From explanations to predictions. Developing a predictive model of pedestrian flows on existing and planned streets.

Citation for the original published paper (version of record):
Stavroulaki, I., Ivarsson, O., Berghauser Pont, M. et al (2024). From explanations to predictions.
Developing a predictive model of pedestrian flows on existing
and planned streets.. Proceedings of the 14th International Space Syntax Symposium, SSS 2024

(article starts on next page)

64

From explanations to predictions

Developing a predictive model of pedestrian flows on existing and planned streets

Ioanna Stavroulaki; Chalmers University of Technology; gianna.stavroulaki@chalmers.se
Oscar Ivarsson; E-Commons & Chalmers Industriteknik, Chalmers University of Technology; oscar.ivarsson@chalmersindustriteknik.se
Meta Berghauser Pont; Chalmers University of Technology; meta.berghauserpont@chalmers.se
Vilhelm Verendel; E-Commons & Chalmers Industriteknik, Chalmers University of Technology; vilhelm.verendel@chalmersindustriteknik.se

ABSTRACT

Promoting sustainable mobility and urban development hinges on understanding and forecasting pedestrian movement. While empirical studies, not least in Space Syntax research, have used explanatory statistical models to identify spatial parameters influencing pedestrian movement, these models face limitations in forecasting pedestrian flows in future or data-scarce areas. Thus, they are not as useful for scenario analysis, assessment and decision making in urban design and planning. Pedestrian route-choice models are equally challenging as they are highly data demanding and depend on predictors too detailed for early design and planning stages. Instead of complex models, this study proposes a parsimonious predictive model based on street network modelling and a few spatial predictors that can be easily defined and calculated during early project stages.

The paper outlines the methodology and results of the model, which employs LASSO regression in machine learning to predict numbers of pedestrians at the street segment level. The model is trained using data gathered in Stockholm and is first tested by predicting full-day pedestrian counts at street segments of central Gothenburg. The model is evaluated both in relation to predicting the absolute number of pedestrians and their relative distribution within the area. This concise yet effective model shows promising results for early forecast of pedestrian flows in development plans and infrastructural changes and can offer a valuable tool for planners and designers to influence and optimise the distribution of pedestrian flows in various urban contexts.

KEYWORDS

Predictive modelling, Pedestrian flows, Machine learning, Space syntax, Urban design and planning

1 INTRODUCTION[1]

Pedestrian movement has always been a primary focus for urban design and planning, but its significance increased in the context of the Sustainable Development research agenda (UN Agenda 2030), particularly Goal 11 'Sustainable cities and communities'. The agenda advocates for sustainable mobility and transport (e.g., public transport, walking, cycling) emphasising its benefits for climate and health. The reduction of private car usage contributes to mitigating greenhouse gas emissions (e.g., Litman 2020), while the promotion of active modes of transport also enhances public health and well-being (e.g., Roe et al. 2020, Bird et al. 2018). Moreover, walking fosters co-presence in public spaces supporting social inclusion and cohesion (e.g., Legeby et al. 2015, Legeby 2013) and pedestrian flows stimulate local economies (e.g., Hillier et al. 1993, Hillier 1996a, Litman 2020).

Understanding how pedestrians move in the city is key in supporting sustainable urban development with urban design and planning. Appropriate methods are needed to model pedestrian flows, both to explain and predict them[2]. There is a specific demand for predictive models that can be directly applicable to urban design and planning practice, facilitating scenario analysis and early impact assessment of development plans and infrastructural changes, thereby guiding decision-making (Stavroulaki 2022). Early estimation of pedestrian movement is crucial as it is during the initial phases of a design and planning process when the fundamental structural decisions are made, such as designing street networks. It is acknowledged that the ability to influence the cost and performance of a development plan is greater in the early project phases and diminishes as planning and design progress (CURT 2004).

Methods to predict pedestrian movement are notably lacking in transport and traffic modelling which remain predominantly car oriented. Pedestrians are occasionally included in traffic modelling as 'vulnerable users' in simulations of vehicle-pedestrian interactions aimed at improving safety (e.g. Rinke et al. 2016, Pascucci et al. 2015, Obeid et al. 2017). While some route-choice models (Prato 2009) have adopted traffic-modelling methodologies to simulate

[1] This work is part of the Digital Twin Cities Centre supported by Sweden's Innovation Agency Vinnova under Grant No. 2019-00041

[2] In simple words, explanatory models explain what has already happened, and identify the relation between an outcome and given variables. Predictive models predict what has not happened yet and find the combination of factors that best forecast a future outcome (response variable).

pedestrian trajectories from origins to destinations (e.g., Basu and Sevtsuk 2022, Sevtsuk et al. 2021), they require extensive data and rely on numerous predictors that are too detailed to specify during the early design and planning stages (e.g., specific attractions, sidewalk width, street lighting, precise land-use mix), or include socioeconomic predictors not predefined in development plans (e.g., income, age). Hence, they cannot be directly applied in early design and planning phases of an urban development project, when the need for scenario analysis and decision support is higher.

Within the urban design and planning research, there is a wealth of empirical studies, not least within the Space Syntax field (e.g. Stavroulaki et al. 2019, Berghauser Pont et al. 2019a, Bolin et al. 2021, Osbil et al. 2011, 2015, Dhanani and Vaughan 2016, Netto et al. 2012; Peponis et al. 1997, 2008; Hillier et al.1993; Penn et al. 1998, Berghauser Pont and Marcus 2015), employing explanatory statistical models to identify the spatial parameters that significantly impact how pedestrians move[3]. These studies aim to understand both individual route choices (e.g. Hiller and Iida 2005, Turner and Penn 2002, Hanna 2021, Conroy Dalton 2003) and aggregated flows of movement in the city. Their main objective is to test the significance of specific parameters, primarily of the street network (e.g. network centrality, connectivity, reach) for pedestrian movement, as well as their combined effect with other spatial variables, for instance, built density, land use and accessibility to attractions (e.g. Stavroulaki et al. 2019, Berghauser Pont et al. 2019a, Bolin et al. 2021, Dhanani and Vaughan 2016, Ozbil et al. 2011, 2015). However insightful, these explanatory models are not directly applicable in predicting pedestrian flows in future areas or in existing areas lacking real-world data for model validation. Thus, they are not as useful for scenario analysis and impact assessment in urban design and planning.

Real-world data on pedestrian movement in urban environments are scarce and usually coarse in spatial and time resolution. Various methods exist (Dong et al. 2020), such as video recording with image recognition or tracking wi-fi mobile phone signals, but they are limited due to the general data protection regulations (GDPR), national restrictions, high post-processing demands for anonymisation and calibration and high financial cost. Even taking these limitations aside, these measurement techniques cannot be used in planned or newly designed places.

Considering the identified need for predictive models suitable for the initial design phases of an urban development project, alongside the highlighted constrains of current methodologies, this paper presents a predictive model that builds on the findings of previous empirical studies (e.g.

[3] Examples of empirical studies outside the Space Syntax field are Sealens et al. 2003, Moudon et al. 2019, 2007.

Stavroulaki et al 2019, Berghauser Pont et al. 2019a, Bolin et al. 2021) and the explanatory models that were developed.

The paper outlines the methodology and results of the model, which employs LASSO regression in machine learning to predict numbers of pedestrians (i.e. pedestrian counts) on the street segment level. The predictive model is based on street network modelling and relies on a few spatial predictors that can be easily defined and calculated during early design and planning stages. The aim was to, instead of a complex model, build a parsimonious model, that is a model that accomplishes a desired level of prediction with as few predictor variables as possible. To assess the generalization of the model, we fitted the model on data collected in Stockholm and tested it by predicting pedestrian counts in Gothenburg. This way we could test if the 'learned' parameters from the built environment features on one city could be transferred and applied on another.

The structure of the paper is as follows: Chapter 2 details the study's setup, the methodology of model construction, the response variable and the predictors, and the evaluation method. Chapter 3 presents the results, including the model's performance, the distribution of errors, the model equations and coefficients, and further tests. Chapter 4 concludes the study and discusses the implications, potential uses, and further improvements.

2 DATASETS AND METHODS

2.1 Overall methodology. Model set-up.

First, the model was fitted using data gathered in 19 different areas (224 street segments) in Stockholm, Sweden, including real-world data on full-day pedestrian counts collected in October 2017 (see section 2.3.). Then the model was tested by predicting full-day pedestrian counts on 75 street segments in central Gothenburg. The model was validated against real-world observations collected on 75 street segments of Gothenburg in November 2018 (see section 2.4.2). Since the Gothenburg real-world data contained only observations on street segments from the central part of the city, we trained two different models for comparison; one was trained in all 224 street segments distributed in 19 neighbourhoods in Stockholm and one was trained in 121 street segments of 6 central neighbourhoods (areas 9-14) (Figure 1). This way, we test if by controlling for the general area type (i.e. city centre) we can improve the predictive performance of the model. Previous studies have shown that, for instance, the built density type of an area (e.g. compact high-rise, spacious low-rise) is a defining factor for the number pedestrians found on its streets (Berghauser Pont et al. 2019a).

In the following sections a detailed description of the model and the training and validation data (predictor variables, response variable) is provided.

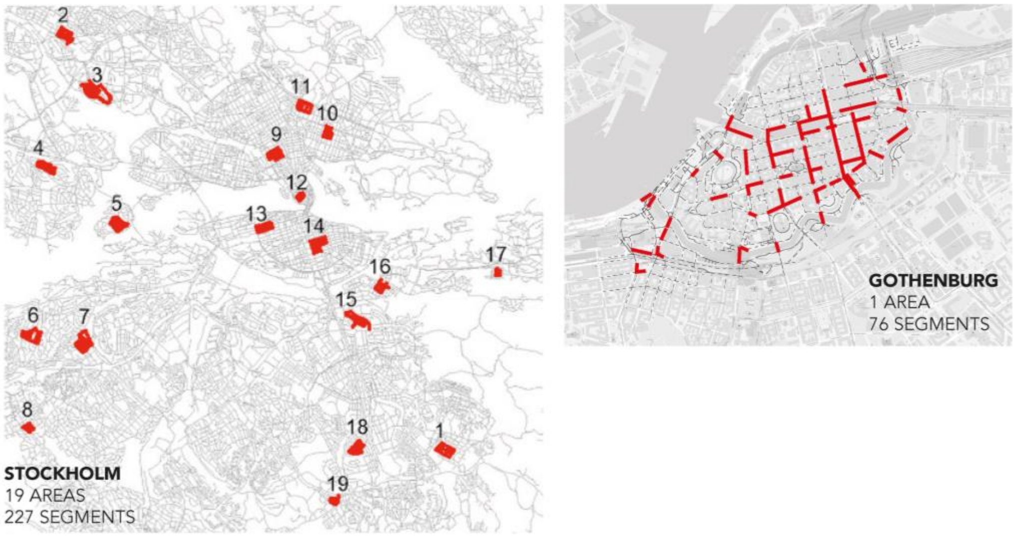

Figure 1: Neighbourhoods included in the training data from Stockholm (a), Street segments included in the model validation in Gothenburg (b).

2.2 Network model.

All variables were calculated for line-segment maps (Stavroulaki et al. 2017) of the non-motorised street networks of Stockholm and Gothenburg. The non-motorised street network includes all streets and paths that are accessible for pedestrians, including those shared with vehicles and bicycles. Streets where walking is forbidden, such as motorways or high-speed tunnels, were not included. The line-segment maps were produced based on road-centre-line maps. The same editing procedures[4], were used for creating the line-segment maps for Stockholm and Gothenburg to ensure comparability and consistency. Place Syntax Tool (PST) was used for all editing[5]. The network datasets, including documentation, are accessible via the Swedish National Database (SND) (Stavroulaki et al. 2020a, 2020b).

[4] This process, before the segmentation of the road-centre-lines to line-segments, included removing duplicate and isolated lines, snapping and generalizing.

[5] PST is an open-source plugin for QGIS. PST documentation (Stavroulaki et al. 2023) is available at https://www.smog.chalmers.se/pst. The code is available at GitHub (https://github.com/SMoG-Chalmers/PST)

2.3 Training data. Stockholm

2.3.1. Predictors

43 spatial variables were initially included in the training data. The aim was to provide a large sample of relevant variables that can be calculated during the early urban design and planning phases, from which the final predictors would be selected. Among them are variables which have shown significant correlations to pedestrian flows in earlier studies, such as Angular Betweenness centrality and Angular Integration (e.g. Stavroulaki et al. 2019, Dhanani and Vaughan 2016, Hillier and Iida 2005), built density and accessibility to attractions as local markets and public transport stops (Berghauser Pont et al. 2019a, Stavroulaki et al. 2019). Besides continuous variables, categorical variables were also included, such as official road classifications used in transport planning and non-motorised multiscalar centrality types (Berghauser Pont et al 2019a).

Place Syntax Tool (PST) was used for all calculations. Below is a detailed description of all the variables, with the abbreviation in brackets:

2.3.1.a. Angular Betweenness Centrality (Bet500-5000)

Angular Betweenness centrality (Hillier et al. 2012, Hillier and Iida 2005, Turner 2007) was calculated in 10 radii, ranging from 500m to 5000m walking distance with intervals of 500m. The radii were chosen to align with the more local scales of pedestrian movement[6], whereas the small interval of 500m ensures a consistent, uniform, and continuous sampling of centrality. The following equation is used:

$$B_{(x)} = \sum_{s \neq x \neq t} \frac{\sigma_{st}(x)}{\sigma_{st}} \qquad (1)$$

where s and t are all nodes (i.e. street segments) in the network different from x

σst = the number of shortest paths from s to t.

σst(x) = the number of shortest paths from s to t that pass-through x

The shortest path from s to t is defined as the shortest angular distance (i.e. accumulated degrees turned).

2.3.1.b. Angular Integration (Int500-5000)

Angular Integration (Hillier and Iida 2005) was calculated for the same 10 radii as Angular Betweenness centrality. The definition of shortest path is again based on the shortest angular distance. The following equation was used:

[6] To reduce the possible "boundary effect" the area which was analysed was at least 5km larger than the area of the study in all directions.

$$AI_{NAIN}(x) = \frac{N^2}{1+\sum_{i\neq x} D(x,i)} \quad (2)$$

where

N = node count or number of reached nodes (including origin node)

D(x, i) = angular depth of i in relation to x. Angular depth is defined as the accumulated degrees turned divided by 90.

2.3.1.c. Accessibility to Local markets (LMarkets_500, LMarkets_Str) and Public transport stops (PubTr_500, PubTr_Str)

To capture both the number of individual attractions on the street that could potentially make it a destination point for pedestrian movement, but also the general number of attractions on each street's immediate local context, which could make it a potential thoroughfare between further destinations, we included two measures for each attraction: first, the number of attractions on each segment and second, the number of attractions accessible within walking distance 500m from each street segment. The list of attraction variables is thus as follows: Accessible Local markets[7] within 500m walking distance from each line-segment (LMarkets_500), Number of Local Markets on each line-segment (LMarkets_Str), Accessible Public transport stops within 500m from each line-segment (*PubTr_500)* and Number of Public transport nodes on each line-segment (*PubTr_Str)*.

To calculate Accessibility within 500m, the cumulative-opportunities accessibility measure (Heyman et al. 2019) is used, following the Place Syntax methodology (Ståhle et al. 2005). The distance threshold of 500m is used as it is commonly recognized to be one that most people are willing to walk (Gehl 2010). The following equation was used[8]:

$$AR_{(o)} = \sum_{a\in A} \Big(f(a)w\big(D(o,a)\big)\Big) \quad (3)$$

where

A = the set of reachable attractions (i.e. local markets, plots) within given radius,

f(a) = attractions value associated with attraction a, or 1 if not attraction value. No attraction value f(a) is used in this case as we aim to count the actual number of points.

D(o,a) = shortest walking distance from origin o to attraction a,

w(x) = attenuation function (no attenuation function is used in this case)

No attraction value f(a) is used in this case as we aim to count the actual number of points.

[7] i.e. all ground floor retail shops, services, restaurants, and cafes

[8] 'Attraction Reach' function in PST

The datasets of attractions are extracted from the Points of Interest datasets (POI) of Open Street Maps[9].

2.3.1.d. Built density (FSI_500, GSI_500)

Built density is often used as a measure of attraction that can generate movement due to its concentration of both people and uses that can act as origins and destinations of movement. Following the work of Berghauser Pont and Marcus (2014), built density is calculated using the cumulative-opportunities accessibility measure (Heyman et al. 2019). Thus, density is not considered as an individual property of each building, but as the amount of built-up space that is accessible from every street, producing a measure of 'accessible' or 'perceived density' from the pedestrian's point of view. Accessible built density is described as the Accessible FSI (Floor Space Index)[10] and Accessible GSI (Ground Space Index) in 500m.

Equation 3 was used for the calculation with the following additions. When calculating FSI(o) the attraction value f(a) is the building's GFA (Gross Floor Area) and when calculating GSI(o) the f(a) is the building's Footprint. D(o,a) is defined as 500m walking distance from the origin (i.e. midpoints of line-segments).

Accessible FSI(o) is then calculated as follows:

$$FSI(o) = AR(o, GFA) / Area(o) \qquad (4)$$

Accessible GSI(o) is then calculated as follows:

$$FSI(o) = AR(o, Footprint) / Area(o) \qquad (5)$$

where Area (o) is calculated as the area of the convex hull, defined by the end-points of all reachable line-segments within 500m from the origin.

The dataset of building polygons including information of building heights, GFA and Footprint was available on request from the Spatial Morphology Group (SMoG)[11].

2.3.1.e. Land division (Plot_500).

Land division is described as the Accessible number of plots in 500m walking distance (Bobkova et al 2017, Bobkova 2019). As in the case of built density, we describe land division not as an

[9] Point of interest (POIs), codes23, 27 (retail, services, food and drinks) https://www.openstreetmap.org
[10] Built density is described, following the work of Berghauser Pont and Haupt (2023), by two measures: FSI (Floor space index) and GSI (Ground space index).
[11] Spatial Morphology Group, Chalmers University of Technology (smog.chalmers.se). The dataset was created in 2016 based on the geodataset 'Översiktskartan' (i.e., General map) accessible via the Swedish Land Survey Authority (Lantmateriet, https://zeus.slu.se/get/). For more information on the creation of the datasets refer to Berghauser Pont et al. (2017a,b; 2019b)

individual property of each block, but as an area-based measure and a street property. The measure is directly related to the size of the plots and the grain of the land division that has been associated with a higher concentration of pedestrian-oriented activities, active frontages and local markets (Bobkova et al 2019, Bobkova 2019, Scoppa et al. 2015). Equation 3 was used for the calculation.

The dataset of plots used was available on request by the Spatial Morphology Group (SMoG)[12].

2.3.1. f. Attraction betweenness for local markets (ATm_500-2000) and built density (ATd_500-2000)

Attraction Betweenness is a weighted Betweenness centrality measure (see Equation 1), assigning weights to the street network line-segments from a table of attractions. Each attraction point is assigned to the closest line segment and selected data is transferred from the point to the line segment. The collected scores on each line segment are then used as weight. Two weights (i.e. attractions) are used in this study, local markets and FSI (Floor Space Index) (i.e. built density, see 2.3.1.d). Two radii are used: 500m, which is a typical walking radius used in accessibility analysis (see 2.3.1.c) and 2km, which is a representative radius for local betweenness centrality analysis. Attraction betweenness has been associated with goal-oriented and attraction-oriented pedestrian movement (Stavroulaki et al. 2020c, Berghauser Pont and Marcus 2015).

2.3.1. g Accessible Population density

Following the Place Syntax methodology (Ståhle et al. 2005) population density was calculated as accessible population reached within walking distance from each line segment, again using the cumulative-opportunities accessibility measure (Heyman et al. 2019) The walking distance threshold used is again 500m. Three measures of accessible population density were calculated: Total Accessible Population (Pop_500), Accessible Residential population (NPop_500) and Accessible Working population (WPop_500).

The population datasets (100x100 grid) used were created by the SCB (Swedish Statistics Agency)[13].

[12] The dataset was created in 2017 by Bobkova (2019) based on the official property geodataset included in 'Översiktskartan' (i.e., General map) accessible via the Swedish Land Survey Authority (Lantmateriet, https://zeus.slu.se/get/). For more information about the processing of the property dataset see (Bobkova, 2019).

[13] The datasets were accessed via Swedish Land Survey Authority (Lantmateriet, https://zeus.slu.se/get/).

2.3.1. h. Street characteristics

The following variables were retrieved from the official road datasets of the Swedish Transport authority (NVDB, Nationell Vägdatabas, Sweden)[14],

- Number of lanes including car and public transport (tram, bus lanes) (TotLanes)
- Number of public transport lanes (TranspLanes)
- Number of car lanes (CarLanes)
- Speed Limit (SpeedLim)
- Segment length (SegLength) that is the metric length of the street segment, from street junction to street junction. Note that a street segment can include many line segments in the line-segment map that is used for all other calculations (i.e. network centrality, accessibility). Street length was calculated in QGIS.

2.3.1.i Street types and road classifications

Various street types and official road classifications were tested as potential predictors.

- Multiscalar centrality types (CenType_): 4 types of streets were identified by Berghauser Pont et al. (2019a) based on their multiscalar angular betweenness centrality profile[15]. The types refer to the non-motorised street network. The City type (CenType_3) includes the most central streets (high streets, main streets) whose centrality increases in higher scales; the Neighbourhood type (CenType_2) includes streets with consistently high centrality on most scales, but dropping clearly on the lowest and highest scales, acting primarily as connectors of neighbourhoods; the Local type (CenType_4) includes street segments with high betweenness centrality only on the very local scale, that are the thoroughfares within each neighbourhood; and the Background type (CenType_1) includes streets that have low centrality in all scales. Berghauser Pont et al. (2019a) found that the street types define the distribution of pedestrians within each neighbourhood, where the Built Density profile of the area define their total volume, a finding that was based on a large empirical study in 53 neighbourhoods in Stockholm, Amsterdam and London. The datasets of street types for Stockholm (Berghauser Pont et al. 2019a) were available on request by the Spatial Morphology Group.
- Funkvagclass: 9 functional road classes are defined by the Swedish Transport authority[16]. Using a hierarchy from Classes 0-3 (primary road network) to Classes 4-5

[14] accessible via https://lastkajen.trafikverket.se

[15] For the detailed methodology of the generation of street types, please refer to the original paper

[16] Dataset: Funktionell vägklass, NVDB, Nationell Vägdatabas (source: Trafikverket, https://lastkajen.trafikverket.se)

(secondary road network) to Classes 7- 9 (very local network, pedestrianized streets), roads are classified based on their importance in the connectivity of the network and their functional capacity. Given that the networks used in this study are non-motorised the higher classes 0-3 (National, International motorways) are not represented in the datasets.

- Vag typ: 7 descriptive types of urban streets are defined by the Swedish Transport authority, such as Main streets, Small Local Streets, Neighbourhood streets, Motorised thoroughfares, Parking streets[17]. Only types 2 to 5 are found in the neighbourhoods included in this study, namely Main streets, Big Local streets, Small Local Streets, Neighbourhood streets and Motorised thoroughfares.

2.3.1.j. Built density types.

6 types of built density were identified by Berghauser Pont et al. (2019a, b) based on a multivariate density profile (i.e. Accessible FSI and Accessible GSI[18]) that can be used to characterize neighbourhoods (Berghauser Pont and Haupt 2023). The types range from the spacious low-rise type (DenType_1) with low FSI and low GSI identifying villa areas, to the dense mid-rise (DenType_3) and Compact mid-rise (DenType_5) identifying city centres with high FSI and high GSI, to Spacious high-rise (DenType_6) with high FSI but low GSI identifying modernistic estates with slabs and point buildings[19].

Berghauser Pont et al. (2019a) found that the density type of a neighbourhood is highly correlated to the general volume of pedestrians found there, while the multiscalar street centrality type (see 2.3.1.i.) accounted for the distribution of that volume between the streets of each neighbourhood.

The dataset of density types for Stockholm (Berghauser Pont et al. 2019a, b) was available on request by the Spatial Morphology Group.

2.3.2. Real-world observations on pedestrian flows in Stockholm

The real-world observations of pedestrian flows included in the training data were collected by Berghauser Pont et al. (2019a). 53 neighbourhoods in Stockholm, Amsterdam and London of different density type (from high-dense urban grids to low-dense suburban areas) with varied street types (from high streets to side streets and small alleys) in areas of diverse land use mix and socioeconomic profile (from business districts and mixed-use neighbourhoods to villa areas)

[17] Dataset: Vagtyp, Categories are translated from Swedish (source: Trafikverket, downloaded at https://lastkajen.trafikverket.se)

[18] See section 2.3.1.d for the measures

[19] For the detailed methodology of the generation of street types, please refer to the original paper

were selected. In Stockholm 19 areas were observed including 227 street segments[20]. The observations include measured pedestrian counts for one day, from 7am to 10pm. The survey for Stockholm was conducted on weekdays in October 2017. The areas were not all measured in the same day but were split from Monday to Friday.

The method used was capturing anonymised Wi-Fi signals from mobile phones. Samples of Wi-Fi signals were collected when devices were searching for wi-fi networks (so called wi-fi probe requests). Each sample included a timestamp, a RSSI (Received Signal Strength Indication) and an anonymized indicator.

The Wi-Fi-signals were monitored at street junctions and were transferred in an Origin-Destination matrix, consisting of Start and End nodes. When a phone-anonymised id was captured in two adjacent street junctions then a count was added to the respective segment between this pair of junctions[21] (Start and End nodes). Both movement directions were added.

The dataset for Stockholm was available on request by the Spatial Morphology Group.

2.4 Test data. Gothenburg

2.4.1. Predictors

The same spatial variables used in the model training with Stockholm data (section 2.3) were tested as predictors for Gothenburg pedestrian counts. The editing and analysis procedure, equations and settings are the same to ensure consistency and comparability.

2.4.2. Response variable. Real-world observations of pedestrian counts

To validate the model, we used real-world observations collected in Gothenburg by the traffic office of Gothenburg municipality (Trafikkontoret 2019) in November 2018. The same method of capturing anonymised Wi-Fi signals from mobile phones was used, as in the case of Stockholm. Also, the same processing procedures were used to ensure consistency between the training and validating data[22]. Monitoring devices were placed in 50 locations in the city centre, including street junctions and bridges. The monitoring took place continuously from Wednesday 6am to Monday 12pm. Since the Stockholm real-world data included day counts from 7am to 10pm, we filtered the same time frames from the Gothenburg dataset. Also, while the Stockholm areas were each monitored for one weekday day ranging from Monday to Friday, Gothenburg centre was monitored for both Wednesday and Thursday. When testing the correlation of pedestrian counts recorded on Wednesday and Thursday, we found a Pearson

[20] For the method of selection, please refer to the original paper. The company used for collection and post-processing was Bumbee Labs, Stockholm.
[21] For the full documentation of processing please refer to the original paper.
[22] The same company, collecting and processing method were used (Bumbee labs, Stockholm).

correlation of 0,989 (p<0.001) which made the choice between days trivial. Finally, we included pedestrian counts from Wednesday from 6 am to 10 pm, as the response variable for the model validation.

Apart from following the processing methodology as described in Berghauser Pont et al. (2019a), for the purpose of this study we took one further step. Since in the case of bridges, the monitoring device was placed at the midpoint of the bridge, the pedestrian counts of that device were directly transferred to the respective street segment.

The datasets, after the final processing, included pedestrian counts for 75 street segments.

2.5 Structuring the datasets

As described in section 2.2 and 2.3, all network centrality measures were calculated for the non-motorised line-segment maps of Stockholm and Gothenburg using Angular Segment Analysis (ASA). The same line-segments were used as the origin points[23] to Equation (3) to calculate Accessible FSI and GSI, Accessible number of Plots, and Accessibility to Attractions (public transport and local markets). As a result, the datasets with the predictor variables were structured per line segment.

As described in sections 2.3.2 and 2.4.2 the datasets containing observed pedestrian were structured per street segment, meaning the street section between a pair of adjacent street junctions. These street segments often include more than one line-segments, especially in curvilinear streets. To deal with that, all values of the predictor variables were transferred from the line-segments to their respective street segments, using a proportion average function[24].

2.6 Machine learning model

The machine learning model is a LASSO regression model (Least Absolute Shrinkage and Selection Operator) that identifies a subset of predictor variables that are most strongly associated with the response variable. LASSO is a regularized regression method that constrains the sum of the absolute values of the model coefficients (i.e. penalization of coefficients). This results in a sparse model where some of the coefficients are exactly zero, indicating that they do not contribute to the prediction of the response variable.

43 predictor variables were initially used for fitting the model, as described in section 2.3. Some additional processing steps were needed before including them in the model. Some continuous variables have a highly skewed distribution (e.g. angular betweenness centrality), and they were

[23] Segment midpoints were use as the actual origin point.

[24] Proportion average takes into account the length of each line-segment.

log transformed to normalize their distribution. Furthermore, the categorical variables (CenType (1-4), DenType (1-6), FunkKlass (4-9), VagTypNo(2-5)) were one-hot encoded[25].

Since the response variable of full-day pedestrian counts is strictly positive and its distribution is right skewed, a log transformation was applied before fitting the model. Further, an inverse transformation before computing any performance statistics was done to compute the error measures in the original scale and avoid negative predictions.

The code is written in Python and is published in GitHub (https://github.com/SMoG-Chalmers/crowd-movement).

2.6.1. Hyperparameter optimization

To select the optimal set of predictor variables, we used a cross-validation approach where in each iteration one neighbourhood area from the training data was left out. Since the pedestrian counts of street segments within each neighbourhood are highly correlated, by leaving out a whole area and not just a set of random streets, we made sure that the model isn't overfitting to the training data (see Figure 2).

The optimal set of predictors was defined by the amount of regularization applied. A grid search was performed for a range of values for the regularization strength and MAPE (Mean Absolute Percentage Error) was used as the evaluation metric for the cross-validation. The value of the regularization strength was chosen based on the minimum cross-validated MAPE.

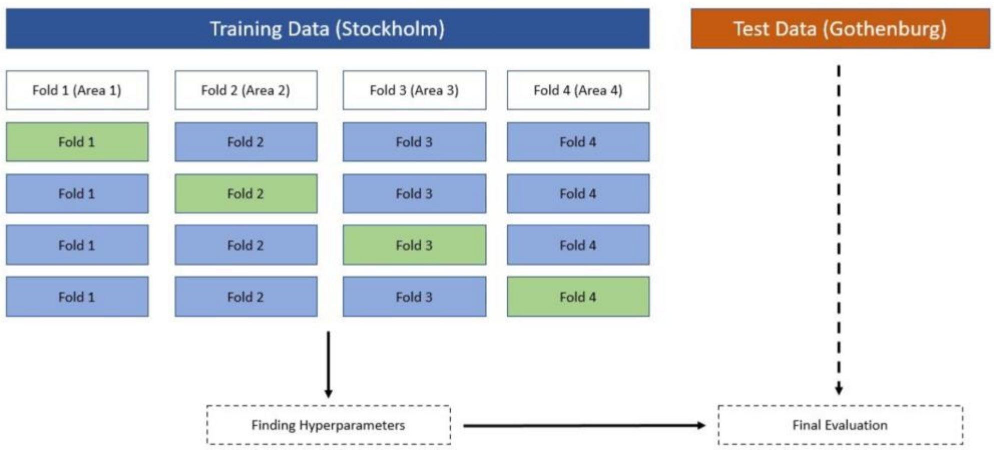

Figure 2: Method of hyperparameter optimization and cross-validation

2.6.2. Performance Evaluation metrics

Three performance metrics are used to evaluate the models: Mean Absolute Error (MAE), Median Absolute Error (Median AE) and R^2. The Absolute Error is calculated per street segment

[25] One-Hot Encoding is a common way of processing categorical variables for machine learning models. It creates a binary variable (0-1) for each category and then each feature (i.e. here each line segment) gets 1 if it falls within each category and 0 if it doesn't.

as the difference between the predicted and observed pedestrian counts. The Mean and Median[26] of the Absolute errors give an indication of the average magnitude of error in the predictions overall. The R^2 (Coefficient of determination) is computed using Equation 6. As a rule of thumb, when R^2 reaches 1 the error of the model is lower. When R^2 gets a negative score, the model performs worse than the intercept-only-model.

$$R^2 = 1 - \frac{MSE\ (model)}{MSE\ (intercept{-}only)} \quad (6)$$

where, MSE (model) is the mean squared error of the model

MSE (intercept-only) is the mean squared error of a model using only the intercept.

3 RESULTS

3.1. Model performance

As described in section 2.1. two models were tested; one was trained in 224 street segments distributed in 19 different areas in Stockholm and will be further called STHLM All, and one was trained in 121 street segments of 6 central areas and will be further called STHLM Central. The results are shown in the plots of Figure 3.

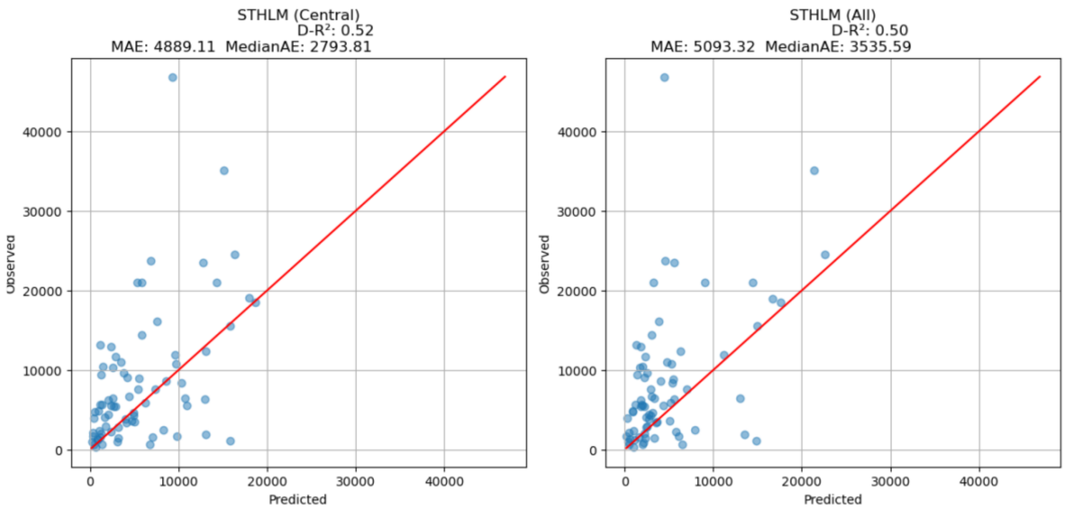

Figure 3: Plots of Observed (y axis) vs Predicted values (x axis) for the two models a. trained in 6 central areas of Stockholm, b. trained in all 19 areas in Stockholm. The values represent daily count of pedestrians, and the dots represent each of the 75 street segments tested. The results show R^2, Mean Absolute Error (MAE), Median Absolute Error (Median AE).

[26] Mean is the average. Median is the value in the middle of a data set, meaning that 50% of data points have a value smaller or equal to the median and 50% of data points have a value higher or equal to the median.

STHLM Central has an R^2 of 0.52 and STHLM All an R^2 of 0.50. Predicting around 50% of the response variable can be considered sufficient for the early-stage estimations of pedestrian counts that the models aim for. Such estimations, albeit rough, can be very useful to guide the initial design phases in an urban development plan, compare different design scenarios and strategies, highlight missteps or potentials, and identify the best ways forward. The Mean and Median Absolute Errors (MAE and MedianAE) are shown in Table 1. To evaluate the magnitude of absolute error in predicting the observed pedestrian counts, we should also consider the range of the observed values. The observed full-day pedestrian counts have Range= 46493 pedestrians, Median=5653 pedestrians and Mean=8344 pedestrians.

Table 1: Performance evaluation metrics

Model	MeanAE	MedianAE	R^2
STHLM Central	4889.11	2793.81	0.52
STHLM All	5093.32	3535.59	0.50

To fully evaluate the model performance in a manner that is relevant to urban design and planning, we should not only consider the absolute predictions, that is the absolute number of pedestrians predicted for each street, but also the prediction of the relative distribution of pedestrians in the area. In other words, do we predict accurately which streets have the highest numbers of pedestrians and which the lowest? This qualitative evaluation is very useful to inform the design process, for example to guide the allocation of pedestrian-oriented uses and active frontages, decide on the relative distribution of built densities and sketch building typologies. Figure 4 shows the relative distribution of observed pedestrian counts versus the relative distribution of predicted pedestrian counts in both models. The thicker lines indicate street segments with higher pedestrian counts and the thinner lines segments with lower counts.

Figure 4: Relative distribution of observed counts (left) vs relative distribution of predicted counts for the STHLM Central (right, top) and STHLM All (right, bottom). The legend ranges show number of pedestrians. In brackets is the number of street segments per range.

While the performance of absolute predictions can be considered as average (R^2=0.50-0.52), the prediction of the relative distribution of pedestrians on the streets is overall quite accurate for both models, however with exceptions.

In the next section, we will shed some light on the streets where the absolute predictions are highly imprecise, as well as the streets that are inaccurately predicted as to their place in the street hierarchy in relation to pedestrian flows.

3.2. Distribution of errors.

It is useful to identify which street segments are better predicted by the model and which are not. Figure 5 shows the 75 street segments coloured by the magnitude of the absolute prediction error. In black are the segments where the predictions are relatively good. The red and dark segments are the ones for which the model underpredicts (negative errors) and in blue are the segments for which the model overpredicts the pedestrian counts (positive errors).
The dark red and red segments correspond to the dots with the very high observed values in Figure 3, with higher than 25.000 observed pedestrians. These are primarily the segments clustered around the most commercial location in the city centre (i.e. Brunnsparken) with observed pedestrian counts exceptionally higher than the average (top dotted circle). Another

highly underpredicted segment is the last section of Kungsgatan, the most commercialised pedestrian street in the city centre, that is outside the Brunnsparken area. The same pattern is seen in both in the STHLM_Central and STHLM_All model. This finding is not surprising as Space Syntax research has shown that streets and locations with high commercial activity, where the 'multiplier effect' has taken place, are not well explained by spatial variables such as the space syntax centrality measures (Hillier et al. 1993). Even the models of the current study which include a larger variety of predictors, including Accessibility to Local markets, albeit all still spatial variables, fail to improve the predictions of pedestrian counts on these highly commercially activated locations. However, another potential reason could be that the OSM 'Points of interest' dataset that was used to locate the commercial activities is incomplete. A test with more accurate data on these activities is needed.

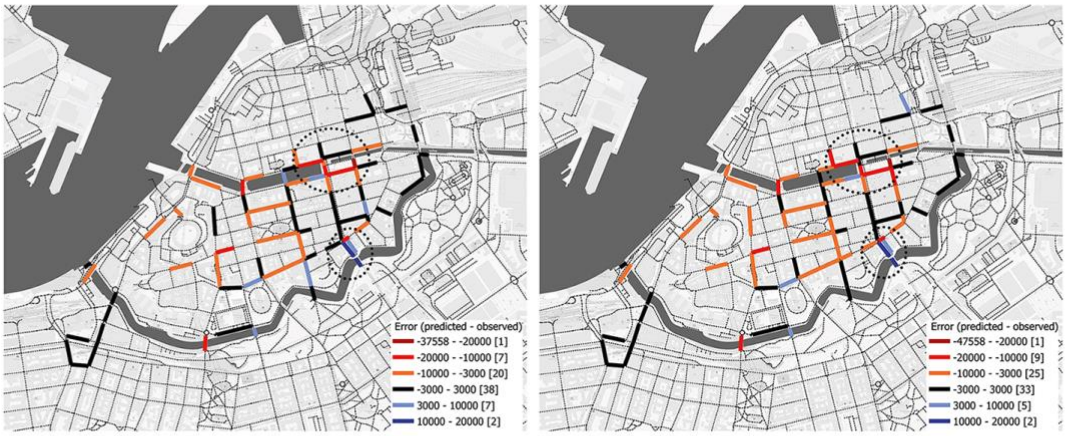

Figure 5: Mapping of the absolute errors (predicted minus observed numbers) for the STHLM Central (left) and the STHLM All (right) models. Negative values show overprediction and positive values show underprediction of pedestrian flows.

Another case worth mentioning is Kungsportsplatsen, a central square with caffes, restaurants and retail. It is surrounded by streets with strikingly different absolute prediction errors, ranging from street segments that are well predicted to highly overpredicted ones (lower dotted circle in Figure 5). Given the complex situation in this public space with a lot of different inflows of pedestrians, this inconsistency may be a result of the placement of sensors. As we see in Figure 4, the observed values show a significant drop on the lower street segments of Kungsportsplatsen (Östra Hamngatan), suggesting that the placement of sensors might have failed to capture the full number of pedestrians. The models on the contrary predict more stable values along the whole Östra Hamngatan from north to south.

The other relatively high prediction errors are harder to interpret. They may be a result of low precision of some of the spatial variables used, as will be discussed further in Chapter 4.

A central conclusion drawn from the spatial distribution of errors is that the spatial variables fail to predict the extreme observed values of pedestrians in highly activated areas (i.e. commercial centres). Future test of the models' performance would be needed in different types of areas with less commercial activities and lower pedestrian flows.

3.3. Predictors and coefficients

Apart from a sufficient predictive performance, a central aim of the models was to include a small number of predictors that could be calculated during the early stages of the design process. Figure 6 shows the optimal set of predictor variables that are included in the final models, from the initial list of 43. These predictors are followed by a blue bar showing a positive or negative association to pedestrian counts. Figures 7 and 8 show the Model equations and the different coefficients for each model.

We see that for both models local Angular Betweenness centrality (radius 2km) is a significant predictor. This is in line with previous empirical studies in the space syntax literature that have shown strong correlations of local Angular Betweenness centrality to pedestrian counts (e.g. Stavroulaki et al. 2019, Dhanani and Vaughan 2016, Hillier and Iida 2005).

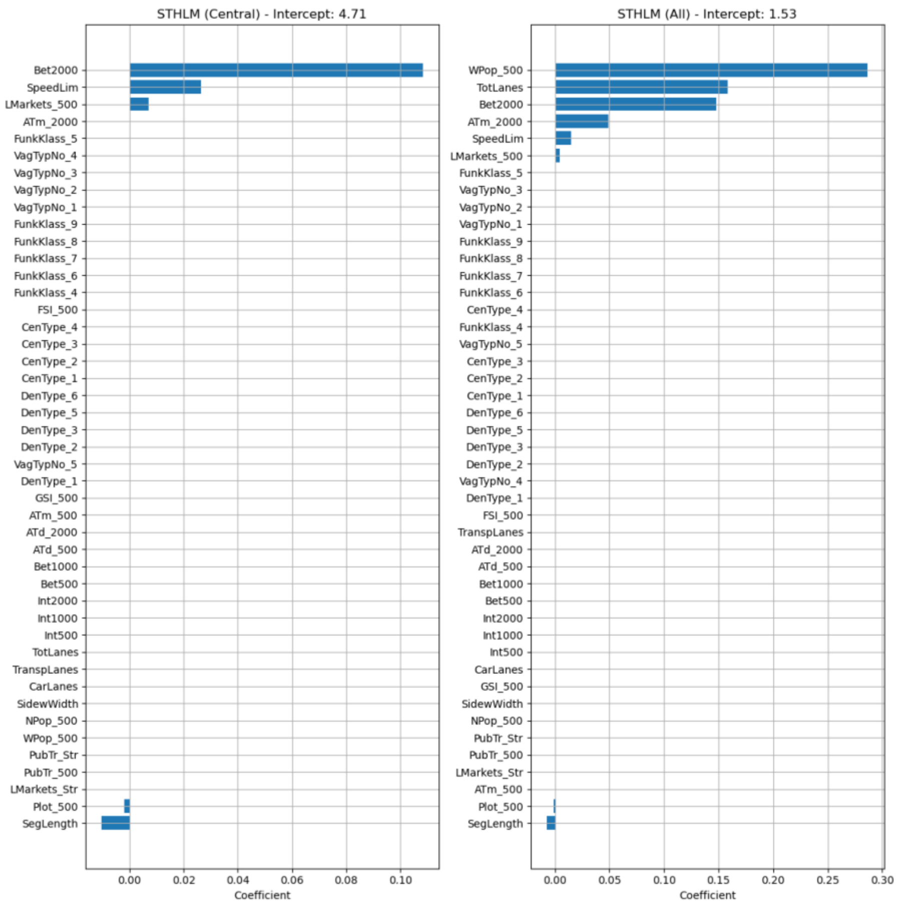

Figure 6: Initial list of variables and in blue the significant predictors included in the final models (left: model trained in 6 areas in central Stockholm, right: model trained in 19 areas in all Stockholm)

$$\ln(FullDayCount) = \beta_0 + \beta_1 Plot\_500 + \beta_2 LMarkets\_500 + \beta_3 SegLength + \beta_4 SpeedLim + \beta_5 \ln(Bet2000) + \epsilon$$

$$\beta_0 = 4.712$$
$$\beta_1 = -0.002$$
$$\beta_2 = 0.007$$
$$\beta_3 = -0.010$$
$$\beta_4 = 0.026$$
$$\beta_5 = 0.108$$

Figure 7: Model equation for the STHLM (Central) model

$$\ln(FullDayCount) = \beta_0 + \beta_1 Plot\_500 + \beta_2 LMarkets\_500 + \beta_3 \ln(WPop\_500) + \beta_4 SegLength + \beta_5 TotLanes + \beta_6 SpeedLim + \beta_7 \ln(Bet2000) + \beta_8 \ln(ATm_2000) + \epsilon$$

$$\beta_0 = 1.530$$
$$\beta_1 = -0.001$$
$$\beta_2 = 0.004$$
$$\beta_3 = 0.287$$
$$\beta_4 = -0.007$$
$$\beta_5 = 0.158$$
$$\beta_6 = 0.015$$
$$\beta_7 = 0.148$$
$$\beta_8 = 0.049$$

Figure 8: Model equation for the STHLM (All) model

Accessibility to Local Markets is also a predictor in both models, a positive association that has been reported in previous studies (Stavroulaki et al. 2019). However, in the STHLM_All model also Attraction Betweenness of Local markets (ATm_2000) is added to the significant predictors. This weighted Betweenness centrality measure identifies the very central shopping streets, since these not only have a high betweenness centrality overall, but also connect a high number of local market points. These are the most central and activated streets in both cities, as Figure 9 shows.

Figure 9: Attraction Betweenness centrality weighted for Local markets in the 6 central areas in Stockholm training data (left) and in the observed streets in Gothenburg (right)

The Speed limit (SpeedLim) is the third common predictor with positive association to pedestrian counts. Although there is a low variation of speed limits within the training and test data, there is a big distinction between suburban residential streets (30km/hour) or pedestrian only streets (5km/hour) and the typical multimodal central streets (50km/hour). The multimodal central streets are associated with higher pedestrian counts. This finding is connected to the positive association of the Total number of lanes, including cars and public transport (TotLanes) to pedestrian counts found in the STHLM All model. The number of Car lanes (CarLanes) or Public Transport Lanes (TrnspLanes) separately are not significant predictors, but their combination is. The total number of lanes is again an indication of the multimodality that characterises the main streets, that apart from attracting the highest pedestrian flows are also the main multimodal transport corridors.

The street's Segment length shows a negative association to observed pedestrian counts in both models. The segment length is also an indication of the block size, meaning that larger blocks are associated with lower pedestrian counts. Accessibility to Plots, an indicator of the degree of land division and of plot size, has a close to zero negative association for both models.

A significant predictor that is added to the STHLM_All model compared to the STHM-Central models is Accessible Working population (WPop_500). A main reason seems to be that working population is the key variable that separates the 6 central areas of Stockholm to the rest of the city as Figure 10 shows, and it captures the common area profile of central areas in Stockholm and Gothenburg. Looking at the values, Gothenburg streets have a minimum of 3500 people, whereas in Stockholm all peripheral areas, except one, fall below that threshold (55 to 2500 people). In the STHLM_Central model this distinction between centre and periphery is already built in the model, as only the 6 central areas of Stockholm are included in the training data. It is interesting to consider this in relation to the findings of Berghauser Pont et al. (2019a), where it was the Built Density type of each area that explained the higher pedestrian volumes in the central neighbourhoods in relation to the low density more peripheral ones. In that empirical study however Working Population density was not included in the explanatory variables. It is worth mentioning that Accessible FSI and Accessible Working Population are often highly correlated. For instance, in the training dataset of Stockholm used in this study, the Pearson correlation between the two variables is 0.959. This is important because in case an urban development plan does not follow a land use plan or a specific program regarding the addition of number of offices or workplaces, making the rough estimation of working population hard, then the Accessible FSI which is much easier to calculate with precision can be used instead.

Figure 10: Accessible Working population for the streets included in the training data of the STHLM_All model

3.4. Further tests to improve model performance.

Throughout the training and evaluation process of the models several limitations were identified that are important to mention. In general, these models are very sensitive to changes.

Modifying certain steps in the pre-processing phase can significantly impact their performance. One notable example is the attempt to apply standardization, which led to models with increased error rates and a tendency to select mostly binary variables. Standardization is typically recommended for LASSO regression, as regularization imposes constraints on the coefficients associated to each variable, which can be influenced by their magnitude. However, most variables were already log transformed prior to modelling, so the negative effects of not using standardisation were partially addressed.

Since one central aim of the study was to test the ability to generalize these types of models to other cities, a test was also conducted by adding data points from the test dataset (Gothenburg) to the training data (Stockholm). The hypothesis was that cities, even within a single country, have different general characteristics and that adding a small set of data points from the city that we aimed to predict, would improve the model performance. This was performed by using two different strategies, one with random selection of data points and another with intentionally selecting streets of different street types. Also, different attempts at weighing transferred data points were evaluated. Unexpectedly, all methods failed to consistently enhance the model performance and instead led to a decay in performance as more data points were included to the training data. However, further tests also showed that a model trained on only data from Gothenburg using a nested Leave-one-out cross validation method doesn't perform as well as a model trained on Stockholm data. Although this seems counterintuitive, it is an interesting finding that requires further testing. It indicates that a street segment from the Gothenburg data is more likely to share similar characteristics with a street segment in the Stockholm data, than in the Gothenburg data. This can be related to the fact that the training dataset is larger and has a larger variation of street profiles and suggests that more training data could have improved the model performance further.

4 CONCLUSIONS AND DISCUSSION

The paper presents the methodology and results of a predictive model, which employs LASSO regression in machine learning to predict numbers of pedestrian at the street segment level. The model is trained using data gathered in Stockholm and tested by predicting full-day pedestrian counts at street segments of central Gothenburg. Two model variations are presented, one trained in 19 areas of different profiles distributed across the whole metropolitan area in Stockholm and one trained in 6 central areas. The models were evaluated for their performance both in predicting the absolute numbers of pedestrians on each street and their relative distribution within the area. While their performance of absolute predictions can be considered as average (R^2=0.50-0.52), the prediction of the relative distribution of pedestrians on the streets is overall quite accurate for both models, however with exceptions. Both models rely on a very small number of spatial predictors which can be easily calculated in the early design phases of an urban development project. To build a concise yet effective model was a central aim of this study.

The results of the spatial distribution of absolute errors, as well as the results regarding the significant predictors for both models led to further interesting findings, which will be summarised and discussed below.

The spatial distribution of absolute errors showed that both models largely underpredict streets with extremely high observed pedestrian counts, which are the highly commercialised streets of the central Gothenburg area. This finding supports earlier empirical studies, not least by Hillier et al. 1993, which concluded that the spatial variables alone, particularly the network centrality measures, do not succeed in explaining the peaks in pedestrian volumes observed in highly commercial streets and locations. More studies are needed to test whether our models' performance would increase in areas with less commercial activation.

An earlier finding that is confirmed by our study is the significance of local Angular Betweenness for pedestrian counts (e.g. Stavroulaki et al. 2019, Dhanani and Vaughan 2016, Hillier and Iida 2005). Significant predictors are also the number of total lanes (cars and public transport) as well as the speed limit, in this context indicating the positive association of street multimodality to higher pedestrian flows. This is an important finding that needs to be followed up with more studies, as it has implications for both urban design and transport planning decisions. The significance of Accessibility to Local markets confirmed earlier studies (Stavroulaki et al. 2019, Bolin et al. 2021), however the significance of Accessibility to Public transport, also an earlier finding, was not corroborated. Finally, the Segment length showed a negative association to pedestrian counts, suggesting that streets around larger blocks attract fewer pedestrians, a finding that needs further testing.

The comparison of the significant predictors included the STHLM Central and STHLM All model led to further noteworthy observations. Accessible Working Population in 500m and Attraction Betweenness weighted with Local Markets are the two predictors that are added to the STHLM All model in comparison to the STHLM Central one. On the one hand, the weighted Betweenness centrality variable captured the common profile of the most central and commercially activated streets in both Stockholm and Gothenburg. On the other hand, Accessible Working Population is the definite variable that distinguished between the central and peripheral areas in the STHLM All model and, thus, captured the general area type of the Gothenburg streets. In the STHLM Central we had already controlled for the area type by including only the central areas of Stockholm in the training data. Earlier empirical studies

(Berghauser Pont et al. 2019a, Berghauser Pont and Marcus 2015) have also argued for the importance of controlling for the neighbourhood type (i.e. central commercial centre vs peripheral residential neighbourhoods) to achieve better explanations of the amount and distribution of pedestrians. While in the study of Berghauser Pont et al. (2019a) it was the Built density type of the neighbourhood that defined the area type, in this study the Accessible Working population seems to play a similar role. This is not unexpected as the two measures prove to be highly correlated and could potentially be used interchangeably as predictors. For instance, in the case of the 19 Stockholm areas their Pearson correlation is 0.959.

Further investigations are needed to increase the model performance and usefulness for urban design and planning. To start with, the model evaluation showed that adding more training data with diverse street characteristics can improve the model performance. More tests are needed regarding the underprediction of the highly commercialised streets using more accurate data on the location of commercial activities. Adding more precise data also on public transport, including the number of lines and frequencies, can give a more accurate account of its significance for attracting pedestrian movement (i.e. walk to transit) that now seems unrecognised.

Another direction for future research is to test predicting the pedestrian volume and distribution in more diverse areas, such as peripheral and suburban areas. Even further, the model should be tested in cities of different countries or cities with greater differences in morphology, population density, and geographical location, given that cities like Stockholm and Gothenburg, share many similarities in these aspects.

While this study presents a model that predicts full-day pedestrian counts, hourly observations are also available for the same areas in Stockholm and Gothenburg, that can be used to develop a model for hourly predictions in the future.

To enhance the use of predictive models of pedestrian flows in urban design and planning, a methodology to combine network-based models predicting aggregated flows, as the ones presented in this paper, to Agent Based Models (ABMs) predicting microscale individual trajectories in public spaces is being developed, that can assist in quantitative scenario analysis in various design scales and phases (Berghauser Pont et al. 2023, Ullrich et al., manuscript in review).

REFERENCES

Basu R and Sevtsuk A (2022) ‘How do street attributes affect willingness-to-walk? city-wide pedestrian route choice analysis using big data from Boston and San Francisco.’ *Transportation research part A: policy and practice* 163: 1–19. DOI: 10.1016/j.tra.2022.06.007.

Berghauser Pont, M and Haupt, P (2023) *SPACEMATRIX. Space, Density and Urban Form*, nai010 publishers, ISBN 978-94-6208-539-8, doi: https://doi.org/10.59490/mg.38

Berghauser Pont, M., Stavroulaki I., Marcus, L., Hunger, F., Ullrich, A., Ivarsson, O. Tarakanov, Y., Gosta A. (2023) Crowd Movement. Research report. Digital Twin City Centre (DTCC) Milestone project, Chalmers University of Technology, doi: 10.13140/RG.2.2.26267.23846

Berghauser Pont, M., Stavroulaki, G., Bobkova, E., Gil, J., Marcus, L., Olsson, J., et al. (2019a) ‘The spatial distribution and frequency of street, plot and building types across five European cities’, *Environment and Planning B: Urban Analytics and City Science*, 46(7), 1226–1242.

Berghauser Pont, M., Stavroulaki, G. & Marcus, L. (2019b) ‘Development of urban types based on network centrality, built density and their impact on pedestrian movement’, *Environment and Planning B: Urban Analytics and City Science*, 46(8), 1549–1564.

Berghauser Pont M, Stavroulaki G, et al., (2017a) Quantitative Comparison of Cities: Distribution of Street and Building Types Based on Density and Centrality Measures, in: *Proceedings of 11th International Space Syntax Symposium*, University of Lisbon, (eds. Heitor T, et al.), Instituto Superior Technico, Departamento de Engenharia Civil, Arquitetura e Georrecursos, Portugal, ISBN: 978-972-98994-4-7

Berghauser Pont M, Stavroulaki G, et al., (2017b) Quantitative comparison of the distribution of densities in three Swedish cities, In: *24th ISUF, City and territory in the Globalization Age Conference proceedings*, Universitat Politècnica de València, DOI: 10.4995/ISUF2017.2017.5317

Berghauser Pont M and Marcus L (2015) ‘What can typology explain that configuration cannot?’ In:*10th Space Syntax Symposium* (eds K Karimi et al.) London: University College London

Berghauser Pont, M. and Marcus, L. (2014). Innovations in measuring density. From area and location density to accessible and perceived density, in: Nordic Journal of Architectural Research, issue 2, pp. 11-31.

Bird E.L., Ige J.O., Pilkington P., Pinto A., Petrokofsky C. and Burgess-Allen J., (2018), ‘Built and natural environment planning principles for promoting health: an umbrella review’, *BMC Public Health*, 18:930, https://doi.org/10.1186/s12889-018-5870-2

Bobkova, E. (2019) *Towards a theory of natural occupation. Developing theoretical, methodological and empirical support for the relation between plot systems and urban processes*, PhD thesis, Chalmers University of Technology

Bobkova E, Marcus L, Berghauser Pont M, Stavroulaki G, Bolin D (2019) Structure of Plot Systems and Economic Activity in Cities: Linking Plot Types to Retail and Food Services in

London, Amsterdam and Stockholm. Urban Science, 3 (3), 66, https://doi.org/10.3390/urbansci3030066

Bobkova, E., Marcus, L. and Berghauser Pont, M., (2017). Multivariable measures of plot systems: describing the potential link between urban diversity and spatial form based on the spatial capacity concept. Lisbon, Proceedings of the 11th Space Syntax Symposium. 47:1-47:22.

Bolin, D., Verendel, V., Berghauser Pont, M., Stavroulaki, I., Ivarsson, O., Hakansson, E. (2021) 'Functional ANOVA modelling of pedestrian counts on streets in three European Cities', *J R Stat Soc Series A*. 2021;00:1–23. https://doi.org/10.1111/rssa.12646

Conroy Dalton, R., (2003), 'The secret is to follow your nose', *Environment and Behavior*, Vol 35, pp.107-131

CURT (2004) "Collaboration, Integrated Information, and the Project Lifecycle in Building Design and Construction and Operation", *Construction Users Roundtable* (WP-1202, August, 2004)

Dhanani, A., Vaughan, L (2016) 'Towards a walkability model for strategic evaluation of policy action and urban active transport interventions.' In: *(Proceedings) 48th Meeting of the Universities' Transport Study Group (UTSG)*, 6-7 January 2016, Bristol, UK. Universities' Transport Study Group (UTSG)

Dong H, Zhou M, Wang Q, Yang X and Wang FY (2020) 'State-of-the-art pedestrian and evacuation dynamics'. *IEEE Transactions on Intelligent Transportation Systems* 21(5): 1849–1866. DOI:10.1109/TITS.2019.2915014.

Gehl, J. (2010) *Cities for People*, Island Press, Washington, DC

Hanna S (2021) 'Random walks in urban graphs: a minimal model of movement.' *Environment and Planning B: Urban Analytics and City Science* 48(6): 1697–1711. DOI:10.1177/2399808320946766.

Heyman, Axel & Law, WPS & Berghauser Pont, Meta. (2019). 'How is Location Measured in Housing Valuation? A Systematic Review of Accessibility Specifications in Hedonic Price Models'. *Urban Science*, 3. 3.

Hillier, B. and Iida, S. (2005) 'Network and psychological effects in urban movement.', In: *Proceedings of the 5th International Space Syntax Symposium*, (ed A van Nes) Delft: University of Technology.

Hillier, B., Penn, A., Hanson, J., Grajewski, T., Xu, J. (1993) 'Natural movement: or, configuration and attraction in urban pedestrian movement.', *Environment and Planning B: Planning and Design,* vol.20, pp. 29-66.

Hillier, B. (1996a) 'Cities as movement economies.' In: *Space is the Machine: A Configurational Theory of Architecture*. Cambridge University Press

Hillier, B. (1996b) *Space is the Machine: A Configurational Theory of Architecture*. Cambridge University Press

Hillier, B., Yang, T., and Turner, A. (2012). 'Normalising least angle choice in Depthmap and how it opens new perspectives on the global and local analysis of city space.', *Journal of Space Syntax*. 3. pp.155-193

Legeby, A., (2013), *Patterns of co-presence: Spatial configuration and social segregation*, PhD thesis, Royal Institute of Technology, ISBN: 978-91-7501-920-8

Legeby, A., Berghauser Pont, M., Marcus, L., (2015), 'Street interaction and Social inclusion', In: Vaughan L. (ed.), *Suburban Urbanities. Suburbs and the Life of the High Street*, UCL Press, London, Ch.9, pp.239-263

Litman, T., (2020) *Evaluating Active Transport Benefits and Costs. Guide to Valuing Walking and Cycling Improvements and Encouragement Programs*, Victoria Transport Policy Institute

Moudon, A.V., Huang R., Stewart, O., Cohen-Cline, H., Noonan, C., Hurvitz, P. and Duncan, G., (2019) 'Probabilistic walking models using built environment and sociodemographic predictors', *Population Health Metrics*, 17:7, https://doi.org/10.1186/s12963-019-0186-8

Moudon, A.V, Lee, C., Cheadle, A, Garvin, C., Johnson D.B, Schmid, T.L, Weathers R.D., (2007), 'Attributes of Environments Supporting Walking', *Health Promoting Community Design; Measurement Issues*, Vol. 21, No. 3

Netto, V., Sabayo, R., Vargas, J., Figueiredo, L., Freitas, C. and Pinheiro. M. (2012), 'The convergence of patterns in the city: (Isolating) the effects of architectural morphology on movement and activity'. In: Greene, M., Reyes, J. and Castro, A. (eds.), *Proceedings of the Eighth International Space Syntax Symposium*, Santiago de Chile: PUC

Obeid, H.; Abkarian, H.; Abou-Zeid, M.; Kaysi, I. (2017) 'Analyzing driver-pedestrian interaction in a mixed-street environment using a driving simulator', *Accident Analysis and Prevention*

Ozbil A, Peponis J, Stone B (2011). 'Understanding the link between street connectivity, land use and pedestrian flows', *Urban Design International*. 6, pp.125-141.

Ozbil A, Yesiltepe D, Argin G (2015) 'Modeling Walkability: the effects of street design, street-network configuration and land-use on pedestrian movement', *A|Z ITU Journal of Faculty of Architecture*, 12, pp.189-207.

Pascucci, F.; Rinke, N.; Schiermeyer, C.; Friedrich, B.; Berkhahn, V. (2015) 'Modeling of shared space with multi-modal traffic using a multi-layer social force approach', 18th Euro Working Group on Transportation

Penn A, Hillier B et al. (1998) 'Configurational modelling of urban movement networks', *Environment and Planning B: Planning and Design,* 24, pp.59-84.

Peponis J, Bafna S, Zhang Z (2008) 'The connectivity of streets: reach and directional distance.', *Environment and Planning B: Planning and Design* 35(5), pp.881-901.

Peponis, J, Ross C, Rashid M (1997) 'The structure of urban space, movement and co-presence: The case of Atlanta.', *Geoforum*, 28 (3–4), pp.341-358.

Pivo, G. and Fisher, J.D, (2011), 'The Walkability Premium in Commercial Real Estate Investments', *Real Estate Economics*, pp. 1–35, DOI: 10.1111/j.1540-6229.2010.00296.x

Prato, C.G (2009) 'Route choice modeling: past, present and future research directions', *Journal of Choice Modelling*, 2:1, pp.65–100

Rinke, N.; Schiermeyer, C.; Pascucci, F.; Berkhahn, V.; Friedrich, B, (2017) 'A multi-layer social force approach to model interactions in shared spaces using collision prediction', *Transportation Research Procedia* 25

Roe, J., Mondschein, A., Neale, C., Barnes, L., Boukhechba, M., & Lopez, S. (2020). 'The Urban Built Environment, Walking and Mental Health Outcomes Among Older Adults: A Pilot Study', *Frontiers in public health*, 8, 575946. https://doi.org/10.3389/fpubh.2020.575946

Saelens, B., Sallis, J., Lawrence F., (2003) 'Environmental Correlates of Walking and Cycling: Findings from the Transportation, Urban Design, and Planning Literatures', *Environment and Physical Activity,* 25:2

Scoppa, M.D.; Peponis, J. (2015) 'Distributed attraction: The effects of street network connectivity upon the distribution of retail frontage in the City of Buenos Aires'. *Environment and Planning.* B, 42, 354–378.

Sevtsuk, A., Basu, R., Li, X., Kalvo, R. (2021). 'A big data approach to understanding pedestrian route choice preferences - Evidence from San Francisco.' *Travel Behavior and Society*. https://doi.org/10.1016/j.tbs.2021.05.010

Ståhle A, Marcus L, Karlstöm A (2005), 'Place Syntax. Accessibility with Axial lines in GIS', In: *5th Space Syntax Symposium* (ed AvanNes), pp.131-144, Delft: University of Technology.

Stavroulaki I., Berghauser Pont M., Fitger, M., Koch, D. Legeby, A., Marcus, L., (2023) PST Documentation_v.3.2.5_20231128, DOI: 10.13140/RG.2.2.32984.67845

Stavroulaki, G. (2022) 'Directions in Space Syntax. Space Syntax modelling of pedestrian flows for sustainable urban development', Keynote, in: *Proceedings of the 13th Space Syntax Symposium, Bergen, Norway*

Stavroulaki I., Berghauser Pont M., Marcus L., Sun K., Legeby, A. (2020a) Non-Motorised network of Gothenburg. Swedish National Data Service. Version 1.0. https://doi.org/10.5878/x49h-pv07.

Stavroulaki I., Berghauser Pont M., Marcus L., Sun K., Legeby, A. (2020b) Non-Motorised network of Stockholm. Swedish National Data Service. Version 1.0. https://doi.org/10.5878/hfww-5y22

Stavroulaki, G., Berghauser Pont, M., Markhede, H., Gil, J., Sun, K. (2020c) Pedestrian Study at Chalmers. Pedestrian movement and co-presence at Campus Johanneberg, Research report, Chalmersfastigheter, Akademiska Hus

Stavroulaki, G., Bolin, D., Berghauser Pont, M., Marcus, L. & Hakansson, E. (2019) 'Statistical modelling and analysis of big data on pedestrian movement.' In: *Proceedings of the 12th Space Syntax Symposium,* pp.79.1–79–24. Beijing Jiaotong University.

Stavroulaki, G., Marcus, L., Berghauser Pont, M., Nilsson, L. (2017) 'Representations of street networks in space syntax: Towards flexible maps and multiple graphs', In: *Proceedings of the 11th Space Syntax Symposium*, Lisbon

Trafikkontoret (2019) WiFi-mätningar av fotgängarflöden i Göteborgs innerstad-slutrapport, Diarienummer 5799/18, Göteborgs Stad

Turner A. (2007), 'From axial to road-centre lines: a new representation for space syntax and a new model of route choice for transport network analysis'. *Environment and Planning B: Planning and Design*, Vol.34, pp.539-555.

Turner, A., & Penn, A. (2002). 'Encoding Natural Movement as an Agent-Based System: An Investigation into Human Pedestrian Behaviour in the Built Environment'. *Environment and Planning B: Planning and Design*, 29(4), 473-490. https://doi.org/10.1068/b12850

Ullrich, A., Hunger, F., Stavroulaki, I., Bilock, A., Jerereg, K., Tarakanov, Y., Gosta, A., Quist, J., Berghauser Pont, M., Edelvik, F., (manuscript in review) 'Towards predictive pedestrian movement modelling: A workflow connecting a network and agent-based model '