Statistical modelling and analysis of big data on pedestrian movement

Citation for the original published paper (version of record):
Stavroulaki, I., Bolin, D., Berghauser Pont, M. et al (2019). Statistical modelling and analysis of big data on pedestrian movement. 12th International Space Syntax Symposium, SSS 2019

79

STATISTICAL MODELLING AND ANALYSIS OF BIG DATA ON PEDESTRIAN MOVEMENT

GIANNA STAVROULAKI[1]; DAVID BOLIN[2]; META BERGHAUSER PONT[3]; LARS MARCUS[4]; ERIK HÅKANSSON[5]

ABSTRACT

This work follows a long line of studies and empirical investigations in Space Syntax research, that, in general, try to conceptualise, describe and quantify the relation between physical space and human agency. How many people share public space is known to affect many socio-economic processes in cities, such as segregation, vitality and local commercial markets. Observing and measuring pedestrian movement through surveys, as well as statistically analysing it have been at the core of diverse investigations not least in the field of Space Syntax, not only a means to validate and measure the dependence of pedestrian movement on spatial configuration, but also as a means to forecast and predict pedestrian flows. However, these studies do not necessarily provide us with comparable, let alone generalisable findings that can lead to generalisable propositions. They remain in most cases specific investigations of particular cities, neighbourhoods or types of areas (e.g. city centres). Another issue, as will be elaborated in this paper, is that the typical statistical methods used, such as multivariate regression models, are not always the optimal or even suitable for modelling pedestrian movement, typically measured in pedestrian counts.

This paper aims therefore, to directly address three methodological challenges: first, construction of comparable GIS-models; second, gathering large scale pedestrian data; third, applying advanced statistical modelling suitable for pedestrian data. The ultimate goal is to estimate the impact of spatial form on urban life in a way that is methodologically sound and can provide robust results that can be

[1] Chalmers University of Technology, Department of Architecture and Civil Engineering, SE-412 96 Gothenburg, Sweden, gianna.stavroulaki@chalmers.se
[2] Chalmers University of Technology, University of Gothenburg, Department of Mathematical Sciences, SE-412 96 Gothenburg, Sweden, david.bolin@chalmers.se
[3] Chalmers University of Technology, Department of Architecture and Civil Engineering, SE-412 96 Gothenburg, Sweden, meta.berghauserpont@chalmers.se
[4] Chalmers University of Technology, Department of Architecture and Civil Engineering, SE-412 96 Gothenburg, Sweden, lars.marcus@chalmers.se
[5] University of Gothenburg, Chalmers University of Technology, Department of Mathematical Sciences, SE-412 96 Gothenburg, Sweden, gushakaer@student.gu.se

generalisable, and allows us to speak of the relation between spatial form and pedestrian movement in a way that is not specific to a certain area, or types of areas or streets, or even to a specific city.

The results show, first, high and consistent correlations between spatial form and pedestrian movement in a study of unprecedented size that comprises three cities, including a large range of neighbourhoods of varying morphological types, from villa areas to urban cores, and offer convincing proof that the tested morphological variables have a strong impact on the spatial distribution of pedestrian flows in cities. Second, the study shows that the model with all explanatory variables has the highest explanatory power and the best model fit where Angular integration and Accessible FSI are the explanatory variables with the largest effect on pedestrian movement, but others are significantly contributing to the predictive power of the model. Third, the study contributes to the advancement of the statistical modelling that is suitable for the specificities of the data used, proposing the use of a negative Binomial model instead of regression models, most common in the field.

KEYWORDS

statistical modelling, pedestrian movement, anonymised pedestrian survey, spatial analysis, spatial morphology

1. INTRODUCTION

This paper describes results of a large empirical study that is aiming to quantify the effect of spatial form on pedestrian movement. We focus on the separate and combined impact of configuration, specifically the space syntax measures of angular centrality, in relation to other key morphological variables, (i.e. build density and land division) and the presence of attractions. The study was conducted in three European cities - Stockholm, Amsterdam and London – in a large variety of neighbourhoods in terms of density and building type, that further included a variety of different street types. The objective is both to test the relation between spatial form and pedestrian movement in a large and consistent study comprising more than one city and to advance the methodological framework used in such empirical studies - from the gathering of empirical data to their statistical modelling - to arrive at both scientifically valid and generalisable results concerning this relation.

1.1. Theoretical and methodological background

This work follows a long line of studies and empirical investigations in Space Syntax research, that, in general try to conceptualise, describe and quantify the relation between physical space and human agency. The interplay between humans and their environment is a question pondered by humanity for centuries. Even so, it has been addressed from the very beginning in space syntax theory (Hillier & Learman, 1973), and developed into descriptions of architectural and urban space, such as axial maps and convex spaces, with the aim to capture this relation (Hillier & Hanson, 1984). These descriptions are on the one hand, based in human affordances and on the other, the configurative (or syntactic) relation between individual spaces (Hillier, 2012). The first means that descriptions relate neither to physical space in itself nor to human behaviour, but to what emerges when they meet; the second means that the descriptions used address the systemic level of space, which opens for linking these descriptions to dynamic affordances, such as movement, and not only static, such as what is perceived from a single point in space.

If these constitute descriptions of the environment, similar care has been taken to understand what aspects of human behaviour need to be captured to establish the link between human agency and the physical environment. Central here has been the notion of generic function (Hillier, 1996), which avoids the entanglement of the innumerable aspects of human activity in cities, by cutting straight to activity that has direct spatial implications. Consequently, it, on the most fundamental level, identifies a distinction of urban space into, first, streets and other open spaces, that primarily have the function of urban movement, and second, city blocks and their buildings, that have the function of urban occupation.

Hence, the fundamentals of a methodology to study the relation between humans and the urban environment is set up in the form of statistical correlation studies between systemic properties of descriptions such as axial maps and movement counts. Such studies have repeatedly reported a consistent relation between the two and have also generated theory, such as the theory of natural movement (Hillier et al., 1993), which has remained a key reference in space syntax research. Based on earlier and new studies (i.e. Peponis et al, 1989; Hillier et al, 1987), it aimed to demonstrate the primacy of spatial configuration over attraction when it comes to the influence on the distribution of pedestrian movement in urban space. However, it also added that since pedestrian movement, in turn, has a strong influence on the distribution of land-use, attractions will add to the influence of configuration on movement, creating a multiplier effect.

More recently, a study by Hillier and Iida (2005), has taken over as key reference to this central relation between configuration and pedestrian movement in space syntax research. Importantly, this study is set in a different theoretical framework that concerns the psychology behind human navigation in space and how different distance measures, such as metric, topological or angular, best reflect this. The aim of the study was the test and validation of angular distance in measures of centrality in street networks, a technique that we will also use in this paper. Building on an earlier study by Penn et al (1998), an extensive empirical study was conducted that proved angular distance as a powerful means for the purpose.

Apart from the referenced papers, numerous studies have to this day been scrutinising the impact of spatial configuration on pedestrian movement, also in relation to other factors, such as land use and built density (e.g. Peponis et al., 1997; Read 1999; Berghauser Pont and Marcus, 2015; Marcus et al, 2017; Ozbil et al. 2011, 2015; Netto et al. 2012; Legeby 2013; Ståhle et al. 2005). For instance, Berghauser Pont and Marcus (2015) and Ståhle et al. (2005) show that correlations vary depending on built density and morphological neighbourhood type, and Read (1999) highlighted the variations between local grids and super-grids.

Hence, observing and quantifying pedestrian movement and statistically correlating it to different explanatory variables of spatial form has been at the core of extensive and diverse investigations aiming to validate and measure the dependence of pedestrian movement on spatial configuration, but it has also been a means to forecast and predict pedestrian flows. Although studies of this kind around the globe have generally reported high and consistent correlations, there have also been some methodological inconsistencies, for instance, in the kind of empirical data used, the explanatory variables tested or the statistical methods applied. Also, they have seldom been based on data samples of a size and distribution that can be said to represent cities as a whole, let alone comprise several cities allowing for comparison. Often, studies are confined to neighbourhoods and use datasets too small for robust statistical analysis.

A reason is that it has been difficult to access, collect or even process data sets of a size that can support large scale comparative studies, something made possible in recent years of increasing computational power and data access. Although, studies of specific areas in cities are very useful for understanding the effect of configuration on movement, since other variables are possible to control (for example, density often being evenly distributed, the demographic and socioeconomic profile being relatively stable, and land use distribution possible to control), they do not provide findings that can support generalisable propositions. A case in point, is the fact that the statistical results are inconsistent when it comes to radii of centrality, where the radius that is statistically significant changes from area to area, even in the same study.

Another issue, which will be elaborated in the methodological part of this paper, is that the statistical methods typically used, such as multivariate regression models, are not always the optimum for modelling pedestrian movement measured by pedestrian counts. Also, particularities of spatial data, such as spatial autocorrelation is not always considered in the statistical modelling, jeopardizing the scientific validity and robustness of the explanatory and predictive power. The above methodological inconsistencies and gaps form a weak link in an unusually well-designed research programme.

This paper aims therefore, to directly address these methodological challenges in a large empirical study. A new set of co-produced GIS-models of London, Amsterdam and Stockholm, based on Space Syntax methodology and constructed in exactly the same way was applied. These models allow us to extract consistent and comparable data on spatial form from all three cities. For equally consistent and comparable data on human activity, we used tracking of anonymized Wi-Fi signals from mobile phones to conduct a large-scale pedestrian survey[6] in all cities – unique to our knowledge.[7] Together this allows

[6] The survey was led by Chalmers University of Technology in cooperation with Bumbee Labs consultancy, who provided the technology and conducted data processing in close discussion with Chalmers.

[7] There are some studies, mostly from computer and information science that use the same method, but in general either the experiments are very small scale aiming to validate and optimise the method,

for an unusually large and consistent study of the critical relation between spatial form and pedestrian movement both for research purposes and practices based upon its findings. Finally, the statistical model used to validate and quantify this relation, introduce methodological advancements more suitable for pedestrian data.

In summary, we contribute to existing Space Syntax literature in three ways. First, we continue a long line of investigations relating spatial configuration to pedestrian movement, focusing on angular centrality measures, with a larger and more consistent study than any of the earlier. Second, we use a unique dataset of pedestrian flows, covering three different cities and hence a large variety of neighbourhood and street types, which provides information and results that are not only specific for certain areas or cities, but that are generalisable and demonstrate the consistent impact of variables of spatial form on pedestrian movement. Third, we advance the statistical methods used to analyse the relation between data on spatial form and movement by proposing predictive statistical models that consider particularities of spatial data, such as autocorrelation, but also particularities of pedestrian data, as measured by pedestrian counts.

1.2. Outline of the paper

The paper is organised in three sections, following this first introductory section. Section 2 describes the methods and datasets used in the study, where we, first, introduce the explanatory variables and describe how they are calculated; second, present the pedestrian survey and its method of data gathering; third, describe the structuring of the datasets; and fourth, describe the method of statistical analysis and modelling used throughout the study. Section 3 presents the results of the statistical analysis, the model fit and the coefficient estimates. Finally, Section 4 highlights the general conclusions of the study and proposes next steps.

2. DATASETS AND METHODS

The overall methodology for the statistical modelling and analysis of pedestrian data aims at testing the independent and relative impact of spatial form on pedestrian flows. We start by testing the explanatory power of angular integration and angular betweenness centrality, as two fundamental measures of spatial configuration. Then we statistically model the predictive power of the same measures in combination to two other key measures of spatial form – built density and land division, building on previous work (Marcus et al. 2017, Berghauser Pont and Marcus 2014, Bobkova et al. 2017). Finally, we include some attraction measures that arguably attract or are associated with increased pedestrian movement throughout the day or intermittently, such as public transportation nodes (i.e. tram, bus, ferry

are indoor or are covering only a specific attraction or event. Examples on outdoor studies are: Schauer et al. 2014; Kurkcu and Ozbay 2017; Abedi et al. 2013, Petre et al. 2017; Barbera et al. 2013.

and train stops, bus and train stations), schools (including kindergartens) and local markets (i.e. all ground floor retail shops, restaurants and cafes).

The datasets used in this study are structured per street segment and cover the metropolitan areas[8] of three European cities, Stockholm, Amsterdam and London. All cities are modelled together, but we include city as categorical variables to control for the effect of the city where data was collected. Because we conducted the pedestrian survey in different days of the week, also weekday was added as control variable.

In the following sections, the main methodological steps and choices taken in this study will be presented. First, all the explanatory variables and method of calculation will be described; second, the pedestrian survey and the method of gathering big data on pedestrian movement will be presented; and third, the statistical methods and models used throughout the study will be introduced.

2.1. Explanatory variables

2.2.1. Angular Integration and Angular betweenness centrality

Angular integration and angular betweenness centrality were calculated for the non-motorised street network of each city, or in simple words, the network, pedestrians use. This includes all streets and paths that are accessible for people walking, including those shared with vehicles and bicycles. Streets where walking is forbidden, such as motorways, highways, or high-speed tunnels, are not included in the analysis.

The line-segment maps used are based on official road-centre-line maps[9] and processed following the method described in Berghauser Pont et al. (2017a). To reduce calculation time, but, most importantly, to increase comparability, the same editing and generalisation procedures (e.g. removing errors and reducing the number of line-segments in a consistent manner[10]), are used for all cities.

[8] The metropolitan areas extend beyond the administrative boundaries and were defined following the Urban Morphological Zones (UMZ) as set by the European Environment Agency (EEA) and the Eurostat. A UMZ is defined as “a set of urban areas laying less than 200m apart”. (source: http://www.eea.europa.eu/data-and-maps/data/urban-morphological-zones-2006 (download date 13-7-2016). The convex hull of each UMZ was used to provide a more regular shaped study area, more appropriate for spatial analyses.

[9] Official Road-Centre-Line maps: NVDB from Trafikverket, Sweden; ITN from Ordnance Survey, UK; and, NWB from Rijkswaterstaat-CIV, the Netherlands. The downloads were done from 05/2016 to 10/2016.

[10] This process, before the final segmentation of the Road-Centre-lines to line-segments, included removing duplicate and isolated lines, snapping and generalizing. The snapping threshold used was 2m (end points closer than 2m were snapped together). The generalizing threshold used was 0,5m (successive line segments with angular deviation less than 0,5m were merged into one). All editing procedures were done with PST.

Because the aim of this paper is to obtain a description of the three metropolitan regions that can predict the intensity of pedestrian flows, we select radii for analysis that are close to the more local scales of pedestrian movement with a maximum of 5km[11]. Also, to provide a uniform and continuous sampling of centrality, the radii are equally spaced and have a small interval (i.e. 500m), resulting in in 10 different radii. All editing procedures were done with PST[12].

For angular betweenness centrality, the following equation is used:

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

where s and t are all nodes 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

For Normalised angular integration (NAIN) (Hillier et al. 2012), the following equation is used:

$$AI_{NAIN}(x) = \frac{N^{1.2}}{1+\sum_{i \neq x} D(x,i)} \qquad (2)$$

where

N= node count or number of reached line-segments in the network

D(x,i) = angular distance of the shortest path between i and x, calculated as the accumulated angular turns needed to get from line-segment x to line-segment i in the network. Angular distance is measured in degrees and then divided by 90 (Hillier and Iida, 2005).

2.2.b. Built density and Land division

Following the work of Berghauser Pont and Marcus (2014), built density and land division are calculated using a measure of accessibility and, more specifically, the cumulative-opportunities accessibility measure (Handy et al. 2016) with the distance threshold set at 500m walking distance. In particular, built density is described as the Accessible FSI (Floor space ratio)[13] in 500m (Berghauser Pont and Marcus 2014) and land division is described as the Accessible number of plots in 500m (Bobkova et al 2017). Thus, for example, the measure of 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

[11] In order for all streets to be calculated within a context of 5km walking distance and reduce the possible "boundary effect" the area which was analysed was at least 5km larger than the study area in all directions.

[12] PST software (Place Syntax Tool, plugin QGIS) was used for editing and calculations. Documentation, including equations is available at https://www.smog.chalmers.se/pst.

[13] Built density is described, following the work of Berghauser Pont and Haupt (2010), by two measures: FSI(Floor space index) and GSI(Ground space index). However, since the two measures are colinear, only one is included in the explanatory variables used in the statistical study.

segment through the street network, adding up to a measure of accessible density, or in other words, human accessibility to built up space within a distance that most people are willing to walk, commonly recognized to be approximately 500 meters (Gehl 2010).

In particular, for the density calculation, the Gross Floor Area (GFA) for each building polygon is calculated by multiplying the area of the building polygon with the average building height[14]. Next, we used the equation for Attraction reach in PST[15] to calculate accessible Gross Floor Area (GFA) in 500m:

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

where
A = the set of reachable attractions within given radius,
f(a) = attractions value associated with attraction a,
D(o,a) = shortest distance from origin o to attraction a,
w(x) = attenuation function.

The attraction value f(a) is GFA when calculating FSI(o). D(o,a) is defined by the chosen radius and is here 500m from the origin. As origin, the midpoints of the line-segments are used. The attenuation function is not used.

Accessible FSI(o) is then calculated as follows:

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

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 measure of Accessibility to plots is directly related to the size of the plots and the grain of the land division. The higher the number of accessible plots within a radius is, the smaller the plots are and the finer the grain of land division is within that radius. Thus, again, as in the case of built density, we describe plot size not as an individual property of each plot, but as an area-based measure of the plot structure and the land division.

[14] The height data used to calculate GFA were received in ready-to–use formats for Amsterdam and London: 3dBAG from ESRI (http://www.esri.nl/) and Ordnance Survey. (https://www.ordnancesurvey.co.uk/) respectively. For Stockholm, height data were extracted from a laser dataset (Lantmateriet, slu.get.se) (see Berghauser Pont et al. 2017 for more details).

[15] PST software (Place Syntax Tool, plugin QGIS). Documentation, including equations is available at https://www.smog.chalmers.se/pst.

The accessible number of plots in 500m was calculated based on comparable plot layers constructed for the three cities, following the work of Bobkova et al (2017)[16]. In Sweden and the Netherlands, the plot layer is based on the cadastral system, while in the UK, the freehold property system is used[17]. To calculate accessible number of plots from each line segment within 500m walking distance, we use again the equation of Attraction reach (3) in PST, where the set of attractions are the plots.

2.2.c. Attraction variables

To evaluate the independent and relative impact of spatial variables on pedestrian flows, we add attraction variables to the explanatory variables list. The datasets of attractions are extracted from the point datasets of Open Street Maps[18]. 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 adjacent streets and 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[19] within 500m walking distance from each line-segment (i.e. midpoint), Number of Local Markets on each line-segment, Accessible Public transport nodes within 500m from each line-segment, Number of Public transport nodes on each line-segment, Accessible Schools within 500m from each line-segment, Number of Schools on each line-segment.

Full name	**Abbreviation**
Angular Integration (NAIN), radius	(for radius 1000m) Int1000
Angular Betweenness, radius.	(for radius 3500) Bet3500
Accessible FSI in 500m walking distance	FSI_500
Accessible Number of Plots in 500m walking distance	Plot_500
Accessible Local markets in 500m walking distance	LMarkets_500
Number of Local Markets on segment	LMarkets_Str
Accessible Public transport nodes within 500m walking distance	PubTr_500

[16] Both cadastral and freehold systems cover all types of land, including road and rail networks as well as water bodies, so the plot layers were constructed based on Hillier's concept of generic function (1996), defined as 'land used for long term stationary functions'. Hence, the final layer of plot polygons consists of land properties that cover all sorts of land except water and movement networks.

[17] Data sources: Fastighet maps from the Swedish Land registry for Sweden, the DKK database for Amsterdam, and the Land Registry Inspire Index polygons for London. The UK has two layers of plot systems, freehold and leasehold properties with only the former accessible to the public, instead of a single cadastral system (freehold property is the ownership of the property and the land it stands on and leasehold property is the ownership of the property for a fixed term without owning the land that it stands on).

[18] https://www.openstreetmap.org

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

Number of Public transport nodes on segment	PubTr_Str
Accessible Schools within 500m walking distance	Schools_500
Number of Schools on the segment	Schools_Str

Table 1. List of variables and abbreviations

2.2. Collecting big data on pedestrian movement

In order to gather comparative empirical data to test the impact of the morphological variables on pedestrian movement, we conducted a large pedestrian survey in all three cities tracking anonymised Wi-Fi signals from mobile phones. The pedestrian survey included around 18 neighbourhoods per city (19 in Stockholm, 18 in Amsterdam, 16 in London) and was done within a three-week period in October 2017[20]. The areas were selected with the main objective to cover all building and street types ranging from small alleys to high streets, in neighbourhoods which differed in building type, from suburban villa areas of low density to central high-density areas with primarily closed building blocks. The selected neighbourhoods included business districts, mixed-use areas and residential areas of different income level (Fig1).

[20] The weather was similar in all cities, mostly moody with short periods of rain within the day.

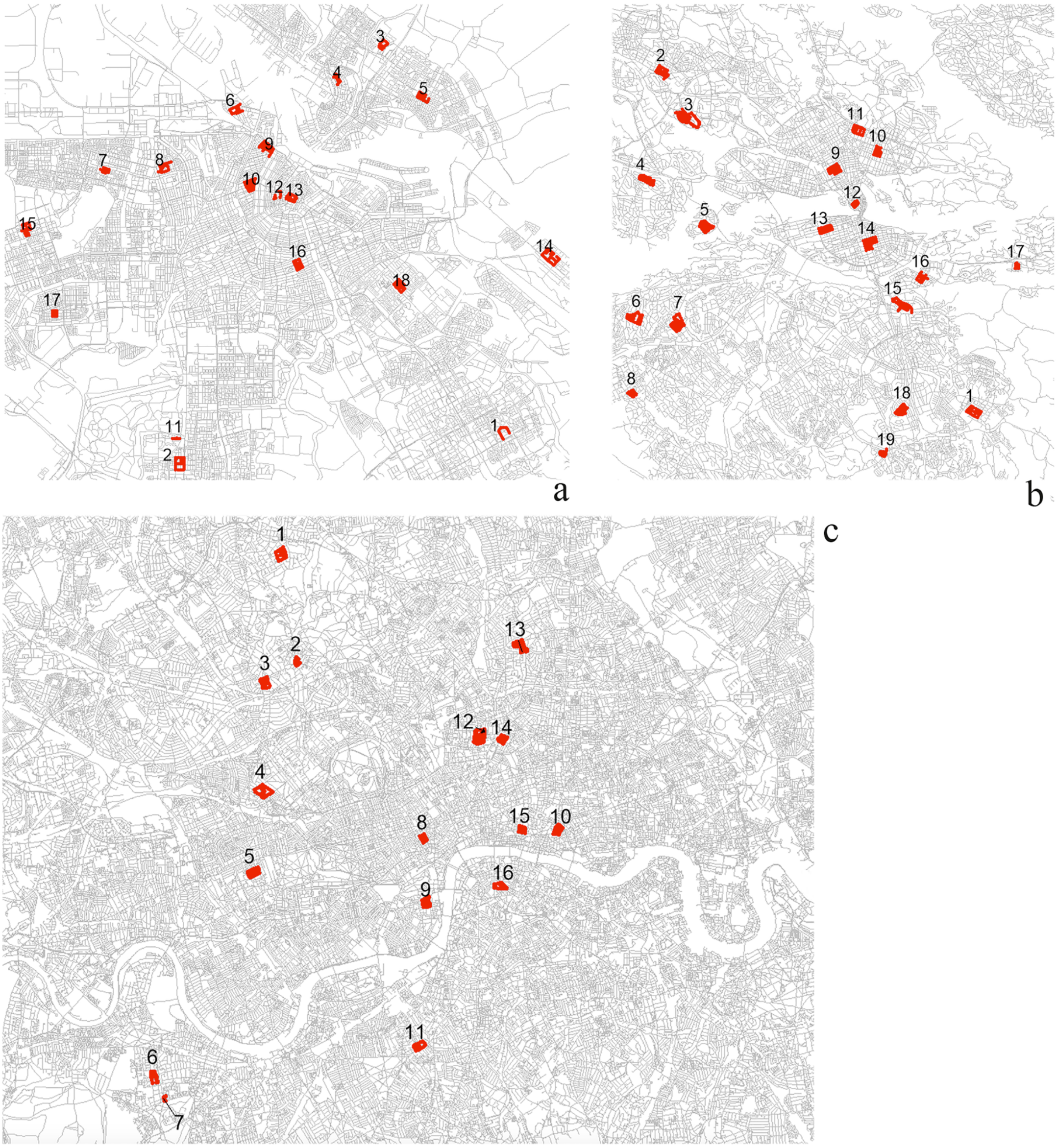


Figure 1. Selected areas for a) Amsterdam, b) Stockholm and c) London.
a)1_Zuidoost, 2_Amstelveen Elsrijk, 3_Noord Kadoelen, 4_Noord, 5_Noord Nieuwendam, 6_Spaarndammer en zeeheldenbuurt, 7_Slotermeer-noordoost, 8_Bos en Lomer, 9_Haarlemmerbuurt, 10_Jordaan, 11_Amstelveen Patrimonium, 12_Burgwallen Nieuwe Zijde, 13_De Wallen, 14_Ijburg West, 15_Osdorp-oost, 16_De Pijp, 17_Nieuw Sloten, 18_Watergraafsmeer
b)1_Skarpnäck, 2_Sundbyberg, 3_Jungfrudansen, 4_Stora Mossen, 5_Stora Essingen, 6_Mälarhöjden, 7_Västertorp, 8_Segeltorp, 9_Norrmalm, 10_Östermalm A, 11_Östermalm B, 12_Gamla stan, 13_Södermalm Maria Församlingen, 14_Södermalm Katarina Församlingen, 15_Hammarby Höjgen, 16_Hammarby Sjöstad, 17_Järlasjö, 18_Tallkrogen, 19_Hökarängen
c)1_Hampstead garden suburb, 2_Hampstead station area, 3_West Hampstead, 4_Maida Hill, 5_Notting Hill, 6_Putney A, 7_Putney B, 8_Soho, 9_Westminster, 10_Cornhill A, 11_Clapham, 12_Barnsbury, 13_Highbury East, 14_Hoxton, 15_Cornhill B, 16_Newington

Because we are interested in the isolated effect of the spatial variables on pedestrian movement, we made sure that no big attractors such as train stations or large shopping malls were located on or near the selected streets.

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. This method was chosen not only because it is technically advanced and appropriate to collect anonymized big data (Schauer et al. 2014), but also because it is GDRP compliant[21] and can be used in in all European cities[22].

The detection devices were positioned at every street crossing in the selected areas. Each neighbourhood was monitored continuously for one workday from 6 AM in the morning to 10 PM in the evening. In total, data for 846 street segments were collected (354 in Stockholm, 296 in Amsterdam and 266 in London), reaching 766.645 pedestrian trips in London, 532.068 in Amsterdam and 789.889 in Stockholm. The collected data record how many people passed each street per hour, but also their average speed and exact paths through the area; revealing both flow patterns and intensities, as well as the microstructure of individual paths.

The main variable used in this paper to represent the pedestrian movement in each street segment is the intensity[23] of pedestrian flows in a whole day, defined as the total amount (count) of people that walked each street from 6 AM to 10PM. We only used street segments where we had full day data available, which resulted to a final dataset of 227 street segments in Stockholm, 296 in Amsterdam and 266 in London.

2.3. Structuring the datasets for the statistical modelling

As already described in section 2.2, all network centrality variables were calculated for line-segment maps using Angular segment analysis. The same line segments were used as origin points in the Attraction reach equation (3) in order to calculate Accessible FSI, Accessible number of Plots, and Accessibility to Attractions (public transport, schools, and local markets). As a result, all datasets with the explanatory variables were structured per line segment.

The Wi-Fi-signals to measure pedestrian movement were monitored at street junctions and, respectively, the pedestrian counts were calculated for each street segment between each pair of adjacent junctions. These street segments often include more than one line-segments, especially in curvilinear streets. To deal with that, all values of the explanatory variables were transferred from the line segments to the street segments, using the proportion average function[24].

2.4. Statistical model

The simplest possible model for the data would be a standard multiple regression, where the pedestrian count Y_i at location i is assumed to be a Gaussian random variable $N(\eta_i \sigma^2)$, where σ^2 denotes the variance and the mean is

[21] GDPR (General Data Protection Regulation) for the EU member states.

[22] For other studies that use wi-fi tracking for outdoor observations see: Schauer et al. 2014; Kurkcu and Ozbay 2017; Abedi et al. 2013, Petre et al. 2017; Barbera et al. 2013. However, they are mostly small-scale experiments aiming to validate and optimise the method, or are covering only a specific attraction or event.

[23] Other terms used in literature are pedestrian flows, pedestrian density, pedestrian movement, pedestrian rates, occupation rates.

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

$$\eta_i = \sum_{k=1}^{K} X_{k,i}\ \theta_k. \quad (5)$$

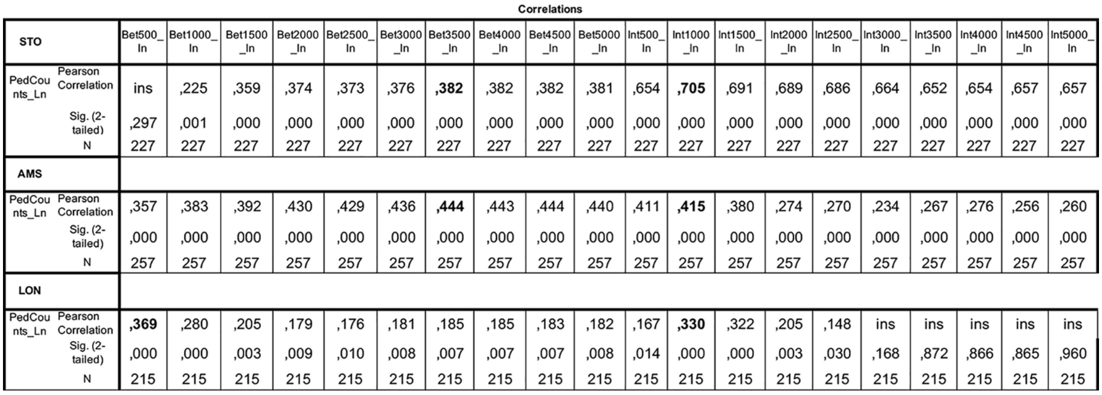
Here the $X_{k,i}$ denotes the k:th explanatory variable evaluated location i, θ_k is the corresponding regression coefficient, and K is the total number of explanatory variables. However, such a model works poorly for the data due to the highly skewed distribution of the counts, and due to the fact that the measurements are positive counts. One possible solution to this problem would be to instead model some transformation of the counts. One could for example let $\log(1 + Y_i) \sim N(\eta_i, \sigma^2)$, where the reason for $1 + Y_i$ is to ensure that those values are strictly positive. Although this works fairly well, as we will see later, it is unsatisfactory for mainly two reasons. The first is that the data are counts and thus integer valued, whereas the transformed regression model is a model for continuous data. The approximation of count data by a continuous distribution is problematic especially for low counts. The second is that the choice of transformation, $\log(1 + Y)$ is a bit ad-hoc.

A more mathematically satisfactory solution is to use a regression model for count data. The standard choice here is Poisson regression, where Y_i is assumed to be Poisson distributed with mean e^{η_i}. A feature of this model is that both the mean and variance of Y_i is assumed to be e^{η_i} due to the Poisson assumption. When testing the model for the data, we found that the variance Y_i in reality was much larger than the mean (a common term for this is over-dispersion) which means that Poisson regression would not give reliable results. A solution to the problem with over-dispersion is to replace the Poisson distribution with a more flexible distribution for count data which can model over-dispersion. A common such choice is the negative Binomial distribution $nBin(\mu, n)$, where $\mu > 0$ is a parameter that determines the mean of the distribution and $n > 0$ is a dispersion parameter. With this parametrisation, the variance is $\sigma^2 = \mu + \mu^2/n$, which means that the parameter n can be used to control the variance independently of the mean. Thus, the main model we use is a negative Binomial regression where we assume that $Y_i \sim nBin(e^{\eta_i}, n)$. The parameters in the model that need to be estimated from the data are the regression coefficients as well as the dispersion parameter.

Another potential problem to address is that there may be spatial correlation in the data which cannot be explained by the mean e^{η_i}. In this case, the estimated regression coefficients may be affected by this correlation and one should be careful with drawing conclusions from the model results. Thus, to get reliable results we include the remaining spatial correlation in the data in the model. We do this by including a random effect that models the spatial dependence. Specifically, we assume that the mean for the count Y_i in neighbourhood j is $e^{\eta_i + U_j}$ where $U_j \sim N(0, \sigma^2)$ acts as a neighbourhood specific intercept. The variance σ^2, which is estimated from data, determines how much these neighbourhood-specific intercepts vary.

Finally, we selected the explanatory variables to include in the formulation of η_i. We tested three different types of negative Binomial models for the mean, namely, Configurational, Spatial and Attraction. The Configurational model included just Angular integration and Angular betweenness as explanatory variables. Both measures were calculated in ten different radii as was explained in Section 2.1.a.

Preliminary Pearson correlations (r) were used to compare and select the most representative variables-radii (Fig2). This step was necessary in order to avoid inputting colinear variables to the statistical model. These preliminary results showed that for both Stockholm and Amsterdam the strongest correlations were found for Angular Betweenness 3500m, and Angular Integration 1000m. In the case of London, Angular Integration 1000m showed again the strongest correlation, but for Angular Betweenness radius 500m was stronger (0,369) than radius 3500m (0,185). However, for consistency and comparability reasons and in order to be able to use one model to arrive to general results, and not only specific for each city, we used Angular Betweenness 3500m for London as well as an input variable. To use, Angular Betweenness 500m, instead, for all cities was not an option, because in Stockholm this showed non-significant correlations.



Correlations

STO			Bet500_ln	Bet1000_ln	Bet1500_ln	Bet2000_ln	Bet2500_ln	Bet3000_ln	Bet3500_ln	Bet4000_ln	Bet4500_ln	Bet5000_ln	Int500_ln	Int1000_ln	Int1500_ln	Int2000_ln	Int2500_ln	Int3000_ln	Int3500_ln	Int4000_ln	Int4500_ln	Int5000_ln
PedCounts_Ln	Pearson Correlation		ins	,225	,359	,374	,373	,376	**,382**	,382	,382	,381	,654	**,705**	,691	,689	,686	,664	,652	,654	,657	,657
	Sig. (2-tailed)		,297	,001	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	N		227	227	227	227	227	227	227	227	227	227	227	227	227	227	227	227	227	227	227	227
AMS																						
PedCounts_Ln	Pearson Correlation		,357	,383	,392	,430	,429	,436	**,444**	,443	,444	,440	,411	**,415**	,380	,274	,270	,234	,267	,276	,256	,260
	Sig. (2-tailed)		,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	N		257	257	257	257	257	257	257	257	257	257	257	257	257	257	257	257	257	257	257	257
LON																						
PedCounts_Ln	Pearson Correlation		**,369**	,280	,205	,179	,176	,181	,185	,185	,183	,182	,167	**,330**	,322	,205	,148	ins	ins	ins	ins	ins
	Sig. (2-tailed)		,000	,000	,003	,009	,010	,008	,007	,007	,007	,008	,014	,000	,000	,003	,030	,168	,872	,866	,865	,960
	N		215	215	215	215	215	215	215	215	215	215	215	215	215	215	215	215	215	215	215	215

Figure 2. Table of Pearson correlations for Angular Integration and Angular Betweenness in different radii for each city. Highest values in bold.

The Spatial model includes, besides the two configurational variables, the variables Built density (i.e. Accessible FSI) and Land division (i.e. Accessible number of plots). Finally, the Attraction model added all the attraction variables described in section 2.2.c. For all three models, we include “weekday” and “city” as categorical variables that control for the effect of the day the pedestrians were counted and the city.

2.5. Estimation and validation methods

The model parameters are estimated using the R-INLA (Rue et al. 2009) package with standard settings. In the estimation procedure, we rescaled all continuous explanatory variables by dividing each variable with its root mean-square deviation in order to improve the numerical stability. R-INLA estimates the model parameters in a Bayesian setting, by assuming that each regression coefficient θ_k has an independent $N(0, \tau^2\ )$ prior distribution. The parameters are estimated by the posterior mean $E(\theta_k|Y)$, i.e. the expected value of the parameter given by the data. Since the explanatory variables were rescaled, the default choice of τ^2=1000 was found to be sufficiently large to not have an effect on the estimates of θ_k (meaning that the estimates did not change if we increased τ^2 further, and thus the estimates are determined by the data and not by the prior distributions).

To assess if an estimated model fits to the data, we examined the residuals $e_i = \ Y_i - \mu_i$, where μ_i is the mean count at location i according to the model. A first thing to check is if there is any spatial structure in these residuals. To get plots that are easier to interpret, we calculated the mean of these residuals for each neighbourhood and plotted this mean on the map of each city. The second thing to check is whether the residuals have the correct distribution according to the model. To do this, we simulate new data $\widehat{Y_i}$ from the model and compute the corresponding residuals $\hat{e}\_i$. We then display a quantile-quantile (QQ) plot[25] of $\widehat{e_i}$ shown against e_i. The curve in this plot should follow a straight line if the residuals have the correct distribution. However, there is uncertainty in this procedure due to the random sampling of new data. To get a better understanding of this uncertainty, we repeat the procedure for 100 different simulated datasets and plot the QQ curve for each simulated dataset in the same figure. This results in 100 different curves that should cover the straight line. If this is the case, we conclude that the model fit is adequate. If on the other hand, all curves are above or below the straight line at some location, this indicates that the residuals have the wrong distribution.

To compare different models in terms of model fit, we use R^2 values as well as the continuous ranked probability score (CRPS) (Gneiting and Raftery, 2007). R^2 is computed as one minus the ratio of mean squared error of a model to the mean squared error of an intercept-only model. The former is computed by taking the mean fitted value (which includes the random effects) for each observation and comparing it to the true data.

The R^2 values thus show how well the model can predict the pedestrian counts. However, another important aspect of predictive models is to also get the uncertainty of the predictions correct (a point prediction does not say anything unless we also have a measure of the certainty of this prediction). Because of this, we also use the CRPS values to compare the model fits. CRPS is a so-called proper scoring rule that measures the correctness of the entire predictive distribution. For a given count y_i, the CRPS value is computed as

$$CRPS(y_i) = 0{,}5\mathbb{E}_{X1,X2}\left(|X_1 - X_2|\right) - \mathbb{E}_X\left(|y_i - X|\right) \quad (6)$$

where X, X$_1$, X$_2$ are independent and follow the distribution of y_i according to the model[26]. Thus, the score compares the expected absolute difference between to potential counts X_1 and X_2 to the expected absolute difference between a potential count X and the actual count y_i. To get a single value which we can use to compare different models, we compute the average CRPS score over all pedestrian counts, CRPS = $\frac{1}{n}\sum_{i=1}^{n} \text{CRPS}(y_i)$. As for R^2, a larger value indicates a better model fit.

3. RESULTS

In this section we present the model fit results of the three negative Binomial statistical models – the Network, Spatial and Attraction model. We will compare them in relation to their explanatory and

[25] That is, we plot the values $\widehat{e_i}$ sorted in increasing size against the sorted e_i values.

[26] The CRPS value is computed by approximating the expected values by Monte Carlo averages. For example, $\mathbb{E}_X\left(|y_i - X|\right)$ is approximated by $\frac{1}{m}\sum_{j=1}^{m} |\mathrm{y_i} - \mathrm{X^j}|$, where X$^\mathrm{j}$ is a draw from the distribution of y_i and we use $\mathrm{m} = 10\ 000$.

predictive power, also considering the number of variables included in each model. We will then compare the coefficient estimates in each model individually. Finally, we will compare the explanatory and predictive power of the two statistical modelling solutions, the negative Binomial and the regular logarithmic regression model, used in a lot of studies in the field. Whenever needed, references to relevant statistical tests which scrutinize the validity of the models will be used.

3.1. Negative Binomial models. Model fit and coefficients

When the three negative Binomial models are compared (Fig3), using the predictive performance indicators, we see that R^2 values are comparable for the three model and the value only slightly rises as we move from the Configuration model (0,649) to the Spatial model (0,652) and finally to the Attraction model (0,661). This suggests that a model including only the two Angular centrality measures can explain a large part of the Intensity of the pedestrian flows. The prediction accuracy improves only slightly as we add more variables. What is also important is that while the simple correlations for the network variables independently in each city did not show consistent high correlations (Fig2), the R2 when both Angular Integration and Betweenness are included increased significantly. The CRPS values show that the Configurational and spatial models have very similar predictive power, but that the Attraction model has a slightly higher predictive power. Thus, in summary there is no gain in using the Spatial model instead of the simpler Configurational model, but the accuracy of the predictive distributions is increased if the Attraction model is used. Looking at the Q-Q plots for each model (Fig4), we see no clear indication for any of the models that they have the wrong residual distributions.

	average CRPS	R^2
Configurational	-522.7396004	0.6494646
Spatial	-522.7752185	0.6528977
Attraction	-510.4907265	0.6616139

Figure 3. Model scores and R^2 for the negative Binomial models. CRPS are estimated from 10^5 simulations.

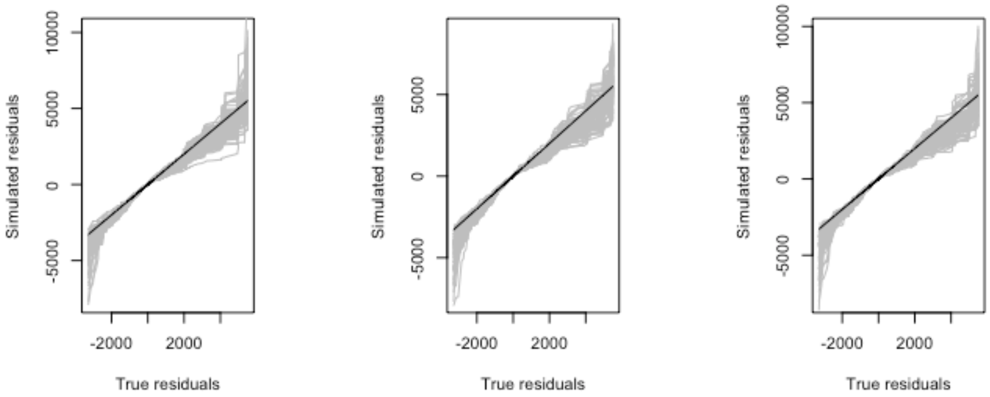
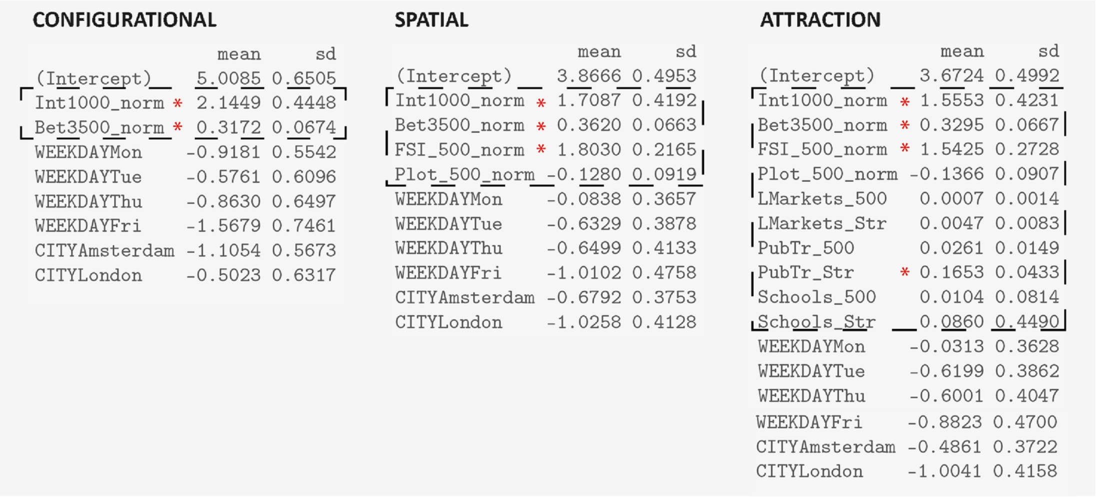


Figure 4. Q-Q plots for the negative Binomial model (left to right: Configurational, Spatial, Attraction)

If we then look at each model separately we can see the coefficient estimates (Fig5). The coefficients which are significant at a level of 0.05 are marked with a star.



CONFIGURATIONAL

```
                    mean     sd
(Intercept)       5.0085 0.6505
Int1000_norm *    2.1449 0.4448
Bet3500_norm *    0.3172 0.0674
WEEKDAYMon       -0.9181 0.5542
WEEKDAYTue       -0.5761 0.6096
WEEKDAYThu       -0.8630 0.6497
WEEKDAYFri       -1.5679 0.7461
CITYAmsterdam    -1.1054 0.5673
CITYLondon       -0.5023 0.6317
```

SPATIAL

```
                    mean     sd
(Intercept)       3.8666 0.4953
Int1000_norm *    1.7087 0.4192
Bet3500_norm *    0.3620 0.0663
FSI_500_norm *    1.8030 0.2165
Plot_500_norm    -0.1280 0.0919
WEEKDAYMon       -0.0838 0.3657
WEEKDAYTue       -0.6329 0.3878
WEEKDAYThu       -0.6499 0.4133
WEEKDAYFri       -1.0102 0.4758
CITYAmsterdam    -0.6792 0.3753
CITYLondon       -1.0258 0.4128
```

ATTRACTION

```
                    mean     sd
(Intercept)       3.6724 0.4992
Int1000_norm *    1.5553 0.4231
Bet3500_norm *    0.3295 0.0667
FSI_500_norm *    1.5425 0.2728
Plot_500_norm    -0.1366 0.0907
LMarkets_500      0.0007 0.0014
LMarkets_Str      0.0047 0.0083
PubTr_500         0.0261 0.0149
PubTr_Str      *  0.1653 0.0433
Schools_500       0.0104 0.0814
Schools_Str       0.0860 0.4490
WEEKDAYMon       -0.0313 0.3628
WEEKDAYTue       -0.6199 0.3862
WEEKDAYThu       -0.6001 0.4047
WEEKDAYFri       -0.8823 0.4700
CITYAmsterdam    -0.4861 0.3722
CITYLondon       -1.0041 0.4158
```

Figure 5. Table of the mean values of the coefficient estimates for the negative Binomial models. Significant variables are marked with a star

In the Configurational model, we see that Angular Integration (1000m) has a larger effect on the intensity of pedestrian movement than Angular Betweenness (3500m), but both are significant[27].

In the Spatial model, when Accessible FSI (indicator of built density) and Accessible number of plots (indicator of land division) were added to the model, Accessible FSI is the more important variable, followed by Angular Integration and Angular Betweenness; Accessible number of plots is not significant.

In the Attraction model, when all the attraction variables are added, only one attraction variable is significant, that is the Number of Public transport nodes on the street segment, but with less impact than both the configurational variables and density. The accessible number of plots and the other five attraction variables are not significant.

3.2. The importance of control variables and the random effect

To control for the possible impact of weekday or variation in the findings between cities, two control variables (weekday and city) were added in each model as discussed in Section 2.4. Figure 5 shows that none of the control variables are significant, meaning that we do not see an effect of the day of the week the survey was conducted, neither do we see significant differences between cities.

Looking at the residuals for the Configurational model in London (Figure 6), the top two plots show the residuals of the negative Binomial model with only fixed effects. There is no indication of a direct linear relation between latitude or longitude and the residuals, but it seems that we have correlation between residuals corresponding to streets in the same area. This is indicated by the fact that residuals for observations with very close long-/latitudes seem to all fall on the same side of 0 (i.e. where residual

[27] Remember that our chosen radii were decided after correlating the pedestrian counts to 10 different radii (see section 2.4). The same radii were used in all models.

value is 0) more frequently than is expected to happen by chance. The bottom two plots show the final model where we have added a random intercept for each neighbourhood in the data, which improves the distribution of residuals; the residuals for each group are (more or less) symmetric around zero. The inclusion of the random effect therefore seems to handle the possible spatial structure (spatial autocorrelation) in the data. Similar results were found in the other cities (Amsterdam and Stockholm) and the other models (Spatial and Attraction model).

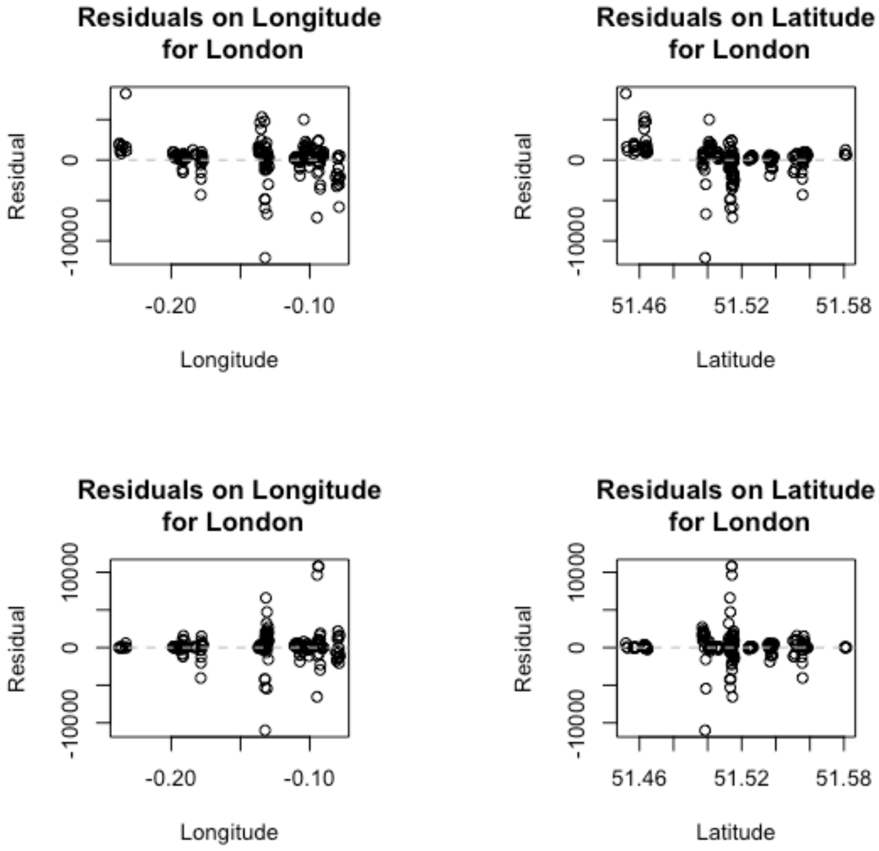


Figure 6. Residuals plotted on coordinates for the Configurational model in London, with only fixed effects (top two plots) and with added random effects (bottom two plots).

Further, Figure 7 shows Q-Q plots for the residuals for the three models without random effects, where we can see that the models seem to have a worse fit than the models with the random effects included (remember Fig4). Here the straight line goes outside the band of simulated residuals for low values (indicating that the lower tail of the distribution is wrong).

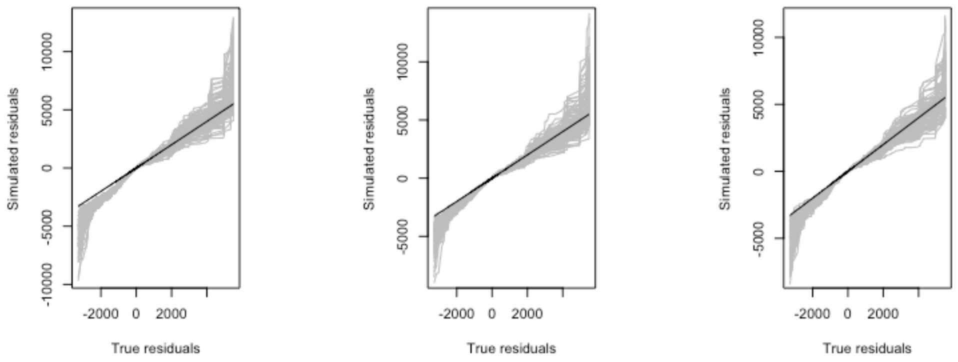


Figure 7. Q-Q plots for the negative Binomial model without the random effects (from left to right: Configurational, Spatial, Attraction)

3.4. Comparison of negative Binomial model results to Logarithmic regression model results

The comparison of the results of the negative Binomial model with the results of a typical logarithmic regression model are discussed here, because regression models are the most commonly used models in the field of space syntax. Although this is not the optimum model in principle as described in section 2.4, it is useful to see if the results change significantly compared to the negative Binomial model. Note that we included the same explanatory and control variables, as well as the random effect.

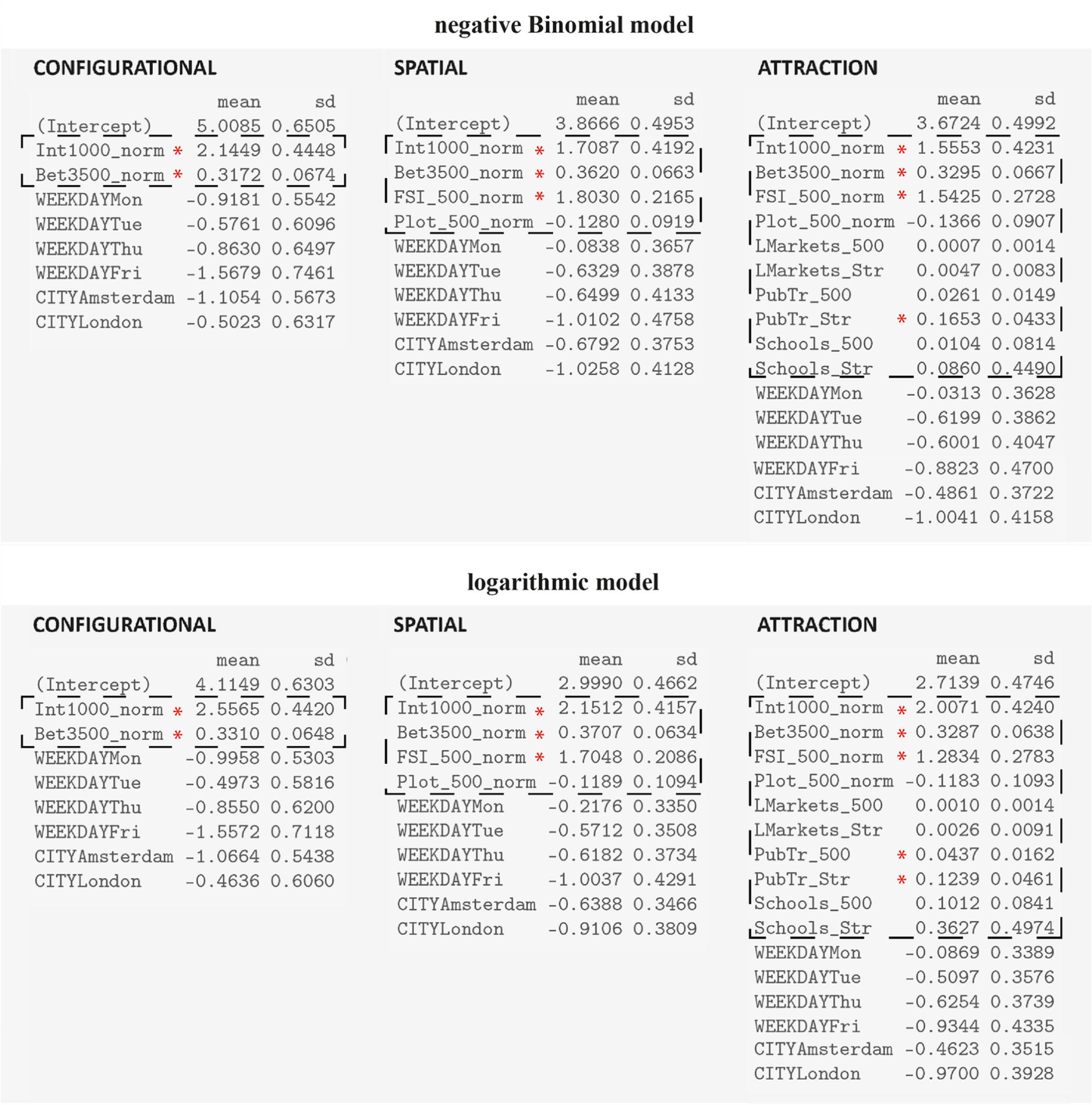
		average CRPS	R^2
negative Binomial	**Configurational**	-522.7396004	0.6494646
	Spatial	-522.7752185	0.6528977
	Attraction	-510.4907265	0.6616139
logarithmic	**Configurational**	-601.4840254	0.6374631
	Spatial	-597.4619654	0.6720471
	Attraction	-591.6143639	0.6463123

Figure 8. Model scores and R^2 for the models. Top half: Negative binomial models. Bottom half: Logarithmic models. The CRPS scores are estimated from 10^5 simulations.

Figure 8 shows the CRPS scores and the R^2 values for the negative Binomial and the logarithmic regression model, for the Configurational, Spatial and Attraction models. We see very similar values between the two modelling solutions. The only thing that changes is that the R^2 increases slightly when we add more variables to the negative Binomial model (i.e. from the Configurational to the Spatial to the Attraction model), while in the regression model, the Spatial model has the highest R^2. Overall, the differences are minimal. However, we see that CRPS values are higher for the negative Binomial models than for the logarithmic models, meaning that the negative Binomial models give more reliable results when also taking uncertainty into account.

Thus, in conclusion the negative Binomial models are preferable. However, it might be of interest to also compare the coefficient estimates of the logarithmic regression model to the negative Binomial model (Fig 9) to see if the worse fit of the logarithmic regression affects the conclusions about the explanatory variables. Both the significance of the variables as well as their relative importance does not change. The only minor changes are, first, that for the Spatial model in the logarithmic regression, Integration has a higher importance than FSI, while in the negative Binomial model, it is the opposite.

Another difference is that for the Attraction model, in the logarithmic regression, one more attraction variable becomes significant; that is the Accessible number of Public transportation nodes in 500m. Thus, if one only is interested in finding which explanatory variables that are most important for explaining the pedestrian counts, it does not matter much whether the logarithmic regression or the negative Binomial model is used.



negative Binomial model

CONFIGURATIONAL

		mean	sd
(Intercept)		5.0085	0.6505
Int1000_norm	*	2.1449	0.4448
Bet3500_norm	*	0.3172	0.0674
WEEKDAYMon		-0.9181	0.5542
WEEKDAYTue		-0.5761	0.6096
WEEKDAYThu		-0.8630	0.6497
WEEKDAYFri		-1.5679	0.7461
CITYAmsterdam		-1.1054	0.5673
CITYLondon		-0.5023	0.6317

SPATIAL

		mean	sd
(Intercept)		3.8666	0.4953
Int1000_norm	*	1.7087	0.4192
Bet3500_norm	*	0.3620	0.0663
FSI_500_norm	*	1.8030	0.2165
Plot_500_norm		-0.1280	0.0919
WEEKDAYMon		-0.0838	0.3657
WEEKDAYTue		-0.6329	0.3878
WEEKDAYThu		-0.6499	0.4133
WEEKDAYFri		-1.0102	0.4758
CITYAmsterdam		-0.6792	0.3753
CITYLondon		-1.0258	0.4128

ATTRACTION

		mean	sd
(Intercept)		3.6724	0.4992
Int1000_norm	*	1.5553	0.4231
Bet3500_norm	*	0.3295	0.0667
FSI_500_norm	*	1.5425	0.2728
Plot_500_norm		-0.1366	0.0907
LMarkets_500		0.0007	0.0014
LMarkets_Str		0.0047	0.0083
PubTr_500		0.0261	0.0149
PubTr_Str	*	0.1653	0.0433
Schools_500		0.0104	0.0814
Schools_Str		0.0860	0.4490
WEEKDAYMon		-0.0313	0.3628
WEEKDAYTue		-0.6199	0.3862
WEEKDAYThu		-0.6001	0.4047
WEEKDAYFri		-0.8823	0.4700
CITYAmsterdam		-0.4861	0.3722
CITYLondon		-1.0041	0.4158

logarithmic model

CONFIGURATIONAL

		mean	sd
(Intercept)		4.1149	0.6303
Int1000_norm	*	2.5565	0.4420
Bet3500_norm	*	0.3310	0.0648
WEEKDAYMon		-0.9958	0.5303
WEEKDAYTue		-0.4973	0.5816
WEEKDAYThu		-0.8550	0.6200
WEEKDAYFri		-1.5572	0.7118
CITYAmsterdam		-1.0664	0.5438
CITYLondon		-0.4636	0.6060

SPATIAL

		mean	sd
(Intercept)		2.9990	0.4662
Int1000_norm	*	2.1512	0.4157
Bet3500_norm	*	0.3707	0.0634
FSI_500_norm	*	1.7048	0.2086
Plot_500_norm		-0.1189	0.1094
WEEKDAYMon		-0.2176	0.3350
WEEKDAYTue		-0.5712	0.3508
WEEKDAYThu		-0.6182	0.3734
WEEKDAYFri		-1.0037	0.4291
CITYAmsterdam		-0.6388	0.3466
CITYLondon		-0.9106	0.3809

ATTRACTION

		mean	sd
(Intercept)		2.7139	0.4746
Int1000_norm	*	2.0071	0.4240
Bet3500_norm	*	0.3287	0.0638
FSI_500_norm	*	1.2834	0.2783
Plot_500_norm		-0.1183	0.1093
LMarkets_500		0.0010	0.0014
LMarkets_Str		0.0026	0.0091
PubTr_500	*	0.0437	0.0162
PubTr_Str	*	0.1239	0.0461
Schools_500		0.1012	0.0841
Schools_Str		0.3627	0.4974
WEEKDAYMon		-0.0869	0.3389
WEEKDAYTue		-0.5097	0.3576
WEEKDAYThu		-0.6254	0.3739
WEEKDAYFri		-0.9344	0.4335
CITYAmsterdam		-0.4623	0.3515
CITYLondon		-0.9700	0.3928

Figure 9. Table of the mean values of all the coefficient estimates for all the models. Top half: Negative binomial models. Bottom half: Logarithmic models. Significant variables are marked with a star

4. CONCLUSIONS

There are three major areas that we see our study contributing to. First, the high and consistent correlations between spatial form and pedestrian movement, in a study of unprecedented size, comprising three cities and including a large range of neighbourhoods of varying morphological types, offer convincing proof that the tested morphological variables have a strong impact on the spatial distribution of pedestrian flows in cities. This is a vital finding, confirming a large number of earlier studies that albeit rigorous within their given frameworks have either been far smaller, not comprised the same range of urban types, or have not offered rigorous comparability between cities.

Second, the study shows that the model with all explanatory variables has the highest explanatory power and the best model fit. The relative importance of the explanatory variables shows that angular integration and accessible FSI effect the number of pedestrians equally strong and significantly more than angular betweenness, while plot size is not significant. The study also showed that the combination of the two angular centrality measures is far more powerful in predicting, than Angular integration alone. Of the attraction variables, only the presence of a public transport nodes in the street has a significant, but small effect.

Third, this study contributes to the advancement of methodology, first concerning the comparability of GIS-models used, second, the method that allows for large scale pedestrian surveys and the statistical modelling that is suitable for the specificities of spatial data, and pedestrian data in particular. Related to the last case, we propose the use of the negative Binomial model instead of linear regression models since it avoids the ad hoc procedure of logging the data. Further, and perhaps more importantly, regression is a model for continuous data, which pedestrian counts are not. We moreover addressed the problem of spatial correlation in the data, which makes the results unreliable if not accounted for. As a solution it is proposed to include a random effect that models the spatial dependence. Finally, we propose to use CRPS for comparing models in addition to R^2-values. While R^2-values show how well the model predicts pedestrian counts, it does not reflect the uncertainty of the predictions, something CRPS can do.

Regarding the choice between using negative Binomial regressions or standard multiple regression models for transformed data, we saw for our data that the negative Binomial models had higher predictive accuracy compared to the standard regression. However, the conclusions drawn regarding the effect of the different explanatory variables did not change much between the two choices. It should be noted here that the importance of using a proper regression model for count data should increase for datasets with higher proportions of low counts, since in that case, the approximation that is done when using a regression model for continuous data on counts will be worse. Thus, the conclusion is that proper model validation checks have to be performed if a standard regression model is used for count data, or for data where there may be remaining spatial dependence in the residuals of the regressions. Blindly using regression models in those cases can lead to inaccurate conclusions.

Besides these three achievements, the study also reveals some new questions that could be studied next, three of which we want to highlight here. First, although this study comprises three cities, including a large range of neighbourhoods of varying morphological types, it is of importance to look beyond the European context and add cities in other continents to study the generalisability of the findings across cities. Related to this, a study of the specificities within cities would be an interesting addition to the current study, where differences between neighbourhoods could be in focus. Second, staying closer to the material at hand, it is of interest to look into the pedestrian survey in more detail and also study the fluctuation of pedestrian flow during the day and variations in speed in relation to urban form. Third, some statistical deeper investigations are needed. A next step for improving the statistical models used here is to replace the independent random effects that were used for each neighbourhood by a random field that also can capture the dependence between these random effects. Again, an interesting topic, also for future statistical research, would be to model the fluctuation of pedestrian flows during the day, which would require replacing the regression models with models for functional data.

REFERENCES

Abedi, N., Bhaskar, A. and Chung, E. (2013). Bluetooth and Wi-Fi MAC Address Based Crowd Data Collection and Monitoring: Benefits, Challenges and Enhancement. *Australasian Transport Research Forum*, ATRF 2013 - Proceedings.

Barbera, V., M., Epasto, A., Mei, A., Perta, C., V., and Stefa, J. (2013). Signals from the crowd: Uncovering social relationships through smartphone probes. *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, IMC. 265-276. 10.1145/2504730.2504742.

Berghauser Pont, M., Stavroulaki, G., Sun, K., Abshirini, E., Olsson, J., Marcus, L. (2017). Quantitative comparison of the distribution of densities in three Swedish cities. *Proceedings 24th International Seminar on Urban Form*, Valencia.

Berghauser Pont, M. and Marcus, L. (2015). What can typology explain that configuration cannot? *The 10th Space Syntax Symposium* (SSS10), 13-17 July 2015 at UCL, 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.

Berghauser Pont, M. and Haupt, P. (2010), *Spacematrix. Space, density and urban form*. Rotterdam: NAi Publishers.

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.

Gehl, J. (2010). *Cities for people.* Washington: Island Press.

Gneiting, T. and Raftery, A., E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* 102(477). 359–378.

Hillier, B., (1996). *Space is the machine. A configurational theory of architecture*. Cambridge: Cambridge University Press

Hillier, B., (2012), Studying cities to learn about minds: some possible implications of space syntax for spatial cognition, Environment and Planning B, Vol. 39, nr. 1,12-32.

Hillier, B., Leaman, A. (1973) The man-environment paradigm and its paradoxes, *Architectural Design*, nr. 8, 507-511.

Hillier, B., Hanson, J. (1984) *The social logic of space*, Cambridge University Press, Cambridge, UK.

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. 155-193.

Hillier, B. and Iida, S. (2005). Network and psychological effects in urban movement. *Proceedings of the Fifth International Space Syntax Symposium,* Delft: University of Technology.

Hillier, B., A. Penn, J. Hanson, T. Grajewski, J. Xu (1993). Natural movement: or, configuration and attraction in urban pedestrian movement. *Environment and Planning B: Planning and Design*, vol. 20, 29-66.

Hillier, B., Burdett, R., Peponis, J., Penn, A., (1987). Creating life: or, does architecture determine anything? *Architecture and Comportement/Architecture and Behaviour*. 3 (3) 233-250.

Kurkcu, A., and Ozbay, K. (2017). Estimating Pedestrian Densities, Wait Times and Flows using Wi-Fi and Bluetooth Sensors. *Transportation Research Record Journal of the Transportation Research Board*. 2644. 10.3141/2644-09.

Legeby, A. (2013). *Patterns of co-presence. Spatial configuration and social segregation*. Stockholm: KTH University

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

Marcus, L., Berghauser Pont, M., and Bobkova, E. (2017). 'Cities as Accessible Densities and Diversities: Adding attraction variables to configurational analysis'. *Proceedings of the 11th International Space Syntax Symposium*. Lisbon: Instituto Superior Técnico.

Ozbil, A., Peponis, J., and Stone, B. (2011) Understanding the link between street connectivity, land use and pedestrian flows. *Urban Design International* 16, 125-141

Ozbil, A., Yesiltepe, D. and 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. 189-207.

Penn, A., Hillier, B., Banister, D. and Xu, J., (1998) Configurational modelling of urban movement networks. *Environment and Planning B: Planning and Design*. 24. 59-84.

Petre, A-C., Chilipirea, C., Baratchi, M., Dobre, C., and van Steen, M. (2017). WiFi Tracking of Pedestrian Behavior. In *Smart Sensors Networks*. (eds. Xhafa F., Leu, F-Y., Hung L-L.), Academic Press, 309-337, 10.1016/B978-0-12-809859-2.00018-8.

Peponis, J., Hadjinikolaou, E., Livieratos, C., and A Fatouros, D. (1989). The spatial core of urban culture. *Ekistics*. 56. 43–55.

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

Read S, (1999), Space syntax and the Dutch city. *Environment and Planning B: Planning and Design* 26. 251-264.

Rue H., Martino S., and Chopin N. (2009), Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion). *Journal of the Royal Statistical Society B*. 71. 319-392.

Schauer, L., Werner, M., Marcus, P. (2014). Estimating Crowd Densities and Pedestrian Flows Using Wi-Fi and Bluetooth. *MobiQuitous 2014 - 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. 10.4108/icst.mobiquitous.2014.257870.

Ståhle, A., Marcus, L. and Karlström, A. (2005). Place syntax—geographic accessibility with axial lines in gis. *Proceedings of the 11th International Space Syntax Symposium*. Delft: TU Delft. 131-144.