Machine Learning for Predicting Pedestrian Activity Levels in Cities

Achituv Cohen*, Sagi Dalyot*, Asya Natapov**
* Technion - Israel Institute of Technology, Israel
**Loughborough University, United Kingdom

Abstract. Analysing and modelling pedestrian activity in built environments allows us to understand, assess, predict, and manage its dynamics. Nonetheless, pedestrian activity data might not be available everywhere. An alternative can suggest predicting pedestrian activity by considering environmental characteristics and the geometrical configuration of the environment. This paper presents a Machine Learning pedestrian activity level prediction model, which is trained and tested using data extracted from smart city sensor systems from multiple cities. The proposed model was applied to Greater London, UK, and the prediction results were compared with pedestrian activity data provided by Transport for London. Our results show that the model has high potential to predict pedestrian activity levels in a city, but that further research is needed to obtain more reliable results.

Keywords. Machine Learning, Pedestrian Activity, Spatial Analysis, Crowdsourcing

1. Introduction

Diverse digital technologies and sensors are used today in smart cities to collect different types of data for better city management aimed to improve citizens' quality of life. The Hystreet platform[1], for example, monitors Pedestrian Activity (PA) by continuously counting the number of pedestrians, using laser scanners positioned on building facades. By modelling and analysing PA, city officials can better predict and manage city traffic and understand the resultant movement dynamics (Duives et al. 2015). Dynamic PA data can assist city officials and improve pedestrian routing services by

[1] https://hystreet.com/en/locations#/

Published in "Proceedings of the 16th International Conference on Location Based Services (LBS 2021)", edited by Anahid Basiri, Georg Gartner and Haosheng Huang, LBS 2021, 24-25 November 2021, Glasgow, UK/online.

https://doi.org/10.34726/1758 | © Authors 2021. CC BY 4.0 License.

suggesting custom routes for pedestrians wishing to avoid empty streets for safety reasons or overcrowded areas for health reasons. Still, these sensors are sparse, and hence PA data is very limited, forcing cities to rely on site and periodic (household) surveys, which are limited and expensive. Research shows that PA can be predicted by analysing the city structure and its features that represent the urban form (Qin 2016, Omer et al. 2015). These, both static and dynamic, can be retrieved from geospatial catalog, such as OpenStreetMap (OSM) (Cohen & Dalyot, 2020). We propose a Machine Learning (ML) prediction model, which is trained and tested using smart sensor data that counts PA from different cities: the Hystreet platform in Germany and Bluetooth (BT) sensor network in Tel Aviv, Israel. The prediction model reveals the relationships between the urban structure and features and the PA in these areas, and then is applied to new areas.

2. Methods

The proposed supervised ML prediction model uses as features OSM's street segments and different city elements, including their attributes. As model labels, the smart sensors' PA counts corresponding to each street segment. PA counts are classified to five categories (*Table 1*), or levels, based on the work of Helbing and Johansson (2009).

PA Levels	Model Labels	Density range
< 0.7 m/p	5	Highest
< 0.95 m/p	4	
< 1.2 m/p	3	
< 1.5 m/p	2	
> 1.8 m/p	1	Lowest

Table 1. PA levels in units of meters per person, according to Helbing and Johansson (2009).

Figure 1 illustrates the implemented ML workflow. OSM street network was downloaded and transformed into a walkable streets graph by examining the OSM highway tags of the segments to retain pedestrian streets only. *Table 2* depicts the feature engineering used in the model: (1) city features, which are derived from OSM tags; (2) centrality features, which are calculated based on OSM's street network; and (3) time-related features. For the city features we adopted the works of Qin (2016) and Omer et al. (2015) that list city and spatial features that effect PA. Centrality features are generated using graph-based measures (Cooper 2015) that help identify the most central street segments within a particular area.

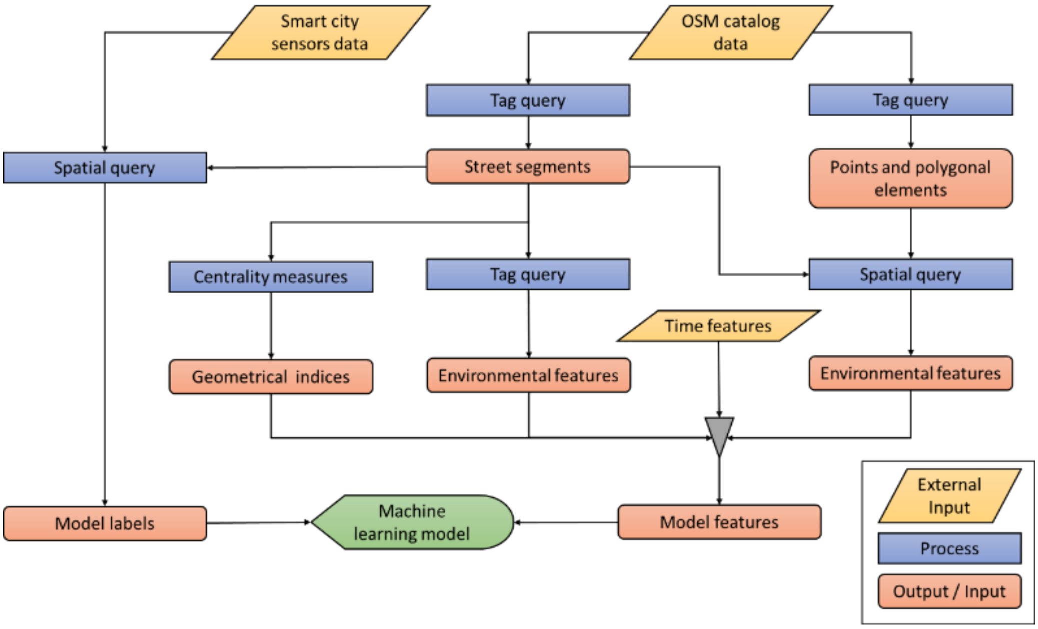

Figure 1. ML prediction model workflow.

Feature Group	Model Features	Calculation method
City Features	Highway	Categorical values based on OSM's highway tag.
	Land use	Categorical values based on OSM's land use tag or nearby element's tag (e.g., residential, retail).
	Amenity	The number of elements with the same tag that are within 20 meters from the street segment.
	Office	
	Tourism	
	Shop	
	Building	
	Natural	
	Leisure	
Centrality Features	Betweenness	Street centrality indices.
	Closeness	
Time Features	Hour	Hour of the day.
	Day	Day of the week order (for example, Sunday=1, Monday=2 ...).

Table 2. Feature engineering and calculation method.

PA label data (pedestrian counts translated to levels) were retrieved from two smart monitoring systems: 1) The Hystreet platform, which continuously counts pedestrians on streets in cities throughout Germany; 2) A BT sensor network, located in Tel-Aviv, Israel, consisting of 65 point-to-point BT sensors, and 78 street segment streets (links) between adjacent sensors. PA is measured in meters per person, while the systems provide the number of people per hour. Therefore, PA is calculated (Equation 1) by dividing the street segment length with the number of people per hour multiplied with the time it takes them to cross the street (speed is defined as the commonly used average walking speed, which is 3,000 meters per hour):

$$PA = \frac{street\ segement\ length}{\left\lfloor number\ of\ people\ per\ hour * \left[\frac{street\ segement\ length_{[meter]}}{pedestrain\ speed_{\left[\frac{meter}{hour}\right]}}\right]_{[hour]} \right\rfloor} \left[\frac{meter}{person}\right] \qquad [1]$$

Our prediction model is based on the Random Forest (RF) classifier. RF combines tree predictors in such a way that each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. Cross validation was conducted on the data to resolve overfitting, with 70% of the samples used for training, and 30% for testing.

3. Preliminary Findings

The ML prediction model includes 141,384 training samples: 83,544 from 9 German cities (Hystreet data), and 57,840 samples from Tel-Aviv (BT data). All cities represent heterogeneous street arrangements that include variety of urban morphologies, land uses and settings. *Table 3* presents the resulting confusion matrix based on 27,379 test samples and label (PA level) predictions. Despite the high F1-score and accuracy values, it should be noted that labels are not evenly distributed, where most samples (90%) belong to Label 1 (least dense streets), thus the prediction of this label is very accurate. However, although the recall and precision values for the other labels is between 44% and 67%, most of the errors are predictions of adjacent labels, indicating the reliability potential of the resulting prediction model. As an example, of the 927 samples from Label 3, 604 were correctly classified (65%), while 282 samples (30%) were classified as Labels 2 and 4, and only 41 samples (less than 5%) were classified as Labels 1 and 5.

		Predicted Labels						
	Labels	*1*	*2*	*3*	*4*	*5*	*Number of samples*	Recall
True Labels	*1*	**24624**	163	74	22	4	24887	98.9%
	2	197	**333**	191	34	4	759	43.9%
	3	37	147	**604**	135	4	927	65.2%
	4	7	8	127	**390**	47	579	67.4%
	5	12	2	5	79	**129**	227	56.8%
	Number of predicted samples	24877	653	1001	660	188	**F1 score**	**99.0%**
	Precision	99%	51%	60%	59%	69%	**Accuracy**	**95.3%**

Table 3. Prediction model confusion matrix.

As PA levels rely on a set of features, which are considered universal, we further evaluate the prediction model on unseen data – Greater London, UK, and generated PA level values for the entire street network. As reference, we use PA level values provided by the Transport for London (TfL) that documented six years of data consisting of 300,000 walking trips. TfL's PA data is organized as 15,477 hexagons covering the Greater London area, each with a measurement of meters walked per square meter. *Figure 2* depicts the two PA results, showing that the centre of London is the area with the densest PA in the Greater London, and as we move away from the centre there is a clear trend of diminishing PA intensity. Other dense areas, which are located within the suburbs, mostly correspond to local shopping streets, central stations, and schools, which tend to be more crowded. Although the prediction model shows an overall resemblance to TfL's dataset, even in London's outskirts, there exists some differences between the two models (Pearson correlation = 0.487).

In conclusion, the developed PA prediction model relies on more than 140,000 samples, where testing showed a high accuracy of about 95%. Using Greater London for model evaluation, the results show robustness of the model and its potential to predict PA for new, unfamiliar areas. We believe that as we use more PA samples from different cities, the prediction model will be adjusted better to other city arrangements and characteristics, producing more reliable results. Overall, this methodology of using ML to predict PA proves to be accurate and reliable for better city management, having the potential to replace on site and periodic surveys, which are limited and expensive.

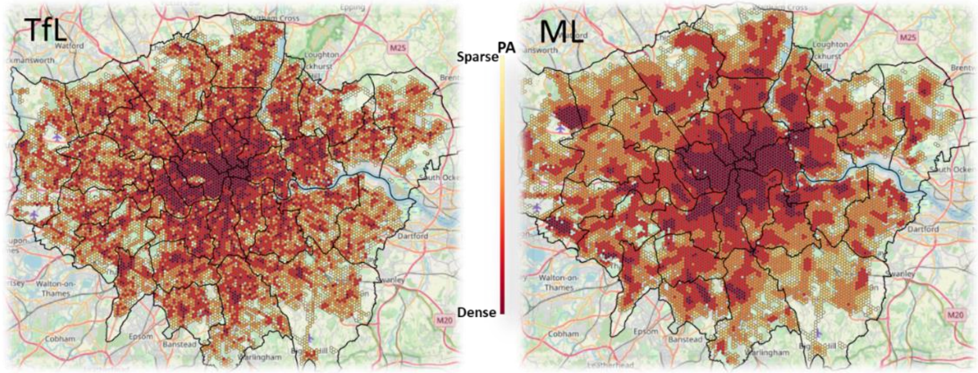

Figure 2. PA in Greater London during the weekdays according to TfL's reference model (left) and the developed ML prediction model (right).

References

Cohen A and Dalyot S (2020) Machine-learning prediction models for pedestrian traffic flow levels: Towards optimizing walking routes for blind pedestrians. *Transactions in GIS* 24(5). Wiley Online Library: 1264–1279.

Cooper CH V (2015) Spatial localization of closeness and betweenness measures: a self-contradictory but useful form of network analysis. International Journal of Geographical Information Science 29(8). Taylor & Francis: 1293–1309.

Duives DC, Daamen W and Hoogendoorn SP (2015) Quantification of the level of crowdedness for pedestrian movements. *Physica A: Statistical Mechanics and its Applications* 427. Elsevier: 162–180.

Helbing D (2009) Pedestrian, Crowd and Evacuation Dynamics. In: *Encyclopedia of Complexity and Systems Science*. New York, NY, pp. 6476–6495.

Omer I, Rofè Y and Lerman Y (2015) The impact of planning on pedestrian movement: contrasting pedestrian movement models in pre-modern and modern neighborhoods in Israel. *International Journal of Geographical Information Science* 29(12): 2121–2142.

Qin Q (2016) *Exploring Pedestrian Movement Patterns with Urban Environmental Factors in Beijing*. The Pennsylvania State University.