can better capture non-linear relationships in the data. The candidate models (linear regression, XGBoost and Random Forest) were evaluated using a 10-fold cross-validation procedure. K-fold cross validation partitions the data into k equally sized subsets, and iteratively uses $k-1$ subsets of the data to train the model, holding out the final subset in order to evaluate model performance. Ultimately the Random Forest Regressor (`RandomForestRegressor`¹ in `scikit-learn`) was eventually selected for use as the predictive model. A random forest is so-called because it is built up from an ensemble of individual decision trees (a commonly-used classification method) and, for a given input, returns the average of the individual decision tree predictions.

The model performance was evaluated through comparison of the counts-per-hour predicted by the model with the real counts in the pedestrian data. We summarise performance using the mean absolute error (MAE), the mean absolute percentage error (MAPE) and the root mean squared error (RMSE) which are calculated as follows:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (1)$$

¹<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>

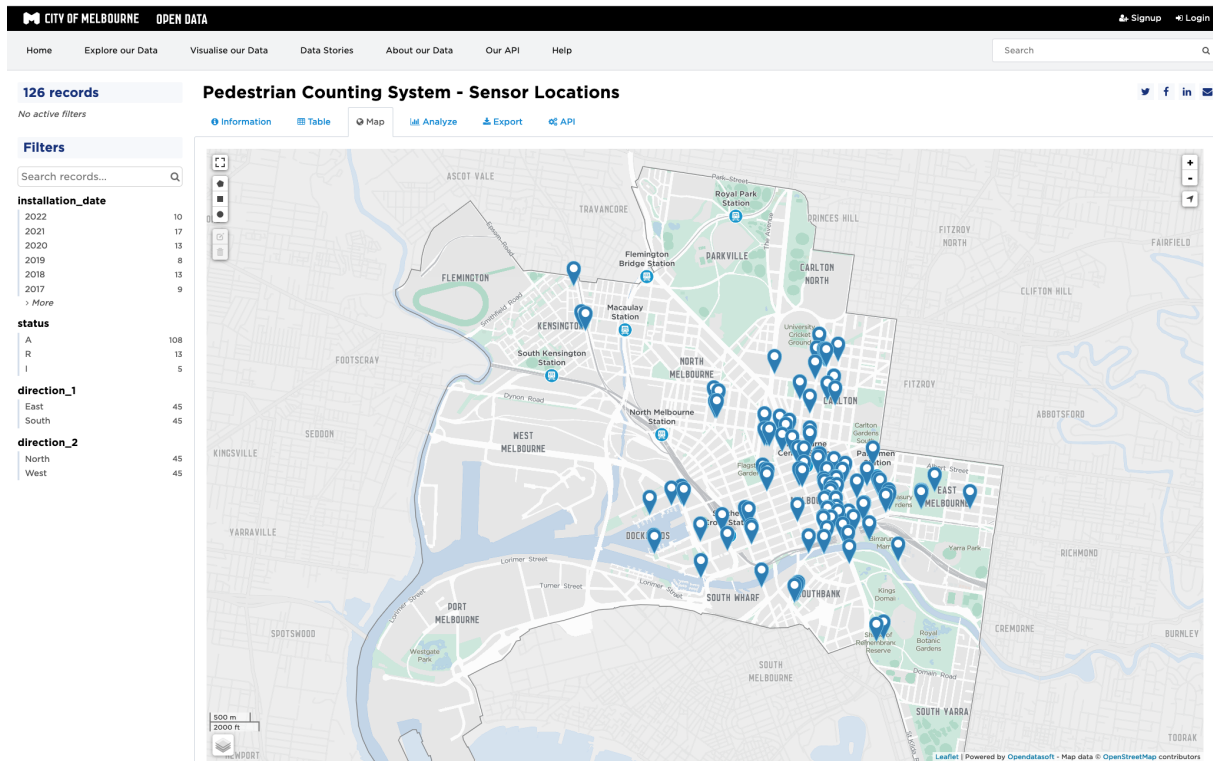


Figure 2. Locations of the pedestrian sensors in the City of Melbourne. Map generated directly from the Melbourne Open Data Portal (<https://data.melbourne.vic.gov.au/explore/dataset/pedestrian-counting-system-sensor-locations/map/>).

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (3)$$

where \hat{y}_i is the model prediction for item i , y_i is the actual value and n is the total number of data points.

3.2 Feature Importance

A longer-term aim of the work is to better understand the impact that different contextual factors (e.g. weather, built environment, etc.) will have on the size of the ambient population. ‘Feature importance’ can reveal information about the features that have the strongest impact on a prediction, and hence which are the most important in driving changes in the dependent variable. The ‘default’ feature importance method for the RandomForestRegressor model in scikit-learn is the ‘impurity-based’ method, but we avoid this because it has been shown to artificially inflate the importance of numerical features². Instead we use ‘permutation importance’ which works by iteratively

²https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html

removing particular features from the model and examining the loss in predictive power that results. Features that have little impact on the quality of the model’s predictions can be considered less important. (Note that the results of the feature importance analysis will be reported in future work).

4 Data and Software Availability

All of the data are publicly available on the Melbourne Open Data Portal. In addition, the code to analyse the data and build the model are available on GitHub (<https://github.com/masher92/footfall/>). At the time of writing the code still requires some additional documentation and testing before it can be considered ‘complete’ and fully reproducible.

5 Results

5.1 Model evaluation

The Random Forest Regressor outperforms the linear regression and XGBoost in terms of the MAE and the RMSE (Table 1). It is thus selected as the most appropriate model for use in the remainder of this analysis, and is hereafter referred to as *the model*. Figure 3 compares the model’s predicted counts-per-hour to the real counts in the pedestrian

data for *all* data points. Although there is some natural variation in the predicted values, most of the predictions fall on the diagonal ($x = y$) or near it, so we can be confident that the model is not biased towards large or small counts.

Table 1. Error metrics calculated on the predicted values (counts of pedestrians per hour) from 10-fold cross-validation of each model against actual values from the sensor data

	MAE	RMSE
Linear Regression	268.40	370.54
Random Forest Regressor	89.88	179.62
XGBoost	121.35	207.40

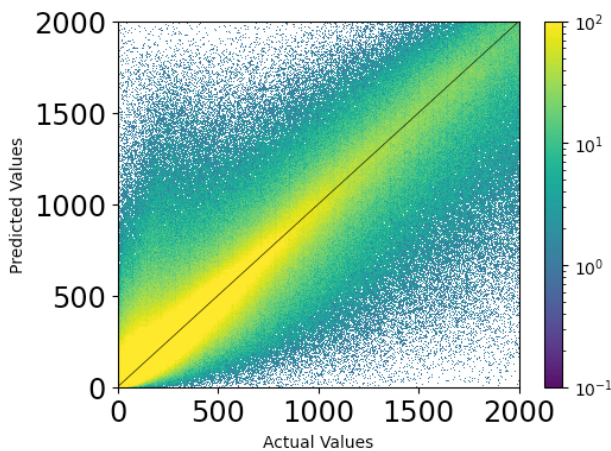


Figure 3. Predicted values (counts of pedestrians per hour) plotted against actual values from the sensor data.

Ultimately we intend to use the model to (i) explore the impact that different factors will have on the size of the ambient population and (ii) predict the size of the ambient population under different conditions. There is insufficient space to cover (i), so we report on the preliminary results with regards to the model error (ii).

Whilst the preceding analysis has considered *all* data points, it is highly likely that the prediction error will vary both temporally and spatially, and so analysis of the model's prediction error is further broken down as such in Section 5.2 and Section 5.3.

5.2 Temporal Variations in Prediction Error

The prediction error may vary depending on the time of day and day of week. For example, perhaps the size of the ambient population is easy to predict during the middle of the day on a weekday as many people take part in their regular employment-related activities, but behaviour on weekends or evenings may be much more sporadic.

To this end Figure 4 plots the mean pedestrian count per hour over seven days (Monday - Sunday) from the observation data as well as the MAE and MAPE in the model predictions.

The graphs of the mean count (from the real data) show a typical pattern for a city centre; there are activity peaks in the morning (commuting), midday (lunch time) and afternoon (commuting) and visibly different patterns on Saturday and Sunday. As might be expected, absolute predictions follow a similar pattern to the mean data; broadly the larger the counts the larger the associated errors. Interestingly, however, the mean absolute percentage error – that should be relatively stable if the model was able to predict all time periods equally well – is larger in the night-time hours. This suggests that behaviour in these times is harder to predict from the factors provided to the model (weather, built environment, etc.), and implies that perhaps the environment might not be driving pedestrian behaviour during those hours in the way it does at other times.

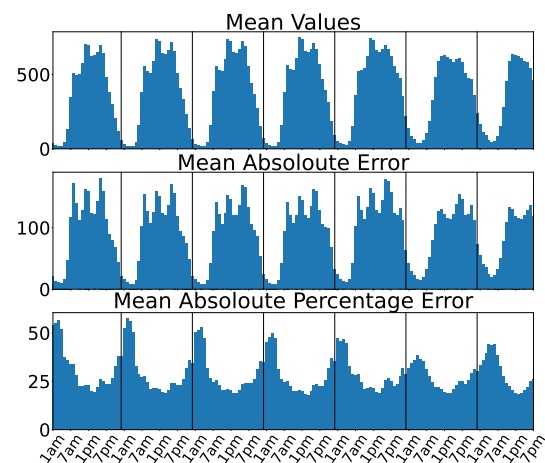


Figure 4. Temporal distribution in the mean number of pedestrians per hour over an average week and associated errors.

5.3 Spatial Variations in Prediction Error

We might expect the model prediction error to be consistent across all sensors, but this is not the case. Figure 5 illustrates the mean count per hour (real data) and the MAE and MAPE as per (1) and (2) respectively. The sensors in the central and southern parts of the city centre clearly capture the largest pedestrian flows, but these sensors do not necessarily exhibit higher or lower error as a percentage. Interestingly there are sensors in the eastern and western parts of the city centre that have particularly high percentage errors. In this case it is likely that their locations cause unusual patterns in visitor behaviour. Although further investigation into the precise locations of these sensors is needed before any conclusions can be drawn with regards to the predictability of particular locations, the maps have revealed interesting patterns that are worthy of further investigation.

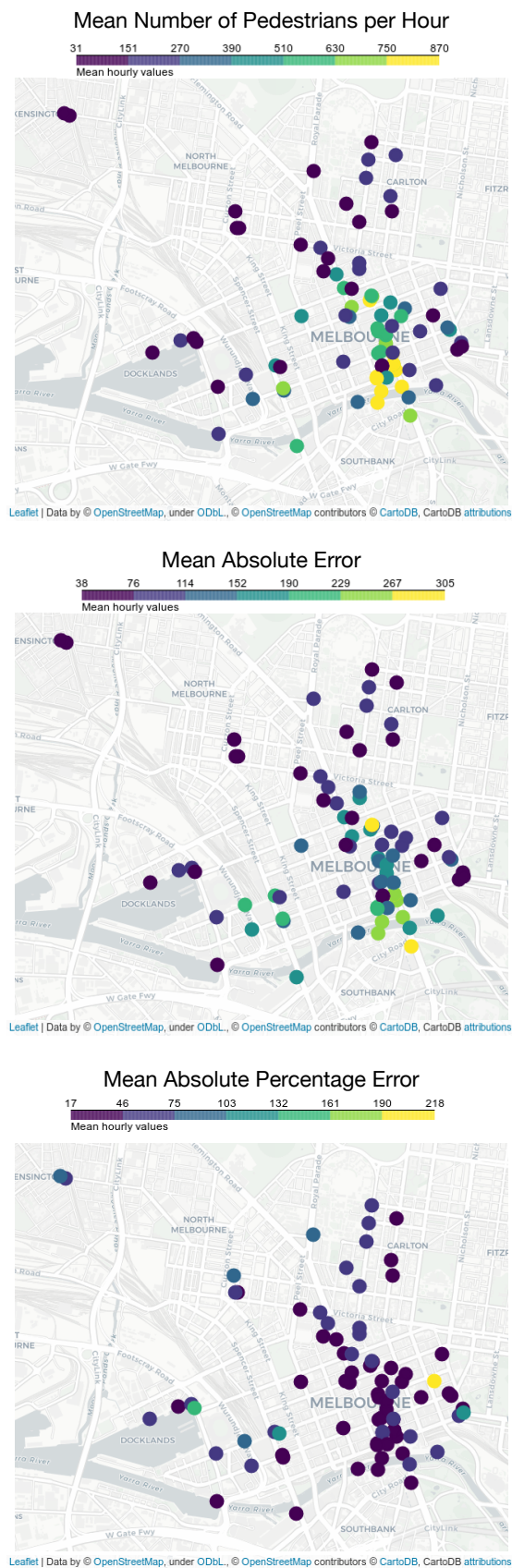


Figure 5. Spatial distribution in the mean number of pedestrians per hour and associated errors.

6 Discussion and conclusions

This paper has presented ongoing work to build a machine learning model that is able to predict the number of pedestrians at different locations of a city at different times and therefore also sheds light on the contextual factors that might drive pedestrian behaviour and urban dynamics overall.

Our approach, which utilizes on-site sensors and machine learning, presents an alternative to previous approaches and potentially exhibits several advantages. For example, previous work made use of mobile devices and location-disclosure to estimate the ambient population and predict crime (Kadar et al., 2017). In comparison, our approach is less reliant on sensitive information and also not limited by the share of people who do not wish to disclose their location data in the first place. Data characteristics and privacy concerns are especially relevant for real-time estimation of the ambient population such as during emergency moments (e.g. floods) and special events (such as festivals). Relying on public cameras/sensors may provide higher accuracy in real-time, while interfering less with the privacy of citizens. Another approach to understand and estimate the ambient population has been iterative agent-based modelling of daily mobility patterns. For example, agent-based models have been applied to dissect footfall data by making minimal assumptions about who constitutes the footfall (Crols and Malleon, 2019). While an iterative agent-based model is of theoretical and explanatory merit, it is computationally expensive and laborious. In contrast, applying machine learning algorithms, like random forest, to the open-source sensor data, does not require myriad assumptions but directly derives a usable model from the data and highlights the most relevant drivers of footfall which in turn also allows characterization of the ambient population and their behaviour.

Of course our approach comes not without limitations. We have seen that there are spatial and temporal variations in prediction error. For instance, the error is generally larger at night time. While this in part may be due to irregularity of nightly events taking place, it might also suggest that the sensor information is generally less reliable at night. Constrained and varying visibility, due to darker light conditions and other nightly interfering light sources, might constitute a general issue with the sensor data, but this speculation warrants further investigation. If true, however, other approaches such as the private device-based estimations discussed above, might be especially relevant to make up for those shortcomings at night. Moreover it is not fully clear if all the measurements taken by the sensors are mutually exclusive or if there is overlap and double-counting.

Further limitations arise from the choice of model and algorithms. While it can be a strength of decision-tree based models, such as random-forest and XGBoost, to arrive at predictive models directly from the data, this also implies dependence on specific data in our understanding of

urban dynamics. For instance, without conducting more case-studies across several cities we do not have sufficient information on whether the structural and environmental drivers of footfall and ambient population ascertained here are Melbourne-specific or can be generalised to other urban environments.

Therefore, in terms of future research, we intend to apply the model to alternative cities, where data are available, to try to understand whether the influence of the different contextual factors varies. Once sufficiently robust, we also intend to evaluate the success of the model with respect to special social events that should cause unusual patterns in pedestrian activity, such as concerts or festivals.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 757455), through a UK Economic and Social Research Council (ESRC) Future Research Leaders grant [number ES/L009900/1], and through an internship funded by the UK Leeds Institute for Data Analytics (LIDA).

References

- Bafna, S.: Space Syntax: A Brief Introduction to Its Logic and Analytical Techniques, *Environment and Behavior*, 35, 17–29, <https://doi.org/10.1177/0013916502238863>, 2003.
- Boivin, R. and Felson, M.: Crimes by Visitors Versus Crimes by Residents: The Influence of Visitor Inflows, *Journal of Quantitative Criminology*, <https://doi.org/10.1007/s10940-017-9341-1>, 2017.
- City of Melbourne Open Data Portal: Pedestrian Historical Data, 2023.
- Crols, T. and Malleson, N.: Quantifying the ambient population using hourly population footfall data and an agent-based model of daily mobility, *GeoInformatica*, 23, 201–220, 2019.
- Kadar, C., Rosés Brüngger, R., and Pletikosa, I.: Measuring ambient population from location-based social networks to describe urban crime, pp. 521–535, 2017.
- Leccese, F., Lista, D., Salvadori, G., Beccali, M., and Bonomolo, M.: On the Applicability of the Space Syntax Methodology for the Determination of Street Lighting Classes, *Energies*, 13, 1476, <https://doi.org/10.3390/en13061476>, 2020.
- Malleson, N. and Andresen, M. A.: Exploring the Impact of Ambient Population Measures on London Crime Hotspots, *Journal of Criminal Justice*, 46, 52–63, <https://doi.org/10.1016/j.jcrimjus.2016.03.002>, 2016.
- Panczak, R., Charles-Edwards, E., and Corcoran, J.: Estimating Temporary Populations: A Systematic Review of the Empirical Literature, *Palgrave Communications*, 6, 87, <https://doi.org/10.1057/s41599-020-0455-y>, 2020.
- Park, Y. M. and Kwan, M.-P.: Individual Exposure Estimates May Be Erroneous When Spatiotemporal Variability of Air Pollution and Human Mobility Are Ignored, *Health & Place*, 43, 85–94, <https://doi.org/10.1016/j.healthplace.2016.10.002>, 2017.
- Whipp, A., Malleson, N., Ward, J., and Heppenstall, A.: Estimates of the Ambient Population: Assessing the Utility of Conventional and Novel Data Sources, *ISPRS International Journal of Geo-Information*, 10, 131, <https://doi.org/10.3390/ijgi10030131>, 2021.