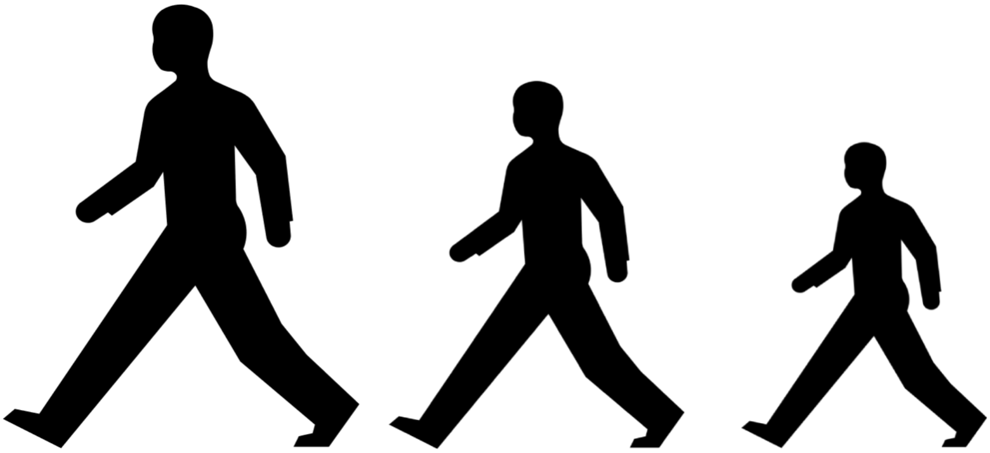

Predicting Pedestrian Counts per Street Segment in Urban Environments

Master’s thesis in Computer science and engineering

SIMON KARLSSON

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2020

Master's thesis 2020

Predicting Pedestrian Counts per Street Segment in Urban Environments

SIMON KARLSSON

UNIVERSITY OF GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2020

Predicting Pedestrian Counts per Street Segment in Urban Environments
SIMON KARLSSON

Supervisor: Selpi, Department of Mechanics and Maritime Sciences
Advisor: Gianna Stavroulaki, Department of Architecture and Civil Engineering
Examiner: Graham Kemp, Department of Computer Science and Engineering

Master's Thesis 2020
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Illustration of people walking.

Typeset in LaTeX
Gothenburg, Sweden 2020

Predicting Pedestrian Counts per Street Segment in Urban Environments
SIMON KARLSSON
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

Cities are continuously growing all over the world and the complexity of designing urban environments increases. Therefore, there is a need to build a better understanding in how our cities work today. One of the essential parts of this is understanding the pedestrian movement. Using pedestrian count data from Amsterdam, London and Stockholm, this thesis explore new variables to further explain pedestrian counts using negative binomial and random forest. The models explored includes variables that represent street centrality, built density, land division, attractions and the road network. The result of the thesis suggests ways for variables to be represented or created to increase the explanatory value in regards to pedestrian counts. These suggestions include: including street centrality measurements at multiple scales, attraction counts within the surrounding area instead of counts on the street segment, counting attractions instead of calculating the distance to the nearest attraction, using network reach to constrain the network at different scales instead of bounding box, and counting intersections in the road network instead of computing the network length.

Keywords: data science, pedestrian movement, machine learning, random forest, negative binomial, spatial morphology, road network, street centrality, built environment, built density, attractions, land division.

Acknowledgements

I am truly grateful Selpi, for the time and energy you have spent on helping me ask the right questions, providing in depth feedback and guiding me through the difficult task of writing a thesis, and all of that on top of learning about a domain completely new to you.

Many thanks Gianna Stavroulaki for sharing data, invaluable knowledge, good feedback and allowing me to do this thesis.

Thank you Meta Berghauser Pont and Evgeniya Bobkova for allowing me to reuse your illustrations. There is no doubt that readers will appreciate them as well.

Simon Karlsson, Gothenburg, March 2020

Contents

List of Figures

List of Tables

1
Introduction

1.1 Background

It is important for urban designers, planners and policy-makers to create lively streets and neighbourhoods because, as stated by Edwards and Tsouros [11], an active city increases public health, social interactions and also contributes to a stronger economy. The means of achieving this are, however, still either unclear or not concrete enough. The complexity is also increasing because of the substantial population growth in cities. In 2016, United Nations [32] estimated more than half of the world's population to be living in cities and that this percentage was increasing. It is therefore important, more now than ever, that we expand our understanding of how urban environments function so that we can make incremental improvements.

There has been many contributions towards this, two of which are Stavroulaki et al. [30] and Håkansson [17]. They focus on understanding pedestrian movement, which is an essential part of urban environments. They did this by building predictive models for pedestrian movement counts per street segment. Stavroulaki et al. [30] predicted the full day pedestrian movement counts using negative binomial models. Håkansson [17] predicted the hourly fluctuations during the day using what they refer to as a *functional ANOVA negative binomial model with logarithmic link*.

These predictive models used meta data about the surrounding area in order to predict pedestrian movement. This meta data included things like, how central a street is, how dense an area is built and how accessible public transport is.

1.2 Motivation

Exploring predictive models for pedestrian movement counts is interesting because it can give insight into how the built environment affects the actual usage and movement within it. A deeper understanding of this relationship would mean a possibility to alter or create built environments to enable an increase of activity within the city.

Even though these type of predictive models have already been created by Stavroulaki et al. [30] and Håkansson [17] there is still room for improvement. The performance of the models in Stavroulaki et al. [30], the predictive models for full day pedestrian counts, achieves an R^2 score of approximately 0.65. This can be interpreted as the models explaining 65 percent of the variance in the pedestrian movement counts.

This then means that 35 percent of the pedestrian movement counts are not yet explained; this is the main reason why it is useful to further explore improvements in these predictive models.

1.3 Objective

The focus of this thesis is to analyse and also extend the work done by Stavroulaki et al. [30] and Håkansson [17]. This is done using the same data, collected by Stavroulaki et al. [29] and Berghauser Pont et al. [3] using a service offered by Bumbee labs, Stockholm. The objectives in extending these predictive models are the following:

- Evaluate random forest as an alternative algorithm to negative binomial.
- Explore new variables to further explain pedestrian movement counts.
- Find, from the variables available, the set of variables that best explain the pedestrian movement counts.

The new variables that are explored in this thesis are still focused on the built environment. Most of them are the same type of variables as used in Stavroulaki et al. [30] but represented in different ways. The aim of the variables used is to describe street centrality, built density, land division and attractions. The meaning of these concepts are described in Chapter 3.

1.4 Outline

This section gives brief explanations of what is presented in each of the following chapters.

Chapter 2 introduces the data used. Chapter 3 presents variables used to represent the built environment, related research and also algorithms and metrics used. Chapter 4 presents common methodology for all the experiments. Chapter 5 reproduces previous work from Stavroulaki et al. [30]. Chapter 6 explores different representations of street centrality. Chapter 7 explores different representations of attractions. Chapter 8 explores new variables based on road network. Chapter 9 combines the findings from each of the previous experiments, then designs and evaluates the final model. Chapter 10 includes discussion of the results and also a section on ethical considerations. Chapter 11 summarizes the thesis with conclusion and future work.

2

Data

This chapter introduces the data used, e.g., what the data is, how it was recorded, the possible limitations in this data and the variables created previous to this thesis.

2.1 Collection

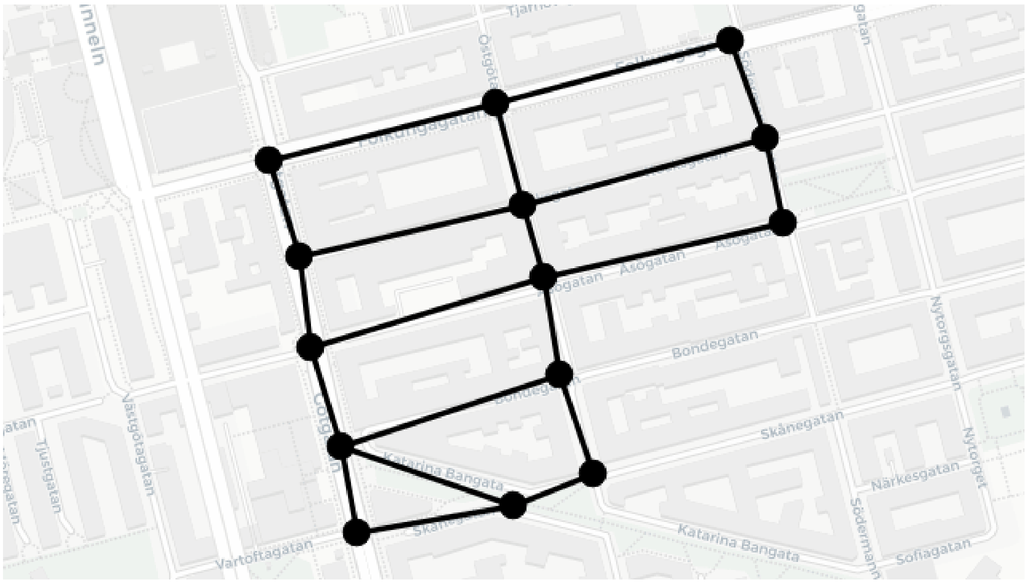

Figure 2.1: Placements of Wi-Fi tracking sensors in Södermalm in Stockholm.
The dots represent the placement of the sensors. The lines represent street segments.
Attribution: Leaflet[1] | mplleaflet[33] | © OpenStreetMap[23] © CartoDB[9]

This thesis will make use of data collected by Stavroulaki et al. [29] and Berghauser Pont et al. [3], using a service offered by Bumbee labs, Stockholm. During three weeks in October 2017, they collected the data by tracking anonymized Wi-Fi signals from mobile phones. They did this by placing Wi-Fi tracking sensors in the intersections in an area. In total, this was done in around 60 areas for one day each. The areas included are from three cities, Stockholm, London and Amsterdam. See Figure 2.1 for an example of how the sensors were located in an area.

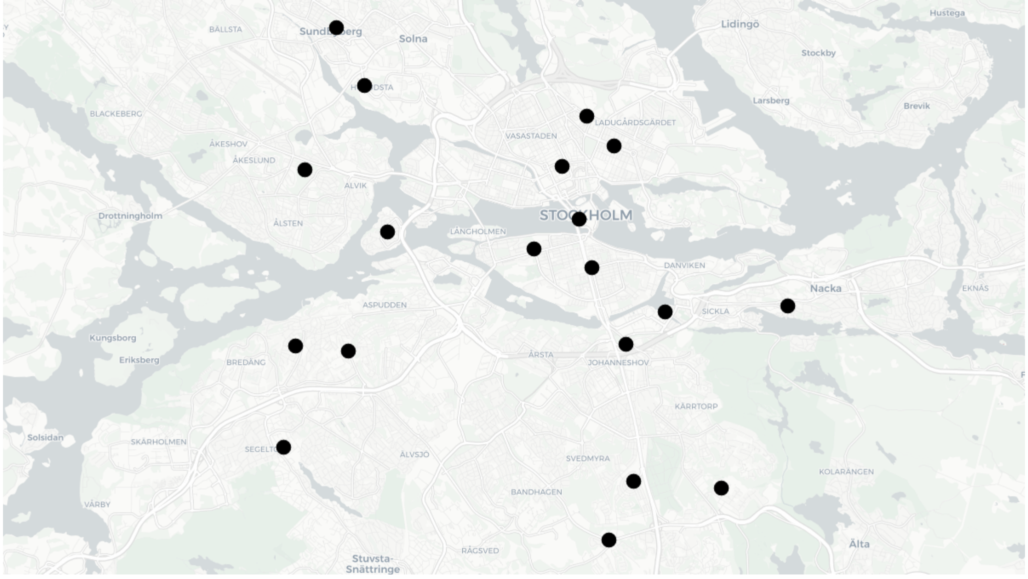

Figure 2.2: Areas included in the measurement in Stockholm.
Attribution: Leaflet[1] | mplleaflet[33] | © OpenStreetMap[23] © CartoDB[9]

Each of the areas were monitored for one day and the areas were selected as to include a diversity in type of area, e.g., how central the areas are. The diversity of areas can be understood when looking at the spread in the distribution of areas in Stockholm as visualized in Figure 2.2.

An example for the data collected in one of the areas can be seen in Table 2.1. As shown in the table, each of the “visits” at a node is recorded with an id of the visitor, an id of the gate/sensor and a position of the gate in form of X and Y coordinates. The coordinates here uses the EPSG:3006 coordinate system, also referred to as SWEREF99 TM.

2.2 Data processing

Using the data with Wi-Fi sensors at each intersection, it possible to calculate which visitors that passed through a specific street segment. This has been done previous to this thesis to create a data set that contains the counts of visitors per street segment. The amount of monitored street segments sum up to approximately 300 in each of the cities.

From knowing which visitors passed through which street segments, a table was created where each row corresponds to a street segment. This aggregation per street segment was for full day counts and hourly counts. It was also done per direction for both full day counts and hourly counts.

During the creation of this table, there was some preprocessing performed. This preprocessing included scaling, extrapolation and filtering. The following Sections

Table 2.1: An overview of the data collected by tracking Wi-Fi signals.

visit_id	gate_id	timestamp	X	Y
0114_1	114197	2017-10-05 06:00:30	675092.947832	6579125.31807
0114_1	114198	2017-10-05 06:02:40	675270.5775479999	6579181.03863
0114_2	114196	2017-10-05 06:02:50	674922.099638	6579073.26476
0114_2	114197	2017-10-05 06:03:40	675092.947832	6579125.31807
0114_3	114197	2017-10-05 06:03:40	675092.947832	6579125.31807
0114_3	114196	2017-10-05 06:03:50	674922.099638	6579073.26476
0114_4	114205	2017-10-05 06:05:10	674986.373212	6578857.9952
0114_4	114201	2017-10-05 06:05:30	674947.89521	6579001.07808
0114_4	114196	2017-10-05 06:05:50	674922.099638	6579073.26476
0114_5	114209	2017-10-05 06:06:00	675002.666861	6578792.841519999
0114_5	114205	2017-10-05 06:06:20	674986.373212	6578857.9952
0114_6	114196	2017-10-05 06:01:40	674922.099638	6579073.26476
0114_6	114201	2017-10-05 06:09:20	674947.89521	6579001.07808
0114_6	114200	2017-10-05 06:11:00	675117.7263859999	6579048.06709
0114_7	114203	2017-10-05 06:08:30	675136.4088229999	6578994.175969999
0114_7	114197	2017-10-05 06:09:40	675092.947832	6579125.31807
0114_7	114200	2017-10-05 06:09:50	675117.7263859999	6579048.06709

visit_id is a unique id of an anonymized pedestrian.
gate_id is a unique id of the sensor, also referred to as gate.
timestamp is the time when the pedestrian was recorded.
X and *Y* marks the position of the sensor.

explain the reason and process for all of these preprocessing methods. Note that this processed and aggregated data is what was used to build the previous models in Stavroulaki et al. [30] and Håkansson [17], it is also the representation of the data that is used to create the predictive models in this thesis.

2.2.1 Scaling

The reason for scaling the data is simply because the gates do not capture all the pedestrians. This is because the measurements are dependent on the pedestrian having a phone with Wi-Fi turned on. So in order to know how many pedestrians that actually walked past an intersection, manual measurements were performed simultaneously for a few select street segments. These measurements then resulted in using a scaling of 2.3 for all the street segments in all the cities.

2.2.2 Extrapolation

The reason for applying extrapolation on the data is that some of the gates were stolen, vandalized, stopped working or missed a pedestrian. Three different extrapolation methods were used: based on time-frames, based on neighbouring gates and based on path. The extrapolation based on time-frames was to assume that the count for a gate during its downtime was similar to the count before and after the downtime, this method was used only for time-frames of one hour or shorter. The extrapolation based on neighbouring gates was to calculate the number of visitors based on surrounding gates. It is worth to note that this method was only used for gates that were completely surrounded by other gates. The extrapolation based on path was done when one visitor seemingly skipped one of the gates on a straight path when there was no other way to go.

2.2.3 Filtering

The reason for filtering the data was two-fold. The first reason was that a gate had too much downtime and none of the extrapolation methods worked, that gate was then removed completely. The second reason was that some of the measurements indicated movement speeds that would not be possible for a pedestrian to reach. Therefore, all the measurements that exceeded a speed of 6 km/h were removed.

2.3 Limitations

Using Wi-Fi sensors to collect this type of data creates a few biases in the data. It does this because of the need of having a phone with Wi-Fi turned on. This means that only people that have phones with Wi-Fi turned on are included in the measurement. Likewise, people that have multiple phones with Wi-Fi turned on are measured multiple times.

These Wi-Fi sensors also capture signals from phones within buildings, as long as it is in within proximity. However, this mostly affects the recorded pedestrians at single gates and not the counts per street segment since the same device would have to be captured at a neighbouring gate as well to be counted.

There are also some uncertainties in the data collected using Wi-Fi sensors. Firstly, there is no clear differentiation between different modes of transport. Secondly, the exact position is not known, the sensor measures within a radius of 25 meters. Thirdly, the exact time is not known, the recorded sensor data is presented with a granularity of 10 seconds, as seen in Table 2.1. These limitations are further explored in Section 2.3.1.

For this specific data collection, each area was monitored for one day (only workdays). This limits the probability that the measured pedestrian movement count on a street segments is representative. E.g., there could be an event happening in an area on the day of the measurement which would greatly affect the pedestrian count.

Another possible limitation with this data collection is the sample size. The data contains only 700 street segments with full day counts, i.e., the sample size is 700. To determine if this is a limitation or not is, however, very difficult. It could be a limitation if the relationship between the pedestrian movement counts and the built environment is complex. This means that it might be difficult for a predictive model to find this relationship between them. If this is the case, a larger data set could help in giving a better indication of what the relationship actually is.

A quick summary of this section gives the following limitations of the collected data:

- People passing the area without a phone or Wi-Fi activated are not counted.
- A person with more than one phone having Wi-Fi turned on is measured multiple times.
- Cannot clearly differentiate between different modes of transport (e.g. walking, biking, driving).
- Position is within a 25 meter radius.
- Time is presented with 10 second granularity.
- Each area is only recorded one day.
- Limited data size.

2.3.1 Speed of walking

Speed of walking was calculated previous to this thesis in order to do filtering on the data, as explained in Section 2.2.3. These calculations are, however, not available for use during this thesis so speed of walking is re-calculated. It is, as previous calculation also was, calculated for each pedestrian and street segment and this is done using the raw data as opposed to the processed data which was described in Section 2.2. This is done by using the length of the street segment and the duration between visiting one of the sensors up until visiting the other sensor.

Unfortunately, the calculated speed of walking has a big error margin. This is because each of the gates that has been used during the measurement has a radius of 25 meters, and the timestamps for the measurements has a granularity of 10 seconds. This means that when calculating the speed for a pedestrian that has walked a street segment that is 100 meters, they could in reality have walked anything between 50 and 150 meters. Similarly if the time spent on the street segment is measured to be 80 seconds, it could in reality be that the duration was anything between 70 to 90 seconds. The reason for this is because each timestamp can be off by 5 seconds and therefore the duration can be off by 10 seconds. Both of these uncertainties contributes to an uncertainty in the calculations of walking speed, and in short distances and/or durations the range of uncertainty can prove to be quite large.

As mentioned, there is a filtering performed based on the speed of movement. The filtering excludes all measurements that show a speed of movement above 6 km/h. This leads to only keeping 23 percent of the raw measurements. See Figure 2.3 which shows a histogram of the movement speeds. The 23 percent of the data that is kept is on the left side of the line which is drawn at 6 km/h. This seems like a reasonable threshold when considering the average walking speed of younger (younger than 65 years) and older people (65 years or older) being 4.90 km/h (1.36 m/s) and 4.10 km/h (1.14 m/s) respectively, as reported in Montufar et al. [20].

2.4 Variables

Values for some variables for each street segment were calculated previous to this thesis. Most of these variables are calculated to describe the surrounding area of the street segment. The surrounding area is limited by a threshold of walking

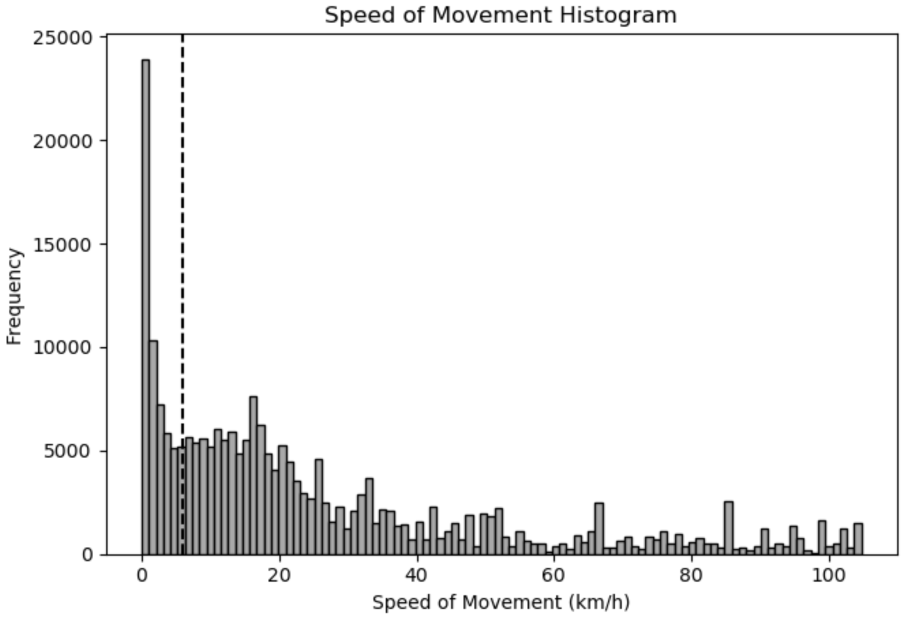

Figure 2.3: Movement speed histogram for the raw measurements. The movement speed is per "pedestrian" and street segment. The dotted line is drawn at 6 km/h, which is the filtering threshold. The top one percent highest values are excluded from the histogram in order to have a more concentrated plot.

distance. The street segments reachable within this threshold are then included in the calculations. For example, when using a threshold of 500 meters, then all street segments possible to reach by walking 500 meters or less are included. In this thesis, and in Stavroulaki et al. [30] amongst others, this threshold is referred to as the radius. See Figure 2.4 for an illustration of how the included area looks for a street segment using three different radii.

The variables that are included for each street segment is shown in Table 2.2, their meaning and their grouping is explained in Chapter 3, Section 3.1. All variables are calculated by Stavroulaki et al. [30]. See a simplified example of the data in Table 2.3 which presents a few selected variables and also the full day pedestrian count, which is to be predicted, as the column TOTAL.

2.4.1 Attraction data

As presented in Table 2.2, attraction variables are included for each street segment in the data. These variables are counting the number of local market, public transport nodes and schools. They are counted both on the street segment and within a 500 meter radius, see Table 2.2.

To calculate these attraction variables, different attractions were collected into a

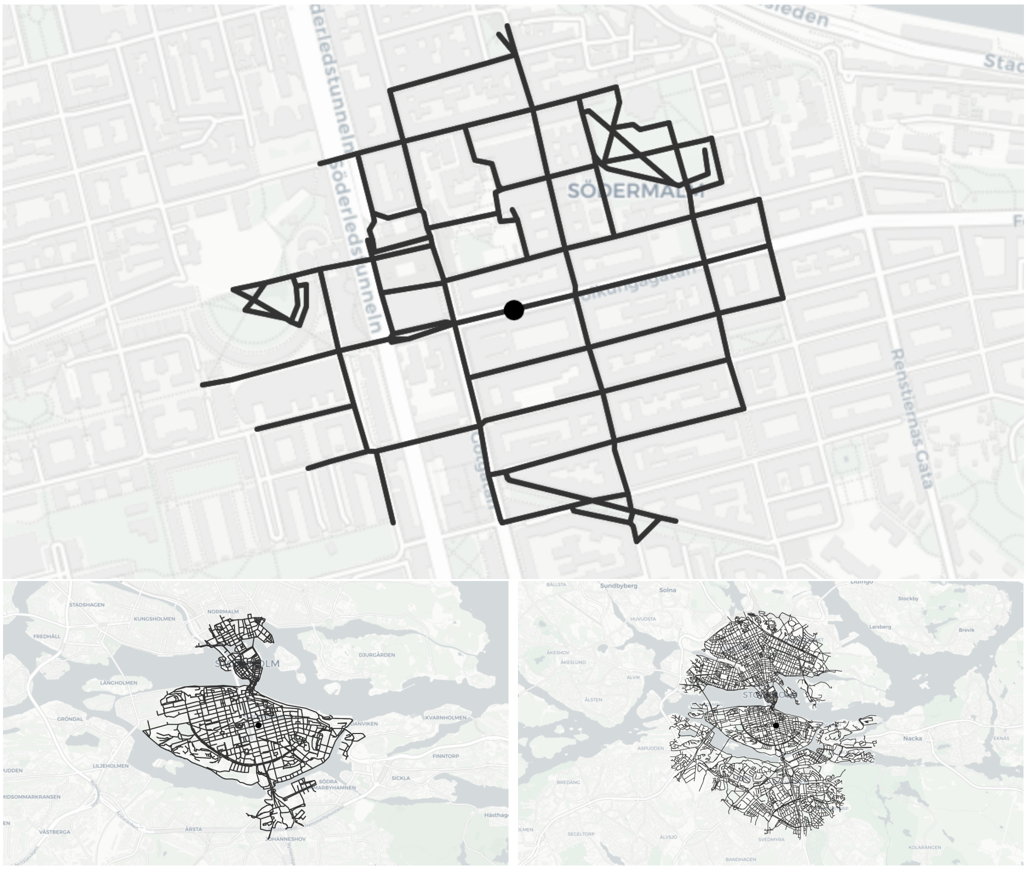

Figure 2.4: 500, 2500 and 5000 meter radius around a street segment in Stockholm.
The dot shows the center of the street segment in question and all the black lines are street segments that are included in the measurement for that specific radius.
Attribution: Leaflet[1] | © OpenStreetMap[23] © CartoDB[9]

Table 2.2: Variables calculated for each street segment in the data.

Name	Radii
Street centrality	
Angular integration	Range[500, 5 000, 500]
Angular betweenness	Range[500, 5 000, 500]
Built density	
Accessible FSI	500
Accessible GSI	500
Land division	
Accessible #plots	500
Attractions	
#Local markets	Street segment, 500
#Public transport nodes	Street segment, 500
#Schools	Street segment, 500

Range[*min*, *max*, *step*]: Represents a range of numbers from *min* to *max* with increments of *step* size.

Street segment: Variable is measured on the street segment itself.

Table 2.3: Example data from table with processed and aggregated counts per street segment.

START	END	TOTAL	Bet500	Int500	FSI_500	PubTr_500
114196	114197	5 316	612,50	1,06	1,65	13,00
114197	114200	1 677	786,00	1,14	1,65	13,00
114201	114202	6 010	1 961,50	1,29	1,72	11,00
114200	114203	835	831,00	1,19	1,68	16,50
114202	114205	4 898	2 084,25	1,39	1,66	14,00
114205	114206	103	629,63	0,96	1,56	16,50
114205	114208	315	601,60	1,02	1,59	16,00
114207	114208	338	684,83	1,15	1,65	12,67
114206	114207	393	1 088,00	1,18	1,69	15,67
114208	114209	498	1 764,33	1,22	1,59	15,33
114199	114200	93	121,00	1,10	1,77	18,00
114197	114198	1 967	196,00	1,13	1,66	17,00
114196	114201	10 069	2 046,50	1,20	1,59	13,00
114202	114203	58	493,17	1,18	1,72	14,00
114205	114209	4 258	2 509,00	1,43	1,52	18,50
114198	114199	217	1 074,00	1,13	1,65	18,00
114203	114204	70	271,00	1,11	1,75	17,00
114199	114204	120	919,75	1,17	1,70	18,50

Only a few select columns are included here to exemplify the table.

data set by Bobkova et al. [7] using OpenStreetMap (OSM)[1]. This data set is also used during this thesis.

Table 2.4: Descriptions of different OSM codes.

Code	Included	Description
10xx	-	Cities, towns, suburbs, villages,...
20xx	X	Public facilities such as government offices, post office, police, ...
21xx	X	Hospitals, pharmacies, ...
22xx	X	Culture, Leisure, ...
23xx	X	Restaurants, pubs, cafes, ...
24xx	X	Hotel, motels, and other places to stay the night
25xx	X	Supermarkets, bakeries, ...
26xx	X	Banks, ATMs ...
27xx	X	Tourist information, sights, museums, ...
29xx	-	Miscellaneous points of interest
41xx	-	Natural features
50xx	-	Parking lots, petrol (gas) stations, ...
52xx	-	Traffic related
56xx	X	Bus, tram, railway, taxi, ...
5601	X	A larger railway station of mainline rail services.

Included column marked with X means that attractions having that code was collected in Stavroulaki et al. [30].

Each of the attractions in this attraction data set is categorized using a code used in OSM. This OSM code is four digits and the first two digits of OSM code are related to general function, e.g., retail and service while the last two digits are related to subcategory of each class, e.g., bakery. See Table 2.4 for some descriptions. See also the Included column which, if marked with an X, means that it is Included in the attraction data set.

Note that the grouping of attraction into local markets, public transport nodes and schools does not follow the OSM codes, these groupings were created by Stavroulaki et al. [30]. Local markets include attractions with OSM codes starting with 23 and 25. Public transport includes attractions with OSM codes starting with 56 and schools include attractions with OSM code 2082.

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

3
Theory

This chapter describes the background, previous work, main variables and predictive model algorithms used in this thesis.

3.1 Variables

In order to understand the problem, solution and lessons learned, it is important to understand the variables that are used. There are conceptually four information categories for the main variables used in the models and these are: street centrality, built density, land division and attractions. What they mean and how they are or can be represented is explained below.

Street centrality is a measurement of how central a street is. In detail, it is a combination of two measurements, betweenness and integration.

Betweenness, is described in Stavroulaki et al. [31] at page 7 in the following way: “Network Betweenness calculates how often a line falls on the shortest path between all pairs of lines in a network, or how many shortest paths pass through it. In other words, lines (axial lines or segments) which control and mediate movement and connections between many other lines in the system have a high betweenness value.”. Lines can in this context be seen as a street segment. See Figure 3.1 for a visualization.

Integration, similar to mathematical closeness, as defined by Hillier et al. [14], is a measurement of centrality which looks at the distance to all other street segments. See Figure 3.2 for a visualization.

Both the betweenness and the integration are in this thesis measured using angular deviation instead of metric distance, based on the findings in Hillier and Iida [15]. Therefore, they are referred to as *Angular betweenness* and *Angular integration* respectively, as defined in Hillier et al. [16]. Whenever betweenness or integration is mentioned in this thesis, it refers to the angular version.

It is also important to note that these measurements can be calculated within different cut-off radii which means that they can, e.g., be calculated at a local as well as a global scale.

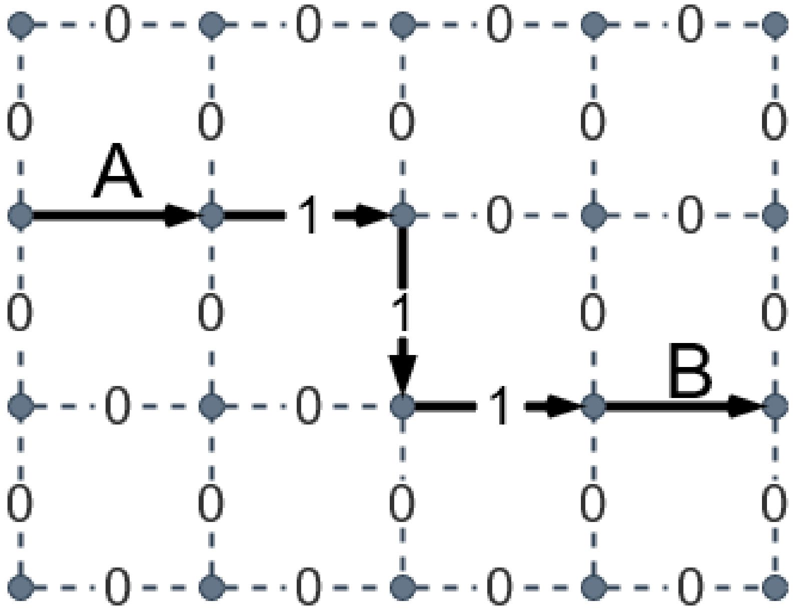

Figure 3.1: Betweenness measurement.
Given the shortest path between A and B, as visualized with black lines and arrows. Then the segments marked with a 1 are then the segments that fall on this shortest path. The more times a segment falls on one of these shortest paths, the higher the betweenness value.

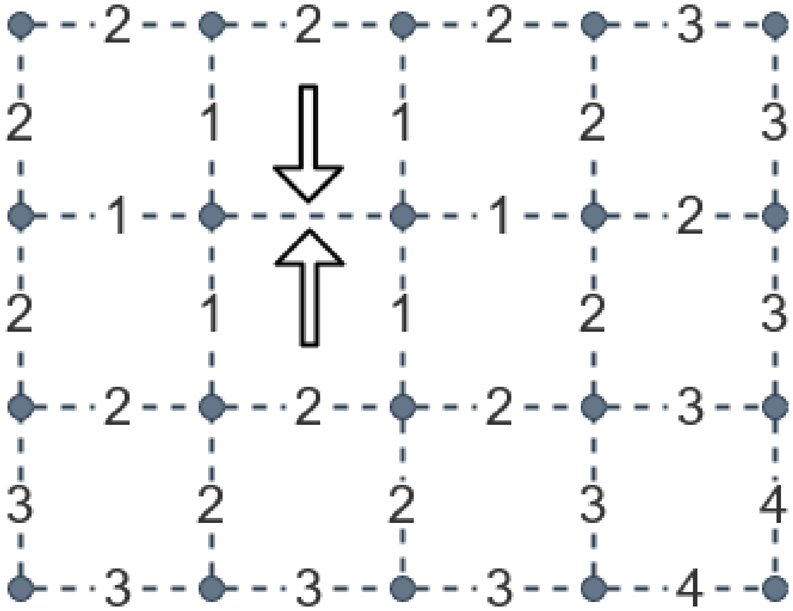

Figure 3.2: Integration measurement.
The street segment marked out with arrows on the top and bottom is the street segment being measured. Each other street segment is marked with the distance between them. The final integration value is an average of these distances.

Built density is a measurement of how densely or sparsely an area is built. One simplified way of looking at it is how big buildings are with respect to how much area they fill up. In detail, as described by Berghauser Pont and Haupt [4], it is a combination of the measurements; Floor Space Index (FSI), Ground Space Index (GSI) and Network density (N).

FSI, also referred to as intensity, is described by Berghauser Pont and Haupt [4] as a ratio between the gross floor area and the base land area, see Figure 3.3 for a visualization. Gross floor area is the total floor area for all the floors within a building. Base land area would for a district be the whole area of the district where the boundaries of the district are drawn in the middle of the streets surrounding the district.

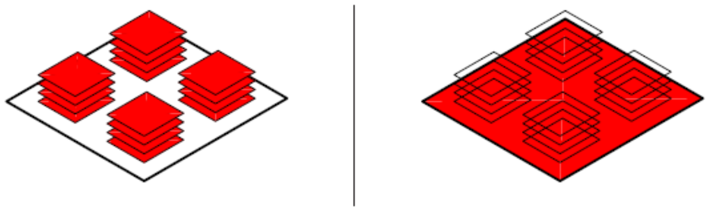

Figure 3.3: FSI measurement.
I.e., the ratio between the gross floor area to the left and the base land area to the right.
Figure source: Page 95 in Berghauser Pont and Haupt [4]
Permission: Meta Berghauser Pont

GSI, also referred to as coverage, is described by Berghauser Pont and Haupt [4] as a ratio between the footprint and the base land area, see Figure 3.4 for a visualization. The footprint, also referred to as built area, for a building is the area of land that it covers.

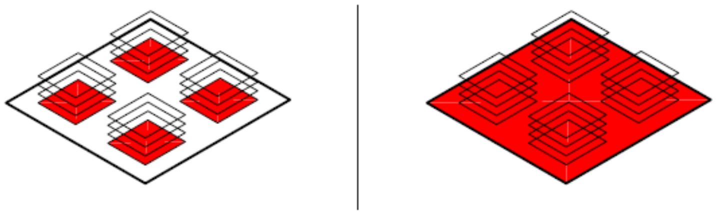

Figure 3.4: GSI measurement.
I.e., the ratio between the footprint to the left and the base land area to the right.
Figure source: Page 95 in Berghauser Pont and Haupt [4]
Permission: Meta Berghauser Pont

Network density is a measurement of network length in relation to the base land area, see Figure 3.5 for a visualization. Network length is simply the length of the network.

Berghauser Pont and Haupt [4] gives a few examples for what the network consists of at the district scale and those are circulation streets, rails, roads and canals.

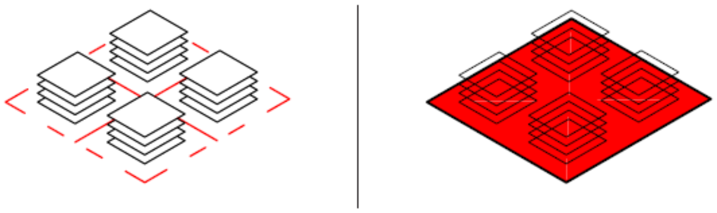

Figure 3.5: Network density measurement.
I.e., the ratio between the network length to the left and the base land area to the right.
Figure source: Page 94 in Berghauser Pont and Haupt [4]
Permission: Meta Berghauser Pont

Land division, also referred to as plot systems, looks at the boundaries between different plots. In detail, as mentioned in Bobkova et al. [6], land division measures accessible number of plots, accessible compactness and accessible openness. Accessible here refers to measurements calculated within a specific distance, e.g., within 500 meters walking distance.

Plots are divided by ownership, i.e., each property is a plot.

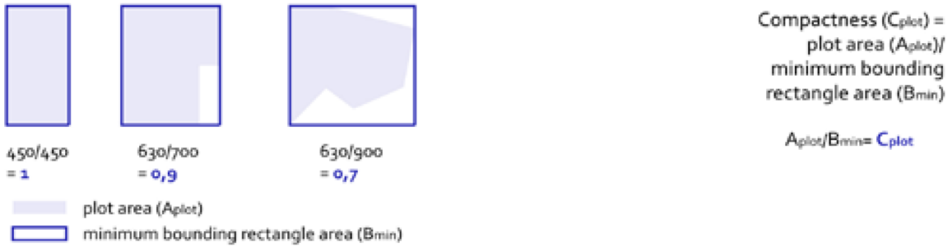
Compactness (C_{plot}) = plot area (A_{plot})/ minimum bounding rectangle area (B_{min})
$A_{plot}/B_{min}= C_{plot}$
450/450
= 1
630/700
= 0,9
630/900
= 0,7
plot area (A_{plot})
minimum bounding rectangle area (B_{min})

Figure 3.6: Compactness measurement.
I.e., the ratio between the plot area, the marked area, and the bounding rectangle area.
Figure source: Figure 5 in Bobkova et al. [6]
Permission: Evgeniya Bobkova

Compactness, is a measurement of how close a plot shape is to a rectangle, see Figure 3.6.

Openness is described in Bobkova et al. [6] as the ratio between the total plot frontage and the total plot perimeter, see Figure 3.7. An example of plot frontage is the length of the lawn for a house that merges with the street and not just another plot.

Attractions, sometimes referred to as activities, refer to non-residential land uses

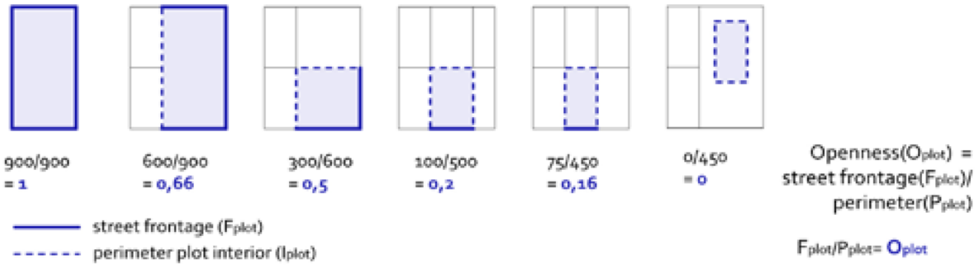
900/900
= 1
600/900
= 0,66
300/600
= 0,5
100/500
= 0,2
75/450
= 0,16
0/450
= 0
Openness(O$_{plot}$) = street frontage(F$_{plot}$)/ perimeter(P$_{plot}$)
street frontage (F$_{plot}$)
perimeter plot interior (I$_{plot}$)
F$_{plot}$/P$_{plot}$= O$_{plot}$

Figure 3.7: Openness measurement.
I.e., the ratio between the plot frontage, marked with a thick solid line, and the plot perimeter, both the thick solid and dashed line.
Figure source: Figure 5 in Bobkova et al. [6]
Permission: Evgeniya Bobkova

such as restaurants, hair salons, grocery shops, bars, bus stops and schools. Attractions can be represented as a count accessible within an area, the distance to specific attractions or possibly many other ways.

3.2 Theoretical background

Some of the variables used and their representations in this thesis are chosen by taking other researchers' contributions in this field into consideration. This section introduces some of the most important contributions.

There is a strong correlation between the pedestrian movement and the street configuration. This is shown in Hillier et al. [14] where they concluded that the more a street is connected with the rest of the city, the higher the pedestrian movement. Note that street configuration is what determines angular betweenness and angular integration for each of the street segments in a city. They also concluded that streets with high pedestrian movement attract attractions which in return attract more pedestrians, as later confirmed by Penn et al. [24] and Stavroulaki et al. [30]. This addition of attractions supposedly acts as a multiplier on the pedestrian movement, more specifically on the pedestrian movement estimated using the street configuration. This is one of the reasons why both street centrality and attractions are included in the pedestrian movement models.

The type of street determines the radius of integration measurement which will give the optimal predictability. Radius of integration measurement is here referring to the how big of a radius the integration measurement is calculated within, e.g., the integration measurement can be calculated within 500 meters or 5 000 meters. The different street types referred to here are categorized as primary, secondary and local street. A primary street is for example long and stretches throughout the city while local would be a short street that only stretches within a neighbourhood. That there is an optimality of radius for the integration measurement is shown for vehicular traffic in Penn et al. [24] and for pedestrians in Read [27] and Pont and Marcus [25]. The difference in optimal radius for the integration measurement is why some researchers, such as Berghauser Pont et al. [5]. calculate street centrality at multiple scales.

Pedestrian appreciation of "distance" is better predicted by the angle needed for navigation rather than the metric distance itself. This is shown in Hillier and Iida [15] which increases our understanding of the individual pedestrians' cognitive reasoning when it comes to choosing paths, it also confirmed by Dalton [10]. This is the reason why street centrality is calculated by the authors of Berghauser Pont et al. [5] using *Angular* integration and *Angular* betweenness. However, metric walking distance is still used to determine the radius for the area to include in the calculations.

Street centrality helps determine the potential character of a street, e.g., residential or commercial. This is shown in Özbil et al. [34] where they also state that land uses[1] is a more significant factor for determining pedestrian movement in an area, while street configuration is a more significant factor for pedestrian movement in individual street segments. Both of these are considered in our models, land use through attractions and street centrality through integration and betweenness.

Attractions on the ground floor and a diversity of attractions correlate positively with pedestrian movement. These were the two factors, found in Netto et al. [22], that had the strongest positive correlation with pedestrian movement. They found this when looking into how to explain the extra variation of pedestrian movement while the street centrality is the same. Therefore, it is probably a good idea to try to represent attractions in different ways in order to try to improve the predictiveness of pedestrian movement.

Street configuration explains the distribution of pedestrian movement but not the volume. This is shown in Özbil et al. [35] where they studied 20 2km x 2km areas in Istanbul, it is also confirmed by Berghauser Pont et al. [3]. Özbil et al. [35] also found that the attractions on the ground floor explained 35 percent of the pedestrian movement as well as that sidewalk width had the strongest correlation to pedestrian movement amongst other street design variables. In this experiment they categorized street segments into four different groups depending on the number of attractions available. The different groups were called: active/friendly, mixture, boring and inactive. This categorization of street segments might be one of the possible ways to represent attractions in order to improve predictiveness.

Built density of an area determines the volume of pedestrians and street centrality determines the distribution of the pedestrians. This is shown in Berghauser Pont et al. [3] were they group street segments based on street centrality and built density. They then compare the intensity and fluctuation of the pedestrian movement flow between these groups. The data used in Berghauser Pont et al. [3] is the same as the data used in this thesis.

[1] Different types of land use can for example be recreational, commercial or residential

Walkability has decreased when adding other means of transport. The reason for this is the imposed barrier on walking that each new means of transport brings. Cars have for example lead to the construction of high speed roads which can be tricky to cross or get past for pedestrians. This is explained in Forsyth and Southworth [12] where they also mention underground subways as the exception to this rule. They also explain walkability quite thoroughly but there is another conclusion that might be more interesting in the context of this thesis. If walkability decreases with other means of transport then having variable(s) that represent this might help explain pedestrian movement.

Route directness and completeness of pedestrian facilities affects pedestrian volumes. This is shown in Moudon et al. [21] where they compare pedestrian volumes in neighbourhoods with similar residential density. They are referring to directness as the ratio between the straight line path and the shortest path and to completeness as the ratio of dedicated pedestrian pathways. In this study, they used two categories of neighbourhoods, urban and sub-urban, between which both route directness and completeness differ. The urban neighbourhoods have both more direct routes and more complete sidewalk systems. This is then compared to the pedestrian volume where urban neighbourhoods are measured to have a three times higher count. Both of these variables can therefore be seen as potential variables in predicting pedestrian counts. However, route directness is presumably, at least to some extent, correlated with street centrality since they both measure the ease of traveling by foot.

In summary, street centrality, built density and attractions have been found to be helpful when predicting pedestrian movement. Street centrality seems to be better represented by the angular deviation and should preferably be measured for multiple radii. Attractions seem to have a strong impact on pedestrian movement where ground floor attractions and a diversity of attractions are supposed to be the most important. Built density seems to give a strong indication of the total number of pedestrians in an area.

3.3 Previous work

This section will explain the previous work in predictive models for pedestrian movement done by Stavroulaki et al. [30] and Håkansson [17].

Stavroulaki et al. [30] created three negative binomial models (negative binomial is explained further in Section 3.4.1) using the full day counts of pedestrians. The three different models were called configurational, spatial and attraction. See Table 3.1 for an overview of the models.

The configurational model takes street centrality into account, using Angular integration and Angular betweenness as explanatory variables. The spatial model included the same variables with the addition of accessible FSI (to represent built density) and accessible number of plots (to represent land division). The attraction model

Table 3.1: Overview of the models used in Stavroulaki et al. [30].

	Configurational	**Spatial**	**Attraction**
Street centrality			
Angular integration	X	X	X
Angular betweenness	X	X	X
Built density			
Accessible FSI	-	X	X
Accessible GSI	-	-	-
Land division			
Accessible #plots	-	X	X
Attractions			
#Local Markets on segment	-	-	X
#Local Markets within 500m	-	-	X
#Public Transport on segment	-	-	X
#Public Transport within 500m	-	-	X
#Schools on segment	-	-	X
#Schools within 500m	-	-	X
Control variables			
Weekday	X	X	X
City	X	X	X
Random effect			
Neighbourhood	X	X	X

included all the variables in the spatial model with the addition of attractions which were represented using the following variables:

- Accessible Local Markets within 500 m walking distance
- Number of Local Markets on segment
- Accessible Public Transport nodes within 500 m walking distance
- Number of Public Transport nodes on segment
- Accessible Schools within 500 m walking distance
- Number of Schools on the segment.

All three models included day of the week and the city as categorical variables.

Stavroulaki et al. [30] also compared these negative binomial models to logarithmic regression models using the same variables and concluded that the negative binomial models were preferable because they gave higher Continuous Rank Probability Scores (CRPS), which can be interpreted as giving more reliable results when probabilities are taken into account.

As an extension to the work done by Stavroulaki et al. [30], Håkansson [17] created a model focused on the pedestrian flow, more specifically the pedestrian count for each hour during the day from 6am to 9pm. Instead of using continuous variables to represent the street centrality and built density as in Stavroulaki et al. [30] the model uses categorical variables for street and density types, developed by Berghauser Pont et al. [3], see Table 3.2. This was to make the problem easier to model and the result easier to interpret. The model also includes variables to represent attractions, specifically public transport stops, schools and local markets.

Table 3.2: Density and street types used in Håkansson [17].

Built density		**Street centrality**	
Type	Description	Type	Description
1	Spacious low-rise	1	Background network
2	Compact low-rise	2	Neighbourhood streets
3	Dense mid-rise	3	City streets
4	Dense low-rise	4	Local streets
5	Compact mid-rise		
6	Spacious mid-rise		

Both the work in Stavroulaki et al. [30] and Håkansson [17] suggested that there is a correlation between the morphological structure, i.e., street centrality, built density and land division, and the number of pedestrians.

The pedestrian count models in Stavroulaki et al. [30] suggest that street centrality which was included in the Configurational model can explain a large part of the pedestrian count. Adding built density and land division, as done in the Spatial model, only had a small increase in model accuracy while the Attraction model had a higher increase. This suggests that the inclusion of attraction variables does make an important difference while the inclusion of land division and built density has a smaller impact. It is, however, uncertain if the attraction variables by themselves, not including built density and land division, would have the same effect. It is also interesting to note that even though the Attraction model increased the accuracy, only the variable for public transport stops on the same street showed a significant effect.

The Angular betweenness and Angular integration variables, which explain street centrality, included in Stavroulaki et al. [30] is calculated using specific distances, as explained in Chapter 2. The exact distance used in these models was chosen by performing Pearson correlations and the results gave Angular betweenness within 3 500 meters and Angular integration within 1 000 meters.

The pedestrian flow model in Håkansson [17] gives an indication of which morphological types in the data that generally correlates with a higher pedestrian count. The comparison between density types and street types was done individually, not in combination. The order for Density types, from higher to lower, was the following: Dense mid-rise, Compact mid-rise, Dense low-rise, Compact low-rise, Spacious low-rise. This indicates that the more densely built an area is, the more pedestrian movement there will be. The order for Street types was: City streets, Local streets, Neighbourhood streets, Background network. This indicates that the more central a street is, the more pedestrian movement there will be. There was also an indication of differences in the number of pedestrians between cities where, when all other variables were considered, Amsterdam was correlated with the highest counts followed by Stockholm and then London.

The results also indicates that markets within 500 meters are correlated with the highest pedestrian count amongst the attraction variables. That is different from the results in Stavroulaki et al. [30] where the public transport stops in the same street was the only variable with significant effect. A difference to note here is that the significance effect was calculated in Stavroulaki et al. [30], but not in Håkansson [17]. The conclusions here for Håkansson [17] are based on the values for the fixed effects within the model, i.e., the coefficients for each variable.

3.4 Predictive model algorithms

Two types of predictive model algorithms are used in this thesis, negative binomial and random forest. Negative binomial is used because of the findings in the previous work done in Stavroulaki et al. [30]. Random forest is mainly used because it is relatively stable, interpretable and easy to use. Using two different models can also help in giving more “robust” results. If both models indicate the same thing then those results are more trustworthy than if it was indicated by just one model.

3.4.1 Negative binomial

In order to understand the negative binomial distribution, it is first important to understand the Poisson distribution. The Poisson distribution models the count of independently occurring events within a specific time frame. Independently here means that the probability of an event happening is independent of if or when other events occur. The Poisson distribution only has one parameter and that is λ, the average number of events in one time frame, also referred to as the mean. The Poisson distribution assumes that the variance is the same as the average, this is where the negative binomial comes in. Negative binomial is a special version of a Poisson distribution where the difference is the addition of a dispersion parameter. That is, a parameter that controls the variance in the data separately from the average. See Figure 3.8 for a comparison between Poisson and Negative binomial. More information about the negative binomial is provided by Hilbe [13], the negative binomial referred to here is the one called NB2.

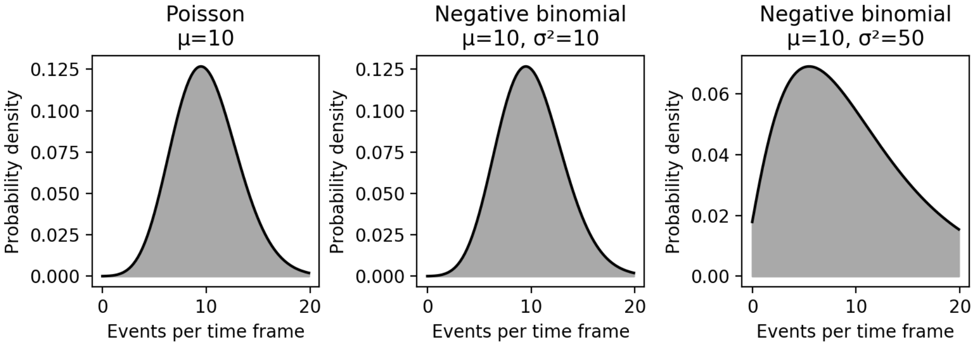

Figure 3.8: Comparison of the Poisson distribution with the Negative binomial distribution.
The Negative binomial distribution in the middle is the same as the Poisson distribution to the left since they both have their variance equal to the mean. The Negative binomial to the right is an example of over-dispersion, i.e., when the variance is higher than the mean.

The negative binomial regression model is a generalized linear model and can therefore be fitted using many different approaches, e.g., maximum likelihood estimation.

3.4.2 Random forest

Random forest is an ensemble of decision trees that makes use of bagging. An ensemble means that there is a collection of multiple models that all contribute to the prediction. A decision tree can be seen as a collection of nested if statements where an if statement either leads to a new if statement or a prediction. Bagging refers to an ensemble technique that creates multiple models where each one is trained on a subset of the training data, see Figure 3.9.

The prediction of the whole model is made by a majority vote or an average of all the sub-models, see Figure 3.10. There is also one extra technique used in random

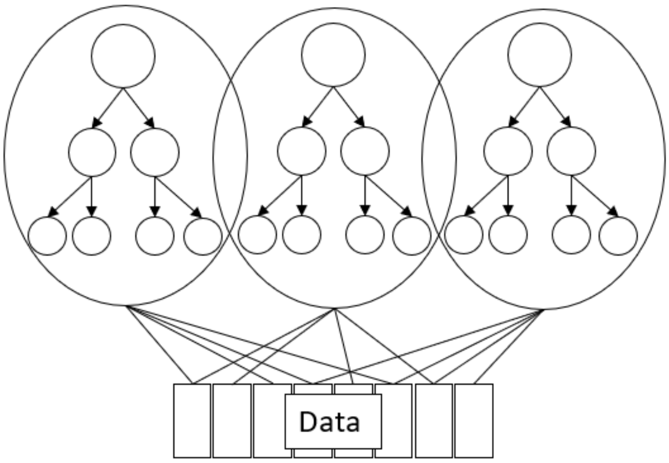

Figure 3.9: Training process of random forest.
Each tree is trained with a subset of data.

forest, which is that the split of each node is done by selecting the best split within a random subset of the variables, see Figure 3.11. This differs from normal decision trees in that the split at each node will for normal decision trees be selected by the best split amongst all variables. In this thesis we make use of the R implementation of random forest, as described in Liaw and Wiener [19].

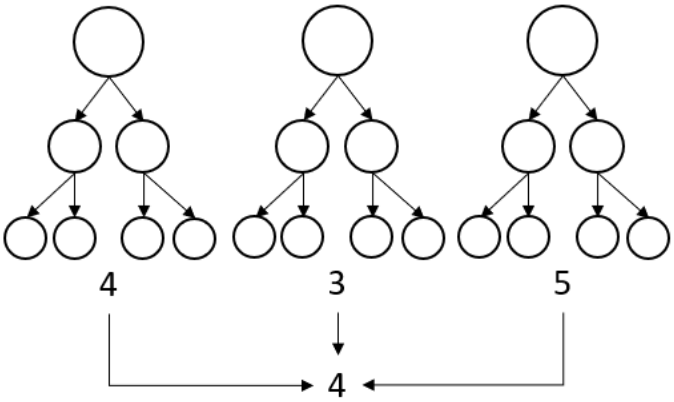

Figure 3.10: Prediction made by random forest.
The prediction is the average of the prediction made by each of the decision trees.

The benefit of random forest is that it is a relatively stable model, it is interpretable and it is simple to use. It is simple to use in the sense that it does not require transformation of any variables. There are two main parameters to tune in random forest: number of trees and number of variables for the random subset of variables for each node. In order to interpret the model, random forest makes use of something

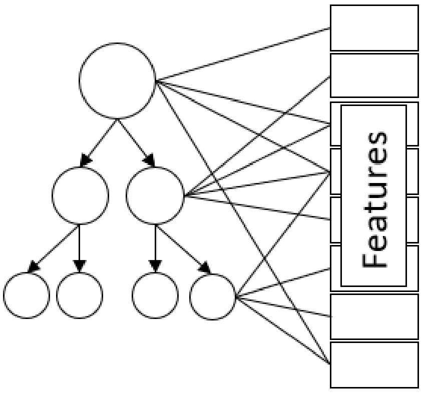

Figure 3.11: Picking of a random subset of features during random forest training.
A random subset of features is picked for each node before finding the best split. This is only visualized for the right side of the tree for simplicity.

called importance. Importance is a measurement for each variable in how important they are for making a prediction. I.e., a high importance for a variable means that it has a heavy weight in determining what the prediction is.

3.5 Metrics

There are four different metrics used for evaluating model performance in this thesis, all of them are introduced in this section.

Mean Absolute Error (MAE) sums up the difference between the predictions and the actual values. See Equation 3.1, where n is the number of samples, y_i is the true value for the i:th sample and $\hat{y}_i$ is the predicted value of the i:th sample.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.1}$$

Root Mean Squared Error (RMSE) is similar with the difference that it penalizes higher values more heavily by squaring the error. See Equation 3.2, where n is the number of samples, y_i is the true value for the i:th sample and $\hat{y}_i$ is the predicted value of the i:th sample.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.2}$$

Coefficient of determination (R^2) is a metric that indicates how well the model explains the variability in the data, simply put it is the mean squared error divided by the variance. See Equation 3.3, where n is the number of samples, y_i is the true value for the i:th sample, $\hat{y}_i$ is the predicted value of the i:th sample and $\bar{y}$ is the average value of all the samples.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} \tag{3.3}$$

Adjusted R^2 is similar to R^2 with the difference that it also takes the number of features used into account, so the more features included the lower the score. See Equation 3.4, where n is the number of samples and k is the number of features used.

$$R^2_{adj} = 1 - \left[\frac{\left(1 - R^2\right)(n-1)}{n-k-1}\right] \tag{3.4}$$

3.5.1 Metrics for statistical models

Statistical models such as the negative binomial do not give predictions in the same way as machine learning models. So, the metrics are for the statistical models calculated using the fitted mean, μ̂. Similar to how R$^2{}_{\mathrm{RES}}$ is calculated in Cameron and Windmeijer [8]. See Equation 3.5, where n is the number of samples, y_i is the true value for the i:th sample, $\hat{\mu}_i$ is the fitted mean for the i:th sample and $\bar{y}$ is the average value of all the samples.

$$R^2_{RES} = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{\mu}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} \tag{3.5}$$

3.6 Feature preprocessing

The previous work, Stavroulaki et al. [30] and Håkansson [17], scaled all numerical features except for attractions. The scaling was performed in R without centering. See Equation 3.6, as described by Becker [2], where n is the number of samples, X is the numerical vector of all the values for a feature and X_scaled is the vector of the scaled values for that feature.

$$\text{X\_scaled} = \frac{X}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n} x_i^2}} \tag{3.6}$$

4

Preparation for experiments

This chapter explain the preparation needed before performing any of the experiments in this thesis. For example, the training and test set split and algorithm parameters.

4.1 Data split

In order to evaluate the final model, mostly how well it generalizes, test data is picked out of the data set. There a four options for how to pick the test data:

- Pick one of the cities.
- Pick a few areas.
- Pick completely random street segments.
- Pick street segments such that there is a similar distribution of the type of street segments within the training and test set.

Pick one of the cities. This limits the relation between the street segments in the training data and the test data, meaning that a good test score would be a strong indication that the predictive model has found a general relationship between the built environment and the pedestrian movement. A drawback is that taking one city out of the data reduces the training data by one third.

Pick a few of the areas. This also limits the relationship between the street segments in the training data. However, it would not give as strong of an indication if the predictive model is generalized between cities, compared to picking one of the cities. Picking out a few areas might result in removing some specific “types” of streets which decreases the possibility for the predictive models to find the underlying relationship between the built environment and pedestrian movement counts.

Pick completely random street segments. This does not limit the relation between the street segments in the training data and the test data. There is also a chance that all street segments of a specific “type” could be picked into only the training data or the test data although more unlikely than when picking out a few areas. It does, however, give the opportunity to choose the exact amount of data that is picked for the test set.

Pick street segments such that there is a similar distribution of the type

of street segments within the training and test set. This choice does not limit the relationship between the street segments in the training and test set. It does however provide the opportunity to pick an exact amount of data for the test set. It also, of course, keeps a similar distribution between street segment “types” between the training data set and test data set.

For a compacted version of these comparisons see Table 4.1.

Table 4.1: Comparison between different data split methods.

Method	Test/Train relation	Test set size	Distribution of types
City	Very limited	One third	Fairly good
Area	Limited	Fairly dynamic	Probably bad
Random	Not limited	Dynamic	Possibly bad
Type	Not limited	Dynamic	Good

The data split is done using the last choice, keeping the distribution between the test set and training set, mostly because of the possible limitation of the data set that was discussed in Section 2.3.

Table 4.2: Number of street segments with full day count per street and density type combination.

		Density					
		1	**2**	**3**	**4**	**5**	**6**
Centrality	**1**	65	36	76	33	90	42
	2	12	18	25	28	18	23
	3	8	18	21	10	23	10
	4	23	36	48	10	15	19

To understand how this test data is picked out, it is important to understand the density and centrality type distribution of the data. There are six different density types and four different centrality types, this amounts to 24 combinations of those types. These are the centrality and density types used in Håkansson [17] and developed by Berghauser Pont et al. [3], as introduced in Section 3.3. Table 4.2 presents the number of street segments with full day counts for each of these combinations. The minimum number of street segments in these type combinations is 8, see Centrality 3 and Density 1, while the maximum is 90, see Centrality 1 and Density 5. It is also possible from this to calculate that the average number of street segments per combination is 29. This means that even though there is a wide spread of the street segments between the categories, the categories still differ noticeably in the amount.

The test data is picked by randomly choosing 10 percent of the street segments within each type combination. This means that the distribution stays roughly the same within the training and test data. The reason for having this relatively low percentage of the data for testing is because of the limitation of the sample size. Keeping a larger part of the data for training and validation can help avoid over-fitting to individual data points, it does however limit the reliability of the final test score.

4.2 Variable evaluation

Pearson correlation is calculated for all of the variables explored in this thesis in order to understand if there is a linear relationship with the pedestrian movement counts. Whenever correlations are presented or mentioned in the following chapters, they are calculated using Pearson correlation.

4.3 Model training

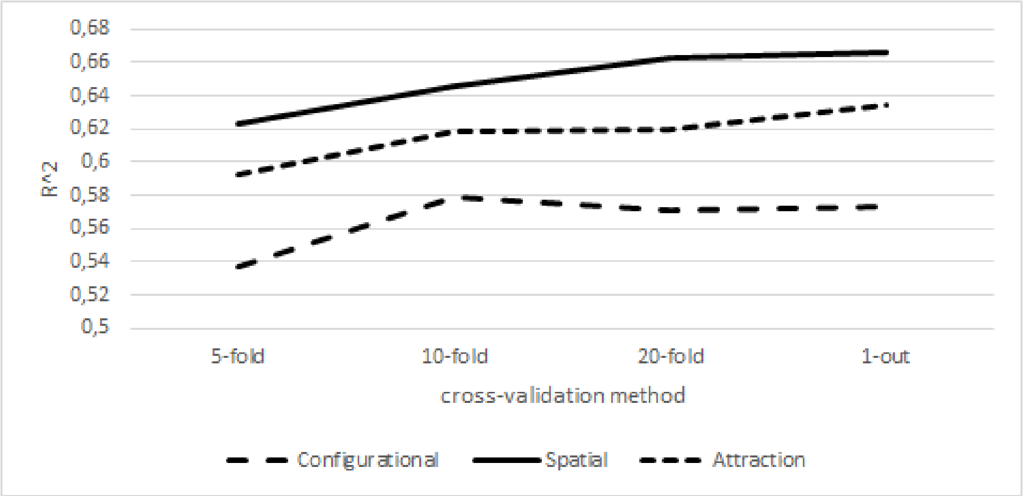

Figure 4.1: R^2 using different methods of cross-validation for the previous work models.

There are two different algorithms used in this thesis. One is negative binomial and the other is random forest. These algorithms train the models in different ways. Negative binomial is in this thesis trained using Integrated Nested Laplace Approximation (INLA), same as in previous work done by Stavroulaki et al. [30] and Håkansson [17]. More details about INLA is provided in Rue et al. [28]. Random forest is evaluated using leave-one-out cross-validation. Leave-one-out cross-validation is chosen above k-fold cross-validation because it performed slightly better when testing different methods on the previous work models created in Stavroulaki et al. [30]. See the results in Figure 4.1.

The implementations used for these algorithms are both libraries in R, R-INLA [26] for the negative binomial and Liaw [18] for random forest.

4.4 Model evaluation

In order to evaluate the predictive models against each other, four different metrics are used and they are MAE, RMSE, R^2 and Adjusted R^2. For negative binomial, these metrics are calculated using the fitted means, as explained in Section 3.5.1. These metrics are always presented in this thesis as averages of 10 runs using different fixed random seeds in order to get more stable result. The results from the negative binomial models are deterministic. However, random forest training is parallelized for speed up and therefore the metrics can differ slightly between runs. The baseline for evaluation are the models created in the previous work, the ones explained in Section 3.3.

The four metrics for evaluation are used because they have slightly different characteristics, the difference between them is explained in Section 3.5. It is important to note that none of these metrics are perfect at explaining how well a model performs. Therefore, when needed, visualizations such as residual plots are used for further analysis.

4.5 Algorithm parameters

The algorithm parameters used in this thesis are chosen following some tests, which are summarized in this sections. The same parameters are used throughout the thesis.

There are two main parameters to tune for random forest, those are the number of trees and the number of variables for the random subset of variables for each node, as mentioned in Section 3.4.2. The latter parameter is referred to as *mtry*.

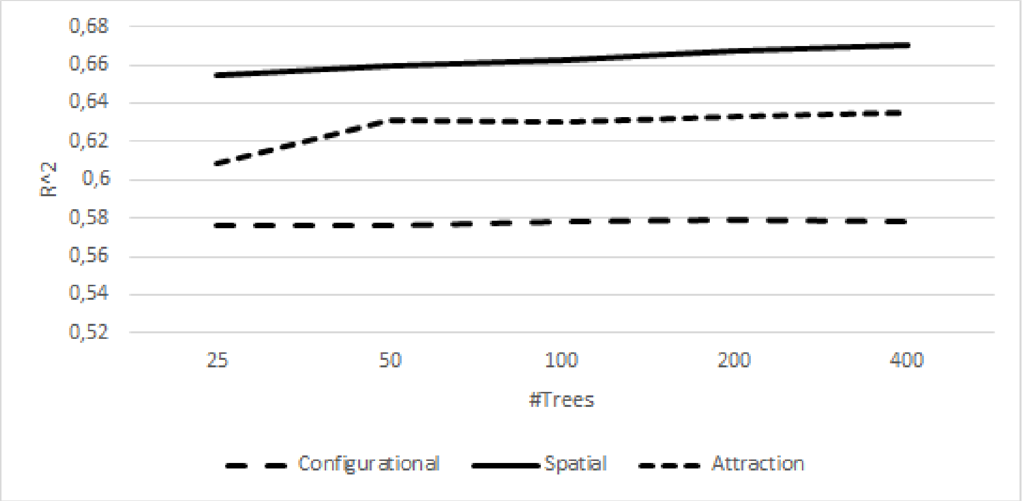

Figure 4.2: R^2 using different number of trees for the previous work models.

The number of trees is the amount of trees created during training and these are then used for prediction. This number of trees to use is chosen from running a test with 25, 50, 100, 200 and 400 trees on the previous work models created by Stavroulaki et al. [30]. The result from this test can be seen in Figure 4.2. In this figure, it is possible to see that there is a continuous, but slight, increase in performance. However, the largest increase, mostly for the Attraction model, is to around 50 trees. With this in mind, 200 trees are chosen for the experiments in this thesis so that there are more than enough trees. The most important is that the models do not suffer any major performance loss because of a lack of trees.

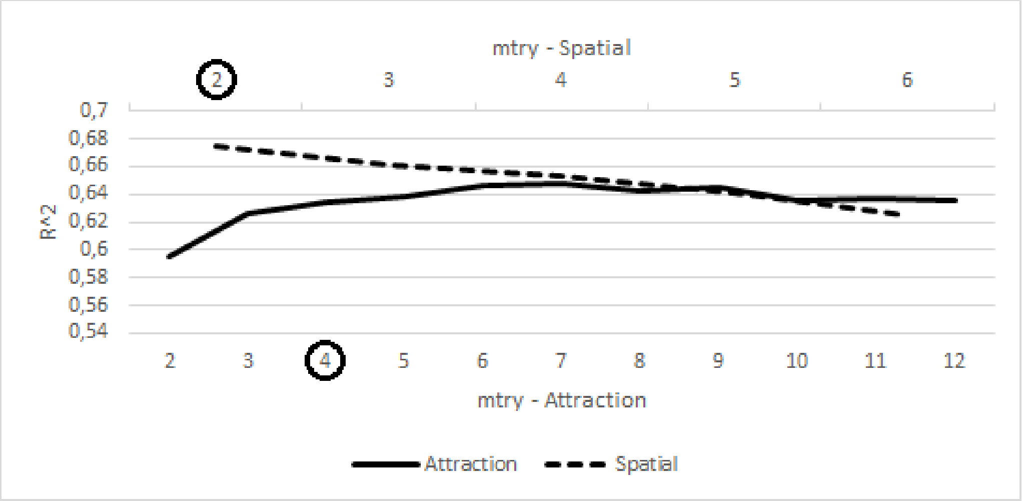

Figure 4.3: R^2 using different mtry for the Spatial and Attraction model.
The Spatial model's default value is the circled two.
The Attraction model's default value is the circled four.
*The Configurational model is left out because of low amount of parameters.

Mtry is explained as the number of variables randomly sampled as candidates at each split. The default value for mtry is the number of variables divided by three. Whether or not this default is good enough is determined in a test on the Spatial and Attraction model using different mtry values. The Configurational model is excluded from this test because it only includes 4 variables which makes it difficult to see any trend in performance between different mtry values. The result of the test is presented in Figure 4.3. For the Spatial model, the default mtry value is two and is marked with a circle on the top horizontal axis. For the Attraction model, the default mtry value is four and is marked with a circle on the bottom horizontal axis. From this figure it seems like none of the models are performing considerably better at any other mtry values compared to the default. Therefore the default value for mtry is used in all the experiments in this thesis. It is also worth noting that higher mtry values increase training time.

There are no optimization parameters to mention for negative binomial.

5

Reproducing previous work

The models created in the previous work, Stavroulaki et al. [30], are used as a starting point for evaluation of the models created during this thesis. Therefore, the first experiment is to train and calculate the metrics for those models. The variables used in each model are specified in Section 3.3 in Table 3.1.

5.1 Variable correlation

Table 5.1: Correlation with pedestrian movement counts for variables used in previous predictive models.

Street centrality		Built density & Land division		Attractions 500m		Attractions Street	
Integration	0.276	FSI	0.411	PT 500m	0.341	PT street	0.127
Betweenness	0.372	GSI	0.384	LM 500m	0.479	LM street	0.390
		Plots	-0.034	Sch 500m	-0.051	Sch street	0.001

PT: Public transport nodes, LM: Local markets, Sch: Schools
FSI: Floor Space Index, GSI: Ground Space Index

The correlation with pedestrian movement counts is calculated for each of the variables used in Stavroulaki et al. [30]. This is done to get insight into the explanatory value of each variable. The correlations are presented in Table 5.1.

From these correlations, it seems like the number of accessible plots is not very informative in terms of pedestrian movement counts. It also seems like attractions within 500 meters are more explanatory than attractions on the same street segment. Moreover, schools do not seem to have a strong correlation for either of the distances.

5.2 Result

The result for the negative binomial models using the previous work variables is presented in Table 5.2. These R^2 scores are surprisingly low considering the result in Stavroulaki et al. [30]. The scores differ for two reasons. The first reason is that the metrics here are calculated using only 90 percent of the data, i.e., the training data, as explained in Section 4.1, whereas 100 percent of the training data was used in

Table 5.2: Metrics for the Negative binomial models following the previous work.

Model	MAE	RMSE	R^2	Adjusted R^2
Configurational	756	1 731	0.664	0.661
Spatial	857	2 685	0.192	0.183
Attraction	797	2 243	0.436	0.424

Stavroulaki et al. [30]. The second reason is that two street segments with high FSI values were excluded in Stavroulaki et al. [30] but they are included here.

When comparing the MAE scores between the models, it seems like the models perform similarly. However, when we look at the RMSE scores, then the Spatial and Attraction model scores much higher than the Configurational model. This means that a few of the prediction errors in the Spatial and Attraction model are relatively large.

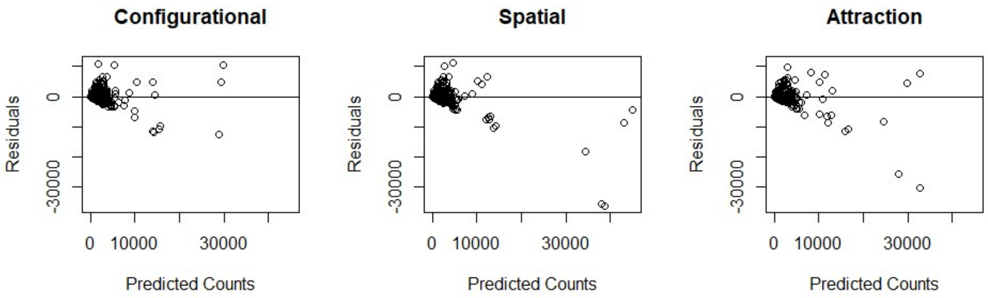

Figure 5.1: Residual plot for the previous work models using negative binomial.

The residual plots for these models using the negative binomial are presented in Figure 5.1. To interpret these residual plots, it is important to know that values above the x-axis are underestimated and values below the x-axis are overestimated. When inspecting the Spatial and Attraction residuals, it is possible to see that a few of the residuals are very highly overestimated.

Table 5.3: Metrics for the Random forest models following the previous work.

Model	MAE	RMSE	R^2	Adjusted R^2
Configurational	962	1 935	0.580	0.577
Spatial	765	1 731	0.664	0.661
Attraction	745	1 808	0.633	0.626

The result for the random forest models using the previous work variables is presented in Table 5.3. The R^2 score improves $0.084 (= 0.664 - 0.580)$ with the Spatial model in comparison with the Configurational model. However, the score decreases when comparing the Attraction model to the Spatial model. This indicates that the additional variables in the Attraction model provides less information gain than it increases the complexity of finding a good fit.

When comparing the results in Table 5.2 and Table 5.3 it seems like the random forest models are generally performing better than the negative binomial models, in terms of these metrics. However, the most interesting in these results are how surprisingly low the Spatial model with negative binomial scores, less than one third of the Configurational model.

5.3 Exploring variable transformation

According to the result in Section 5.2 the Spatial model was scoring very low. This is mostly because of an overestimation of two street segments having high FSI values. One way of dealing with this could be to transform the FSI variable as to decrease the influence of these extreme values.

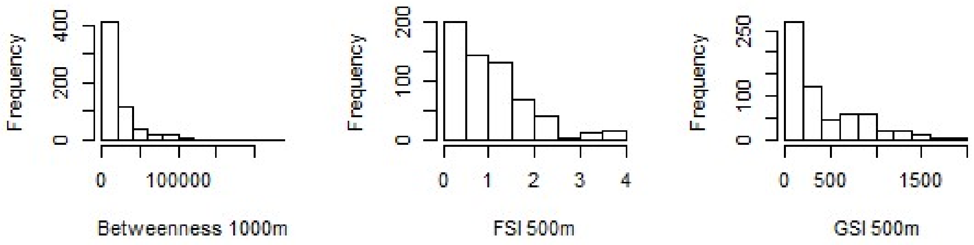

Figure 5.2: Histogram of skewed variables.

Looking at the distribution of betweenness, FSI and GSI, see Figure 5.2, it is possible to see that these variables are asymmetrically distributed, i.e., skewed. They are not presented in plots here but Angular integration and accessible number of plots are fairly symmetrically distributed.

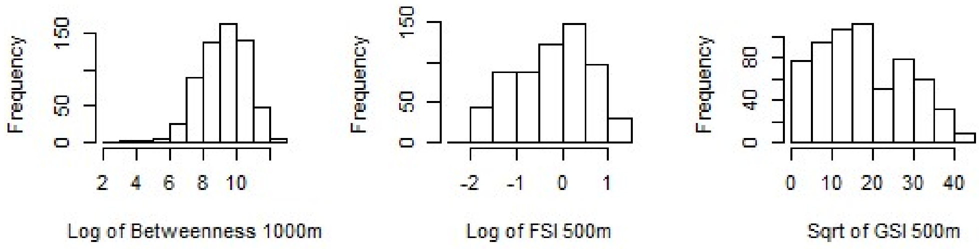

Figure 5.3: Histogram of skewed variables from Figure 5.2 after transformation.

Using different transformation, the logarithm and the square root, then betweenness, FSI and GSI can be represented in ways that are more symmetrical, see Figure 5.3. This could allow the negative binomial models to be more arithmetically stable.

Table 5.4: Correlation with pedestrian movement counts before and after variable transformation.

	Betweenness	FSI	GSI
Before Transformation	0.372	0.411	0.384
After Transformation	0.275	0.375	0.338
Difference:	0.097	0.036	0.046

When comparing the correlation of the variables before and after transformation, see Table 5.4, it is possible to see that the correlations are lower after the transformation. This indicates that there is some information loss in terms of linear relationship during the transformation of these variables. However, it can still be useful to transform these variables in order to decrease the influence of extreme values.

5.3.1 Result using variable transformation

Table 5.5: Metrics for the negative binomial models following the previous work with transformed betweenness and FSI.

Model	MAE	RMSE	R^2	Adjusted R^2
Configurational^T	726	1 796	0.638	0.635
Spatial^T	730	1 793	0.640	0.636
Attraction^T	735	1 845	0.618	0.610

^{T}: Angular betweenness and FSI are transformed.

The results for the negative binomial models using transformed versions of Angular betweenness and FSI are presented in Table 5.5. All the metrics are here fairly similar between the models. The Spatial and Attraction models are, with these variables transformed, scoring much higher in R^2 than before the transformations, as presented in Table 5.2.

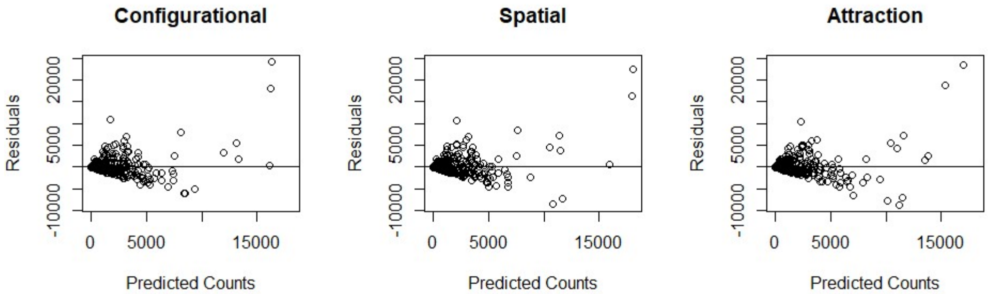

Figure 5.4: Residual plot for the negative binomial following the previous work with transformed betweenness and FSI.

The residual plots are presented in Figure 5.4. In these plots, there are two highly underestimated counts, in the top right corners, for all of the models that stick out from the rest of the residuals. It is, however, worth questioning how much these counts actually represent the average pedestrian count for these street segments. Since the measurements were performed for only one day in each area, it could for example be that there was a specific event happening on that day that increases the number of pedestrians.

5.4 Analyzing high counts

According to the results in Section 5.3.1 all the previous work models were underestimating two high counts. This section therefore investigates these high counts to see if there is something that explains why these counts are so much higher than the rest. It is important to understand why these street segments have much higher counts than the other street segment because the metrics used, i.e., MAE, RMSE, R^2 and Adjusted R^2, are affected more by high counts than low ones.

Table 5.6: The top ten highest pedestrian counts.

ID	City	Pedestrian count
348	Stockholm	40 537
349	Stockholm	34 149
329	Stockholm	18 709
340	Stockholm	16 418
318	Stockholm	15 963
317	Stockholm	15 214
332	Stockholm	15 044
389	London	12 714
755	Stockholm	10 069
444	London	10 005

The two top street segments are approximately double the amount of the third highest count. These two street segments are connected and part of Drottninggatan (a major shopping street) in Stockholm.

The two very high counts are in a specific area in Stockholm, Norrmalm, see Table 5.6. Therefore, it might have been an event during the measurement that caused the count to be very high. Otherwise it might be that the area is very active and that it is not uncommon with very high number of pedestrians.

5.4.1 Events in Stockholm

There were at least two events happening in the centre of Stockholm during the day, 2017-10-04, of the Norrmalm measurement. One of the events was the “Nobel Calling” which lasted for a week, from 2017-10-02 to 2017-10-08. However, this event was not happening close (Hotel Rival, Mariatorget 3) to Norrmalm during that day. The other event during that day, or actually in the evening, was “Gunnar Wennerberg 200 år” which included live music in memory of Gunnar Wennerberg. This was happening relatively close (Konstakademien, Fredsgatan 12) to Norrmalm but it is doubtful that it would have serious impact on the number of pedestrians in the measured area.

It was difficult to find out if any of the stores on these street segments had any major sale during this day which could also be a factor in why so many people visited the street.

5.4.2 Attractions and spatial layout

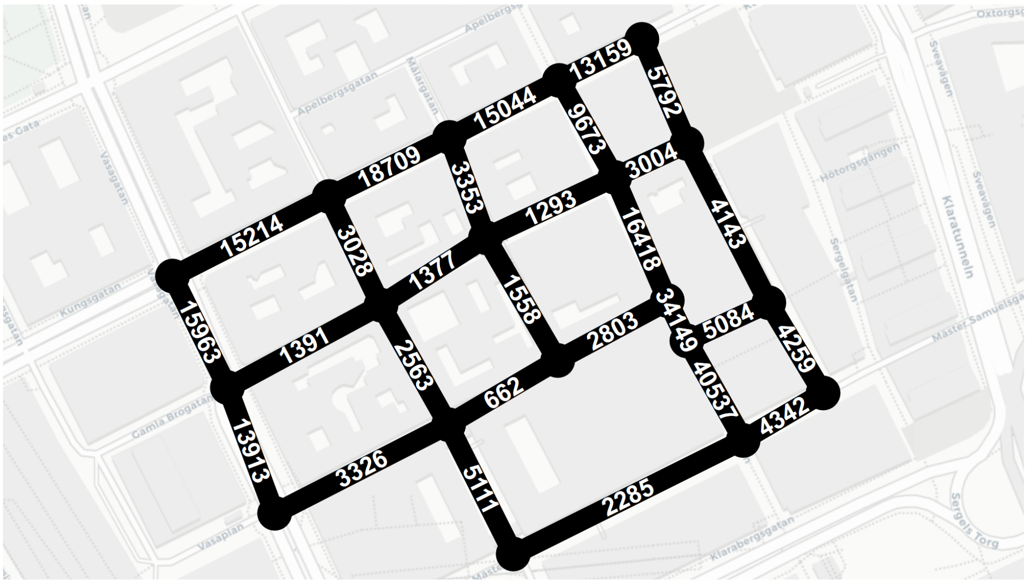

Figure 5.5: Overview of Norrmalm in Stockholm with the exact pedestrian counts presented.
The most notable counts are 40 537 and 34 149, down to the right.
Attribution: Leaflet[1] | © OpenStreetMap[23] © CartoDB[9]

The street segments with high counts can be seen in Figure 5.5. The street segments in question are in the south east going north west, they are in the beginning of the second street from the right. From the figures it is possible to see that the pedestrian count quickly goes down in the street segments above or to the side of the ones in question. Looking at the pedestrian counts from the south east and up, the progression is 40 537, 34 149 and then 16 418. The drop from 34 149 to 16 418 is almost half of the pedestrians that have turned around or walked in another direction.

Along the whole street which the street segments are part of, there are stores on the ground floor level. The difference with these two street segments, within the measured area, is that the stores are larger and probably also enjoy a larger number of customers, even though the number of them are roughly the same.

It is also worth considering that Norrmalm has a big shopping area, the middle of that area is just south east of these high count street segments. So to sum everything up, this area is very central and dense with attractions.

When this is considered, it seems plausible that these two street segments attract many pedestrians that are there to visit one or more stores and then walk back in the other direction.

5.4.3 Result excluding high counts

Table 5.7: Metrics for the previous work models using negative binomial with two high counts excluded.

Model	MAE	RMSE	R^2	Adjusted R^2
Configurational^T	664	1 346	0.612	0.609
Spatial^T	673	1 417	0.570	0.565
Attraction^T	671	1 406	0.576	0.567

^{T}: Angular betweenness and FSI are transformed.

The results for training negative binomial when excluding the high counts is presented in Table 5.7. These scores seem to be worse than when the two high counts were included in the data, as presented in Table 5.5.

Table 5.8: Metrics for the previous work models using random forest with two high counts excluded.

Model	MAE	RMSE	R^2	Adjusted R^2
Configurational	876	1 578	0.467	0.463
Spatial	679	1 357	0.605	0.601
Attraction	662	1 333	0.619	0.612

The results for training random forest when excluding these counts is presented in Table 5.8. These scores are lower than when the two high counts were included, same as with the negative binomial.

Training the previous work models when excluding the two high counts seem to decrease the scores for the metrics used. This suggests that there is something in the models capturing the fact that these street segments have high pedestrian counts.

Comparing Table 5.7 to Table 5.5, it is possible to see that the Configurational model is not affected by the removal of these points as much as the Spatial and Attraction model. This most probably means that one of the variables included in both of these models is the one that mostly explains the high pedestrian counts. There are two candidate variables, FSI and the number of accessible plots. Amongst these two, FSI has a much higher correlation to pedestrian movement counts, as can be seen in Table 5.1. This suggests that FSI is the best candidate.

6

Exploring different representations of street centrality

An extension of predictive models for pedestrian count is here explored by comparing different representations for street centrality. The reason for this is two-fold. Firstly, previous research, such as Hillier et al. [14], indicates that street centrality has a high influence on pedestrian movement. Secondly, these variables have been calculated within multiple radii by Stavroulaki et al. [30] and is readily available.

6.1 Method

This experiment is performed by creating a few variables which are referred to as centrality polynomials and then by designing new models based on different centrality variables. The creation of the centrality polynomials is explained in Section 6.1.1. The design of the models is provided in Section 6.1.3 together with a short description for each model. The results of the new models are presented in Section 6.2 and they are there also evaluated in comparison to the Spatial model, which is the highest scoring of the original models created in Stavroulaki et al. [30].

6.1.1 Variable: Centrality polynomials

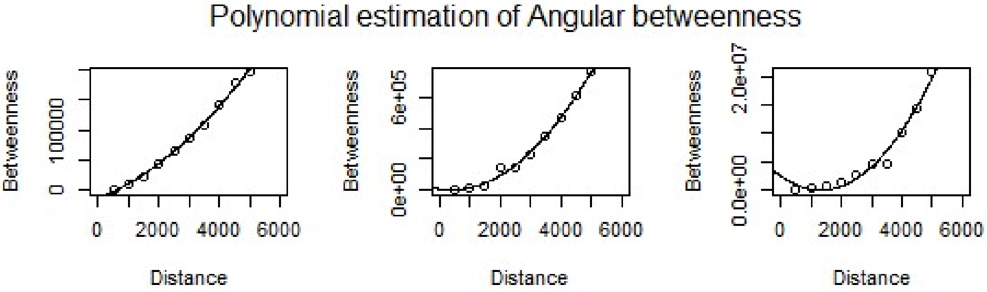

Figure 6.1: Three examples for polynomial estimation of Angular betweenness.

$$ax^2 + bx + c \tag{6.1}$$

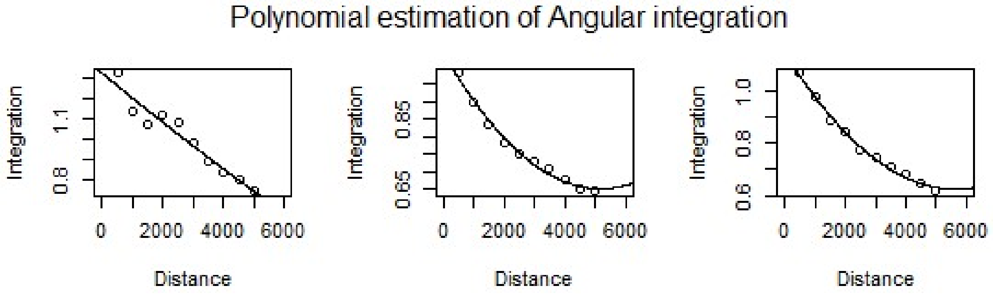

Figure 6.2: Three examples for polynomial estimation of Angular integration.

Angular integration and angular betweenness is calculated at ten different distances by Stavroulaki et al. [30]. This means that there are 20 variables to include in order to keep the calculations for each distance. However, these calculations can be estimated using a polynomial representation, see Figure 6.1 and 6.2. They are then represented using the coefficient for each term, i.e., a, b and c in Equation 6.1. This leads to both angular integration and angular betweenness being represented using only 6 variables.

6.1.2 Variable correlations

In order to explore the explanatory value of the available centrality variables regarding pedestrian movement counts, correlation is calculated for each one of them. These correlations are presented in Table 6.1. In Stavroulaki et al. [30] Angular betweenness at 3 500 meters and Angular integration at 1 000 meters were picked because they correlated highest with the pedestrian counts. In Table 6.1 Angular betweenness at 3 500 meters correlates the highest but Angular integration at 1 000 meters does not. There are two possible explanations for this. The first and probably most likely explanation is because the correlation in Stavroulaki et al. [30] were calculated per city which is not the case here. The second explanation could be that only the training data is used to calculate the correlations here, i.e., only 90 percent of the data.

The centrality polynomials for Angular integration has a low correlation with pedestrian movement counts, except for the constant term. The centrality polynomials for Angular betweenness, however, all correlate higher than Angular betweenness at 500 meters. The highest correlated term in these is the 2nd term in Equation 6.1. This is the term that determines the rate of change in Angular betweenness when increasing the radii of the measurement.

6.1.3 Models

Three new models are explored here, see Table 6.2. The new models are All, Three and Polynomials. The model referred to as All simply includes all measurements for angular integration and angular betweenness. The model referred to as Three includes

Table 6.1: Correlation with pedestrian movement counts for centrality variables.

Integration		Betweenness		Polynomials	
Int5000m	0.151	Bet5000m	0.253	Int Constant	0.245
Int4500m	0.179	Bet4500m	0.269	Int 1st term	-0.030
Int4000m	0.202	Bet4000m	0.277	Int 2nd term	-0.057
Int3500m	0.127	Bet3500m	0.372	Bet Constant	0.151
Int3000m	0.236	Bet3000m	0.279	Bet 1st term	-0.179
Int2500m	0.266	Bet2500m	0.275	Bet 2nd term	0.224
Int2000m	0.284	Bet2000m	0.260		
Int1500m	0.289	Bet1500m	0.205		
Int1000m	0.276	Bet1000m	0.160		
Int500m	0.253	Bet500m	0.108		

three measurements each for angular integration and angular betweenness, within 500 meters, 2 500 meters and 5 000 meters. The model referred to as Polynomials represent angular betweenness and angular integration by using polynomial estimations, in the way described in Section 6.1.1.

6.2 Result

The result of the negative binomial for the centrality models is presented in Table 6.3. These results indicate that the Three model is the best of the models evaluated, it scores best in all the metrics, it is also the only new model scoring higher than the Spatial model.

The result of the random forest for the centrality models is presented in Table 6.4. These results show that both the Three model and the Polynomials model scores higher than the Spatial model in all metrics. The Three model, however, scores considerably better than the Polynomials model.

The results presented in Tables 6.4 and 6.3 both indicate that the Three is the best model among those tested. This could be because it informs the model about how central a street segment is in regards to different radii. An example is that a small street in a neighborhood could be a very central part for navigating within and out of the neighborhood while it is not a central part for navigating within the city as a whole. It is then probably more informative to know both of these facts instead of only one of them to understand the pedestrian movement count of a street like this.

The same information as in the Three model is of course also included in the All model. However, the All model is evidently scoring lower than the Three model. This indicates that the information gain with including all radii is less than the increase of complexity in finding a relationship to the pedestrian movement count when training on this data. If given a larger set of data, there is a possibility that the All model could perform better because it is then possible for the model to learn

Table 6.2: Overview of the centrality models.

	Spatial	All	Three	Polynomial
Street centrality				
Angular integration				
5 000m	-	X	X	-
4 500m	-	X	-	-
4 000m	-	X	-	-
3 500m	-	X	-	-
3 000m	-	X	-	-
2 500m	-	X	X	-
2 000m	-	X	-	-
1 500m	-	X	-	-
1 000m	X	X	-	-
500m	-	X	X	-
Angular betweenness				
5 000m	-	X	X	-
4 500m	-	X	-	-
4 000m	-	X	-	-
3 500m	X	X	-	-
3 000m	-	X	-	-
2 500m	-	X	X	-
2 000m	-	X	-	-
1 500m	-	X	-	-
1 000m	-	X	-	-
500m	-	X	X	-
Integration polynomial				
Constant	-	-	-	X
1st term	-	-	-	X
2nd term	-	-	-	X
Betweenness polynomial				
Constant	-	-	-	X
1st term	-	-	-	X
2nd term	-	-	-	X
...				

... represents the other variables used that are not shown because they are the same as the Spatial model in Table 3.1.

Table 6.3: Results for negative binomial centrality models

Model	MAE	RMSE	R^2	Adjusted R^2
Spatial^T	730	1 793	0.640	0.636
All^T	732	1 878	0.605	0.588
Three^T	724	1 747	0.658	0.652
Polynomials^T	807	2 103	0.504	0.495

^{T}: Angular betweenness and FSI are transformed.

Table 6.4: Results for random forest centrality models

Model	MAE	RMSE	R^2	Adjusted R^2
Spatial	761	1 722	0.668	0.664
All	748	1 738	0.661	0.648
Three	687	1 571	0.723	0.719
Polynomials	721	1 651	0.694	0.689

more complex relationships.

7

Exploring different representations of attractions

An extension of predictive models for pedestrian count is here explored by comparing different representations for attractions. The reason for this is three-fold. Firstly, previous research, such as Netto et al. [22], indicates that attractions have a high influence on pedestrian movement. Secondly, the highest correlated variable to the pedestrian movement counts is an attraction variable, see local markets within 500 meters in Table 5.1. Thirdly, more attraction data than was used in the previous predictive models has already been collected by Bobkova et al. [7] from OSM and is available for use.

7.1 Method

This experiment is performed by first creating two different types of variables which are referred to as OSM Counts and OSM Distance, these are explained in Section 7.1.1. Then, new models are designed based on different attraction variables. The design of the models is provided in Section 7.1.3 together with a short description of each model. The results of the new models are presented in Section 7.2 and they are also evaluated in comparison to the Spatial model, which is the highest scoring of the original models created in Stavroulaki et al. [30].

7.1.1 Variable: OSM Attractions

Previous to this thesis, six attraction features had been calculated. These six features are made up of three categories with two different “distances” for each. The three categories are; Local Market, Public Transport and School. Each of them are represented with a count from the actual street segment in question and a count within a 500 meter walking radius. Note that these are the variables used in the previous work, as mentioned in Section 3.3.

Considering the background knowledge from Section 3.2, it is clear that attractions are important when predicting pedestrian movement. Therefore more granular data for attractions are created in this thesis. This is done using the attraction data collected by Bobkova et al. [7] as mentioned in Section 2.4.1. The distances to the surrounding attractions from each street segment are then calculated using GeoPandas[1] for storing geographical data, OSMnx[2] for downloading OSM road networks and NetworkX[3] for calculating the shortest path.

Two new types of variables are then created from these calculations. One of the types is the count of attractions within 500 meters for each OSM group. The other type is the distance to the nearest attraction for each OSM group. A simplified walk-through of the steps involved is provided below:

- Download the road network for walking from OSM using OSMnx with a 1 000 meter radius around every area.
- For each street segment in an area, calculate the distance to all the attraction within this 1 000 meter radius using NetworkX.
- Calculate the number of attractions for each OSM group within 500 meters for each street segment, using GeoPandas.
- Find the distance to the nearest attraction for each OSM Group for each street segment, using GeoPandas. (Some attractions are further away than 1 000 meters, therefore extra calculations for longer distances are performed when necessary.)

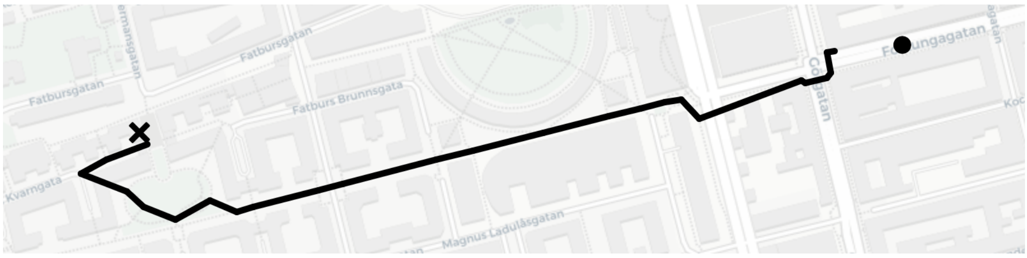

Figure 7.1: Shortest path to an attraction from a street segment calculated from the nearest node of the center.
Attribution: Leaflet[1] | © OpenStreetMap[23] © CartoDB[9]

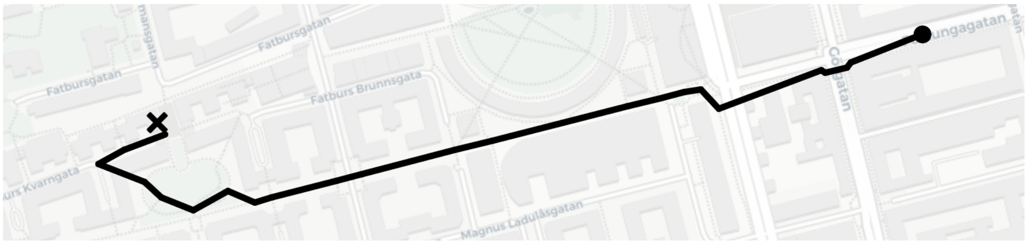

Figure 7.2: Shortest path to an attraction from a street segment calculated from the center of the street segment.
Attribution: Leaflet[1] | © OpenStreetMap[23] © CartoDB[9]

The distances to attractions is calculated in two different ways. One way is to calculate the distance from the nearest node from the center of the street segment,

[1] http://geopandas.org/

[2] https://osmnx.readthedocs.io/

[3] https://networkx.github.io/

see Figure 7.1. The other way is calculate the distance from the center of the street segment by adding it as a node in the road network, see Figure 7.2. In both ways of calculating, the distance is calculated to the node that is closest to the target attraction.

Only the result for calculating from the nearest node, as in Figure 7.1, is included here, the results for calculating from the center of the street segment, as in Figure 7.2, is instead presented in Appendix A.1. The variables calculated from the nearest node are chosen because, on average, they scored higher in R^2 values and had higher correlations with the pedestrian movement counts.

There is one important note regarding the differences in the road networks used for the distance calculation to attractions. The calculations here are performed on the OSM walking road network which is different from the road network used by Stavroulaki et al. [30] for calculating the variables presented in Section 3.1, i.e., the variables that were calculated previous to this thesis. The main difference between the networks is that the non-motorized does not include lanes while the OSM walk network does.

7.1.2 Variable correlation

Table 7.1: Correlation between attraction variables and pedestrian movement counts.

Previous attractions		OSM Counts		OSM Distance	
PT 500m	0.341	OSM-20	0.171	OSM-20	-0.199
LM 500m	0.479	OSM-21	0.413	OSM-21	-0.227
Sch 500m	-0.051	OSM-22	0.383	OSM-22	-0.251
PT street	0.127	OSM-23	0.427	OSM-23	-0.180
LM street	0.390	OSM-24	0.409	OSM-24	-0.238
Sch street	0.001	OSM-25	0.495	OSM-25	-0.190
		OSM-26	0.281	OSM-26	-0.256
		OSM-27	0.258	OSM-27	-0.210
		OSM-56	0.343	OSM-56	-0.133
		OSM-5601	0.400	OSM-5601	-0.141

PT: Public transport nodes, LM: Local markets, Sch: Schools
The description of each OSM code is found in Table 2.4.

The correlation between all the attraction variables and the pedestrian movement counts is presented in Table 7.1. These correlations indicate that the OSM Count variables have higher explanatory value than the OSM Distance variables. The variable with highest correlation is OSM-25 when counted within 500 meters. This is not too surprising when considering that the OSM-25 group contains supermarkets, see Table 2.4 for details on OSM-25, which is a frequent destination for many people.

Table 7.2: Overview of the attraction models.

	Spatial	Attraction 500m	OSM Group	OSM Railway	OSM Distance	OSM Railway Top	OSM Distance Top	OSM All Top
Attractions								
Local Markets								
Segment	-	-	-	-	-	-	-	-
500m	-	C	-	-	-	-	-	-
Public Transport								
Segment	-	-	-	-	-	-	-	-
500m	-	C	-	-	-	-	-	-
Schools								
Segment	-	-	-	-	-	-	-	-
500m	-	C	-	-	-	-	-	-
OSM codes								
20 (Public facilities)	-	-	C	C	D	-	-	-
21 (Hospitals,...)	-	-	C	C	D	C	D	C+D
22 (Culture,...)	-	-	C	C	D	-	D	D
23 (Restaurants,...)	-	-	C	C	D	C	-	C
24 (Hotel,...)	-	-	C	C	D	C	D	C+D
25 (Supermarkets,...)	-	-	C	C	D	C	-	C
26 (Banks,...)	-	-	C	C	D	-	D	D
27 (Tourism)	-	-	C	C	D	-	D	D
56 (Public transport)	-	-	C	C	D	-	-	-
5601 (Railway station)	-	-	-	C	D	C	-	C
...								

C: count of attractions.

D: walking distance to the nearest attraction in meters.

... represents the other variables used that are not shown because they are the same as the Spatial model in Table 3.1.

The full description of each OSM code is found in Table 2.4.

7.1.3 Models

Seven new models are explored using the available attraction variables, see Table 7.2. The Attraction 500m model only includes the attraction variables at 500 meters calculated previous to this thesis, these attraction variables are also calculated at street segment level which is included in the original Attraction model created by Stavroulaki et al. [30]. The OSM Group model includes all the unique OSM groups, i.e., the two digit codes. The OSM Railway model is an extension of OSM Group where OSM-5601 is also included. OSM-5601 is the code for railway stations. The OSM Distance model includes OSM Distance variables. The OSM Railway Top model includes the OSM Count variables that has a correlation above 0.4. The OSM Distance Top includes the OSM Distance variables that has a correlation below −0.2. The OSM All Top model includes all of the variables included in both OSM Railway Top and OSM Distance Top.

7.2 Result

Table 7.3: Results for negative binomial attraction models

Model	MAE	RMSE	R^2	Adjusted R^2
Spatial^T	730	1 793	0.640	0.636
Attraction 500m^T	752	1 854	0.615	0.609
OSM Groups^T	742	1 855	0.614	0.604
OSM Railway^T	745	1 867	0.609	0.598
OSM Distance^T	714	1 763	0.652	0.642
OSM Railway Top^T	736	1 829	0.625	0.617
OSM Distance Top^T	717	1 768	0.650	0.643
OSM All Top^T	726	1 820	0.629	0.618

^{T}: Angular betweenness and FSI are transformed.

The result of the negative binomial for the attraction models is presented in 7.3. Only the models using OSM Distance variables scored better than the Spatial model.

Table 7.4: Results for random forest attraction models

Model	MAE	RMSE	R^2	Adjusted R^2
Spatial	762	1 720	0.668	0.665
Attraction 500m	725	1 719	0.669	0.664
OSM Groups	725	1 782	0.644	0.635
OSM Railway	715	1 766	0.650	0.641
OSM Distance	776	1 843	0.619	0.609
OSM Railway Top	708	1 679	0.684	0.678
OSM Distance Top	760	1 828	0.625	0.619
OSM All Top	729	1 751	0.656	0.647

The result of the random forest for the attraction models is presented in 7.4. Both the Attraction 500m model and the OSM Railway Top model scored higher than the

Spatial model. However, only the OSM Railway Top model is worth considering in this result since the increase in R^2 is so low for the Attraction 500m model.

Results in Tables 7.4 and 7.3 indicate that the negative binomial and random forest react quite differently to the new models. Negative binomial scores the best when using distance while random forest scores the best when using counts. One of the reasons for this could be the zero inflation in the count variables, i.e. that the variables contain a high percentage of zero values. In the case of zero inflated variables, random forest is not affected in the same way because it is trained using many binary decisions, i.e. yes or no questions. This means that the model can treat the zero values completely differently from how it treats the non-zero values. The same is not true for the negative binomial since it is a type of linear regression model. Instead, it tries to find a linear relationship between the variables and predictor. This means that the relationship found between the variable and the predictor can be highly influenced by zero values if they are many. This is true for the OSM Count variables, the average of zero values is 41.3% which is a large part the values.

8

Exploring new variables based on road network

An extension of predictive models for pedestrian count adds variables created based on the road network. The reason for this is based on previous research. The first insight considered from previous research is that route directness and completeness of pedestrian facilities was found to affect pedestrian volumes, as stated by Moudon et al. [21]. Route directness and completeness are not the exact variables that are created in this experiments, however, the variables here explain the connectivity of the network to some extent which is similar to these variables. The second insight considered from previous research is that walkability decreases when other means of transport is added, as stated by Forsyth and Southworth [12].

8.1 Method

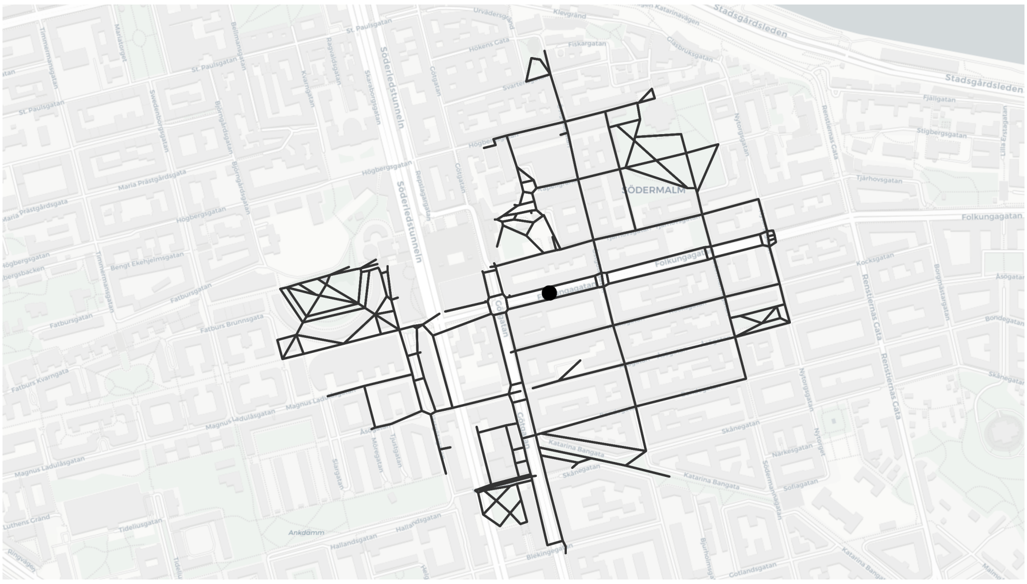

Figure 8.1: OSM walk network around a street segment with 500m reach.
The center of the street segment is marked with a circle.
Attribution: Leaflet[1] | © OpenStreetMap[23] © CartoDB[9]

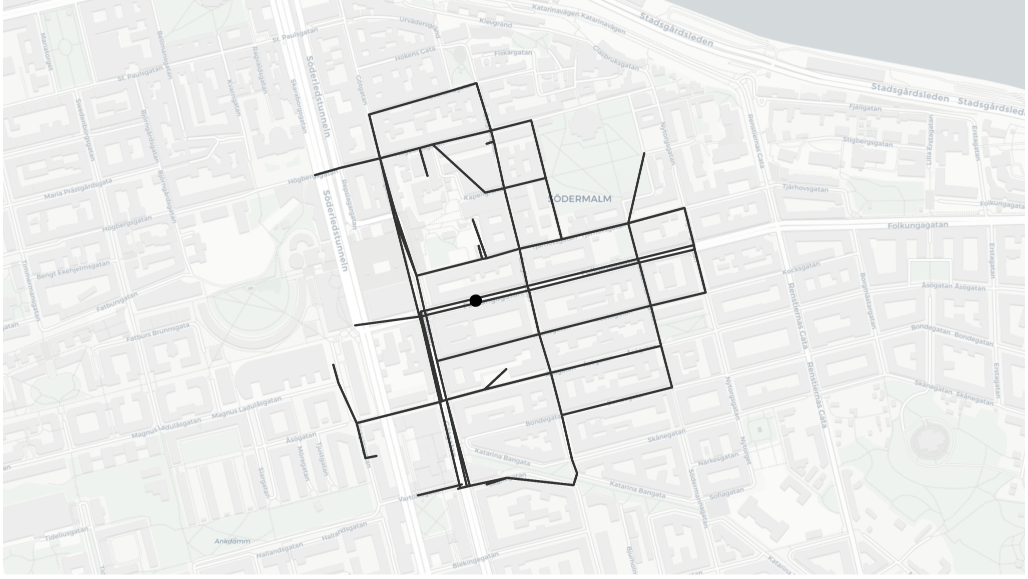

Figure 8.2: OSM bike network around a street segment with 500m reach.
The center of the street segment is marked with a circle.
Attribution: Leaflet[1] | © OpenStreetMap[23] © CartoDB[9]

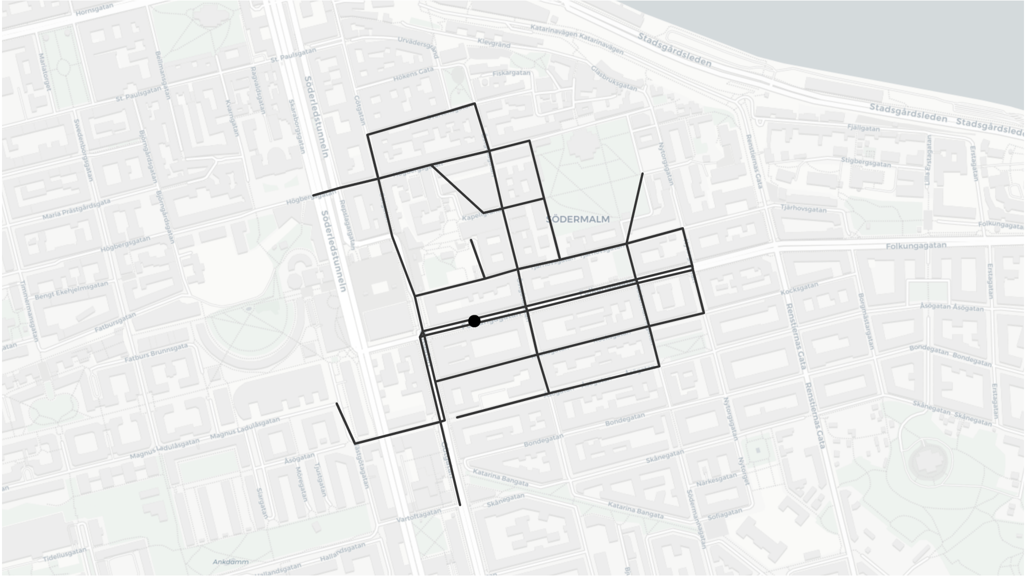

Figure 8.3: OSM drive network around a street segment with 500m reach.
The center of the street segment is marked with a circle.
Attribution: Leaflet[1] | © OpenStreetMap[23] © CartoDB[9]

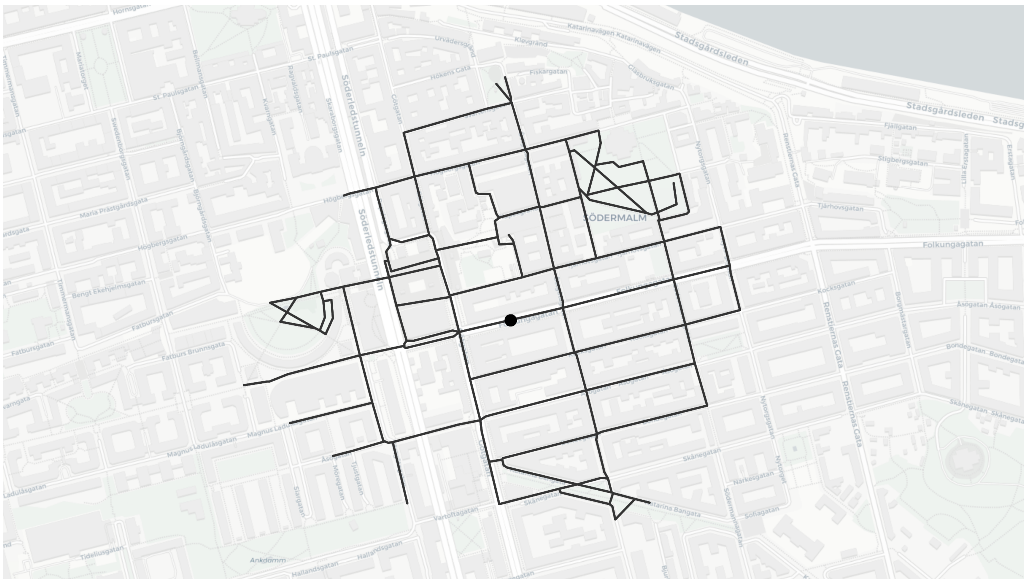

Figure 8.4: Non-motorized network around a street segment with 500m reach.
The center of the street segment is marked with a circle.
Attribution: Leaflet[1] | © OpenStreetMap[23] © CartoDB[9]

Four different road networks are used. Three are OSM road networks with different modes of transport, walk, bike and drive, see Figure 8.1, 8.2 and 8.3. The last one is a non-motorized network that is provided by Berghauser Pont et al. [3], see Figure 8.4. The non-motorized network includes all streets and paths that are accessible for people walking or cycling, including those that are shared with vehicles. From the aforementioned four figures, it is possible to see that the OSM road networks are more detailed than the non-motorized network. The difference is that the OSM road networks include "lanes" while the non-motorized network does not.

Using the OSM road networks, it is possible to compare the road networks for different modes of transport, i.e., walking, biking and driving. This could be useful because of the findings in Forsyth and Southworth [12], that suggested that an addition of other means of transport decreases walkability.

The road networks used are constricted using two different distance measures for a distance of 500 meters. One is network reach which restricts the network to the street segments reachable within 500 meters distance of travel. The other is called bounding box which includes the road network inside a square with each side being the 500 meters from the center. Figure 8.5 shows the walk network using bounding box for 500 meters, this can be compared to Figure 8.1 which instead uses network reach for the same network.

8.1.1 Variable: Network length

A few variables are created by summing up the length of the road network. This could give an indication of how well connected the network is. When the total

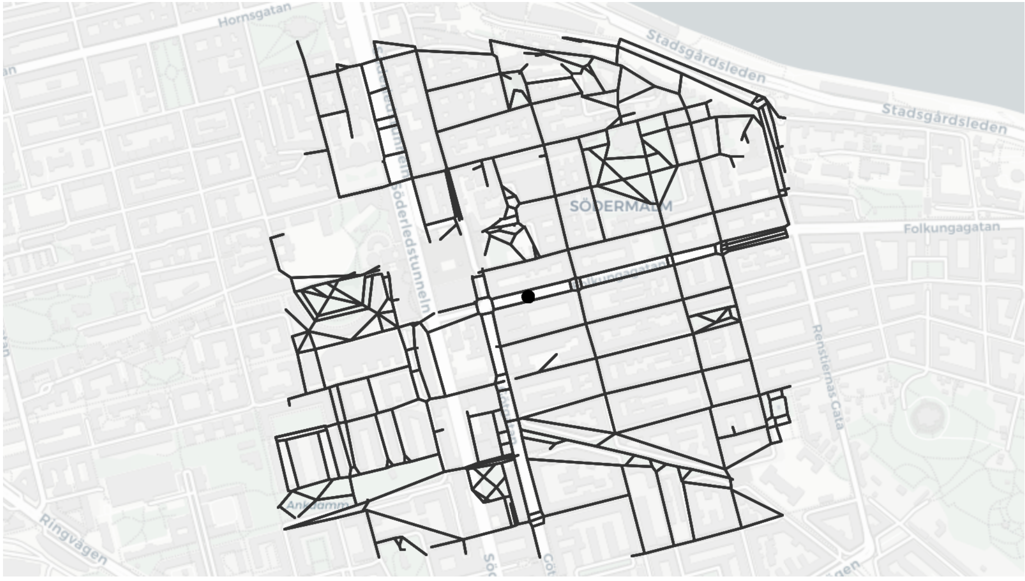

Figure 8.5: OSM walk network around a street segment with 500m bounding box.
The center of the street segment is marked with a circle.
Attribution: Leaflet[1] | © OpenStreetMap[23] © CartoDB[9]

network length is larger, it is likely that there are more connected roads. These variables are created using the OSM road networks, i.e., walk, bike and drive. They are created using both network reach and bounding box.

There are also two ratio variables created, one for each of the distance measures. The variables are calculated as the ratio between the walk network and drive network. This is done with the assumption that the longer the network length is for walking in comparison to driving, the better the walkability.

8.1.2 Variable: Intersection density

Some variables are created by calculating the number of intersections in the road network. The reason is partly the same as for the network length variables, that it gives an indication on how well connected the road network is.

Four intersection variables are created for each road network, the road networks are here limited by using network reach. The first three variables are created by calculating three way, four way and more way intersections. The last variable is the total count of intersections, i.e., three way + four way + more way.

An extra variable is created by taking the ratio between the intersections in the walking network and the driving network, similar to the ratio variables for the network length.

Table 8.1: Correlation between road network variables and pedestrian movement counts.

NMZ Intersection		OSM Intersection		OSM Network Length	
Three Way	0.108	Walk Three Way	0.314	Walk Reach	0.335
Four Way	0.207	Walk Four Way	0.295	Bike Reach	-0.046
More Way	0.076	Walk More Way	0.264	Drive Reach	-0.017
Intersections	0.150	Walk Intersections	0.327	Walk/Drive Reach	0.127
		Bike Three Way	0.098	Walk Box	0.288
		Bike Four Way	0.099	Bike Box	-0.006
		Bike More Way	0.162	Drive Box	0.076
		Bike Intersection	0.118	Walk/Drive Box	0.115
		Drive Three Way	0.045		
		Drive Four Way	0.101		
		Drive More Way	0.158		
		Drive Intersections	0.074		
		Walk/Drive Intersections	0.168		

8.1.3 Variable correlation

The correlations with pedestrian count for all of the road network variables are presented in Table 8.1.

The walk network (Walk Three Way, Walk Four Way, Walk More Way and Walk Intersections) seems to correlate the highest with pedestrian counts which is not too surprising. It is worth noting here, however, that the walking network has a noticeably higher correlation than the non-motorized network (Three Way, Four Way, More Way, Intersections). This indicates that the extra details included in the OSM road networks are valuable when predicting pedestrian counts.

The bike and drive network length variables, to the right in Table 8.1, have low correlations with pedestrian movement counts and they are mostly negative. The negative correlation here supports the claim made by Forsyth and Southworth [12], that the addition of other means of transport decreases walkability and therefore also the pedestrian counts. However, the correlations here are too small to draw any conclusions. Moreover, the bike and drive intersection variables, in the middle of Table 8.1, all have a positive correlation.

The network length variables calculated using road networks constrained by network reach have higher correlation than those constrained by bounding box. This could be because of the simple fact that the network reach constrains the road network in the same way as people actually travel. Therefore, then also being more representative of pedestrian counts.

8.1.4 Models

The new models created in this experiment are presented in Table 8.2. Six of the models are created using the intersection variables (NMZ Intersection Detailed, NMZ Intersection, Walk Intersection Detailed, Walk Intersection, All OSM Intersection

Table 8.2: Overview of the road network models.

	Spatial	NMZ Intersection Detailed	NMZ Intersection	Walk Intersection Detailed	Walk Intersection	All OSM Intersection	Intersection Ratio	All OSM Reach	All OSM Box	Walk Reach	Walk Box	Ratio Reach	Ratio Box
Road network													
Intersections													
Walk Three Way	-	-	-	X	-	-	-	-	-	-	-	-	-
Walk Four Way	-	-	-	X	-	-	-	-	-	-	-	-	-
Walk More Way	-	-	-	X	-	-	-	-	-	-	-	-	-
Walk Intersections	-	-	-	-	X	X	-	-	-	-	-	-	-
Bike Intersections	-	-	-	-	-	X	-	-	-	-	-	-	-
Drive Intersections	-	-	-	-	-	X	-	-	-	-	-	-	-
Walk/Drive Intersections	-	-	-	-	-	-	X	-	-	-	-	-	-
NMZ Three Way	-	X	-	-	-	-	-	-	-	-	-	-	-
NMZ Four Way	-	X	-	-	-	-	-	-	-	-	-	-	-
NMZ More Way	-	X	-	-	-	-	-	-	-	-	-	-	-
NMZ Intersections	-	-	X	-	-	-	-	-	-	-	-	-	-
Network length													
Walk Box	-	-	-	-	-	-	-	-	X	-	X	-	-
Bike Box	-	-	-	-	-	-	-	-	X	-	-	-	-
Drive Box	-	-	-	-	-	-	-	-	X	-	-	-	-
Walk/Drive Box	-	-	-	-	-	-	-	-	-	-	-	-	X
Walk Reach	-	-	-	-	-	-	-	X	-	X	-	-	-
Bike Reach	-	-	-	-	-	-	-	X	-	-	-	-	-
Drive Reach	-	-	-	-	-	-	-	X	-	-	-	-	-
Walk/Drive Reach	-	-	-	-	-	-	-	-	-	-	-	X	-
...													

... represents the other variables used that are not shown because they are the same as the Spatial model in Table 3.1.

and Intersection Ratio). The six other models are created using the network length variables (All OSM Reach, All OSM Box, Walk Reach, Walk Box, Ratio Reach and Ratio Box).

The NMZ Intersection models include the variables from the non-motorized network. The NMZ Intersection Detailed model includes the intersection count variables for each intersection type and the NMZ Intersection model includes the total intersection count.

The Walk Intersection models (both Walk Intersection Detailed and Walk Intersection) include the variables created from the walk network. The Walk Intersection Detailed model includes the intersection count variables for each intersection type and the Walk Intersection model includes the total intersection count.

The All OSM Intersection model includes the intersection variable for the walk, bike and drive network.

The Intersection Ratio model includes the intersection ratio variable, i.e., the ratio between walk and drive intersections.

The All OSM models (both All OSM Reach and All OSM Box) include the network length variables for all the OSM road networks, i.e., walk, bike and drive. The All OSM Reach model includes the variables created using the network reach constraint and the All OSM Box model includes the variables created using the bounding box constraint.

The Walk models (both Walk Reach and Walk Box) include the network length variables created using the walk networks. The Walk Reach model include the variables created using the network reach constraint and the Walk Box model includes the variables created using the bounding box constraint.

The Ratio models (both Ratio Reach and Ratio Box) include the network length ratio variables, i.e., the ratio between the walk and drive network length. The Ratio Reach model includes the ratio variable created using the network reach constraint and the Ratio Box model includes the ratio variable created using the bounding box constraint.

8.2 Result

The result of the negative binomial for the road network models is presented in Table 8.3. There does not seem to be any models that score noticeably better than the Spatial model. The ones that do score better than the Spatial model (Intersection Ratio, Walk Reach, Ratio Reach and Ratio Box) all score less than one percent higher in R^2. One model that attracts attention is the OSM All Reach model which scores considerably lower than all the other models.

The result of the random forest for the centrality models is presented in Table 8.4.

Table 8.3: Results for negative binomial road network models

Model	MAE	RMSE	R^2	Adjusted R^2
Spatial^T	730	1 793	0.640	0.636
NMZ Intersection Detailed^T	742	1 819	0.629	0.623
NMZ Intersection^T	741	1 815	0.631	0.626
Walk Intersection Detailed^T	744	1 811	0.633	0.626
Walk Intersection^T	741	1 803	0.635	0.631
All OSM Intersection^T	740	1 802	0.636	0.630
Intersection Ratio^T	729	1 785	0.643	0.638
All OSM Reach^T	738	1 917	0.588	0.581
All OSM Box^T	732	1 780	0.637	0.631
Walk Reach^T	736	1 782	0.644	0.639
Walk Box^T	732	1 798	0.637	0.633
Ratio Reach^T	726	1 773	0.647	0.643
Ratio Box^T	730	1 789	0.641	0.636

^{T}: Angular betweenness and FSI are transformed.

Table 8.4: Results for random forest road network models

Model	MAE	RMSE	R^2	Adjusted R^2
Spatial	760	1 725	0.666	0.663
NMZ Intersection Detailed	754	1 756	0.654	0.649
NMZ Intersection	751	1 726	0.666	0.662
Walk Intersection Detailed	729	1 708	0.673	0.668
Walk Intersection	728	1 660	0.691	0.687
All OSM Intersection	717	1 671	0.687	0.683
Intersection Ratio	755	1 715	0.670	0.666
All OSM Reach	748	1 694	0.678	0.673
All OSM Box	743	1 685	0.682	0.677
Walk Reach	747	1 712	0.671	0.668
Walk Box	747	1 689	0.680	0.677
Ratio Reach	745	1 729	0.665	0.661
Ratio Box	753	1 741	0.660	0.656

There are a few models that score better than the Spatial model, the most noticeable are the Walk Intersection and the OSM All Intersection models. Between these two models, the Walk Intersection model scores worse in MAE but better in the other three metrics. This means that the OSM All Intersection model, comparatively, has one or more bigger errors.

The results presented in Tables 8.4 and 8.3 show that road network models only score noticeably better when using random forest.

9

Selection and evaluation of final model

In the previous experiments a few models are found that score higher than the Spatial model, the highest scoring model from the experiment of reproducing previous work which is why it is used as a baseline. In this experiment, the highest scoring models from the experiments in Chapters 6, 7 and 8 are combined into one new model. This model is then inspected to see if any of the variables included can be removed without any major performance impact. After this filtering of variables, the final model is evaluated using the test data.

The final model is created using random forest since the models evaluated using random forest generally score better than the ones using negative binomial.

9.1 Models

Table 9.1: Variables included in the Combination model.

Street centrality	Built density & Land division	Attractions	Misc
Integration 5 000m	FSI	OSM 21	Walk intersections
Integration 2 500m	Plots	OSM 23	City
Integration 500m		OSM 24	Weekday
Betweenness 5 000m		OSM 25	
Betweenness 2 500m		OSM 5601	
Betweenness 500m			

FSI, Plots, City and Weekday are variables kept from the Spatial model as mentioned in Section 3.3.
The description of each OSM code is found in Table 2.4.

The Combination model includes the variables from each of the best performing models from each of the experiments in Chapters 6, 7 and 8. The highest scoring street centrality model is the Three model. The highest scoring attraction model using random forest is the OSM Railway Top. The highest scoring road network model using random forest is the Walk Intersection model. The variables from all these models sum up to the variables presented in Table 9.1.

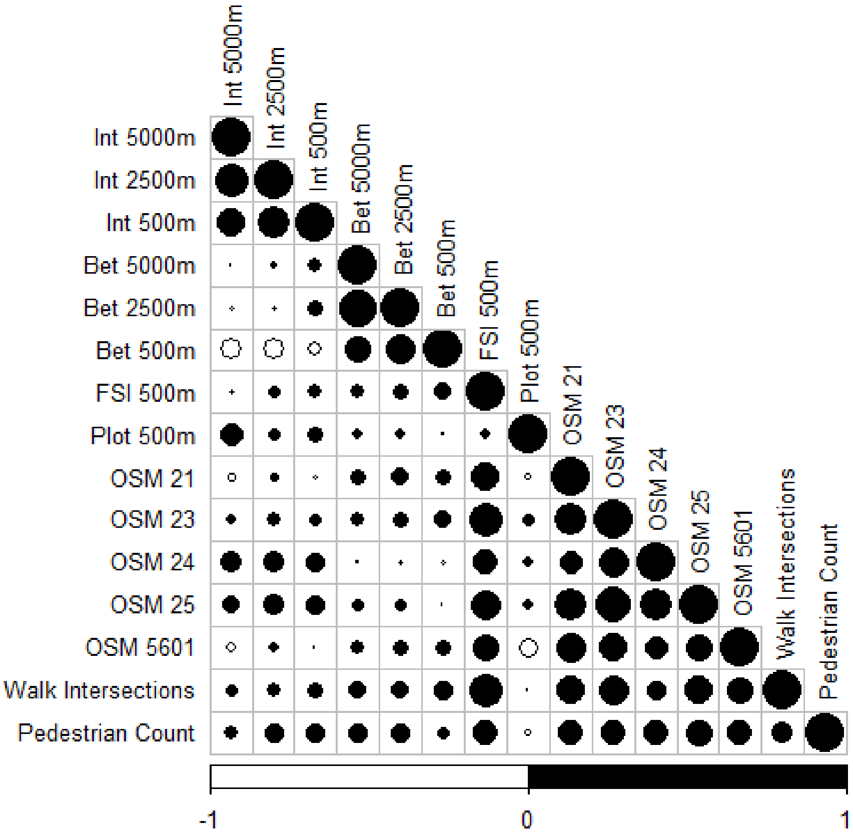

Figure 9.1: Correlations for Combination model variables.

The correlations between these variables and also the pedestrian count are presented in Figure 9.1. The most obvious correlations to point out are the correlations within groups of variables, i.e., the group of attraction variables, the group of betweenness variables and the group of integration variables. Other than these, there are a few other correlations to point out:

- OSM 21 correlates more with betweenness than integration. This indicates hospitals, pharmacies and other attractions included in OSM-21 are generally close to street segments that mediate movement within the area.
- OSM 24 and 25 correlates more with integration than betweenness. This indicates that hotels, motels, supermarkets, bakeries, etc. are generally close to street segments that are relatively easy to reach from the other street segment in the area.
- Betweenness at 500 meters correlates negatively with all the integration variables. This indicates that street segments that mediate movement in a smaller area tend to not be placed in such a way that it is easily accessible from all street segments in the area.
- FSI correlates with attractions, walk intersections and pedestrian count. This supports the fact that FSI is a very informative variable in regards to pedestrian

movement. It also indicates that attraction variables and the walk intersection variable are not as important when FSI is included, and vice versa.

- Plot correlates with integration at 5 000 meters and negatively with OSM 5601. This indicates that street segments that are relatively easy to access from most other street segments in a larger area tend to have more accessible plots around them. It also indicates that street segments close to railway station tend to have less accessible plots around them.
- Walk intersection correlates highest with FSI, attractions and pedestrian count. This indicates that densely built areas are generally more “connected”. It also indicates that more highly “connected” areas generally enjoy a larger number of pedestrians.

9.2 Selection

Table 9.2: Combination model variable importance.

Percentage of MSE Increase		Node Purity Increase	
FSI	10.93	Integration 2 500m	760 957 540
OSM 23	7.84	OSM 25	712 498 917
Integration 5 000m	7.78	FSI	544 217 204
Integration 2 500m	7.31	Betweenness 2 500m	397 781 844
Betweenness 5 000m	6.61	Betweenness 5 000m	390 716 238
Betweenness 2 500m	6.46	OSM 23	378 126 925
OSM 25	6.33	Integration 500m	367 889 341
Walk intersections	6.31	Integration 5 000m	359 413 367
Integration 500m	5.07	OSM 21	278 078 407
OSM 24	4.91	OSM 5601	242 076 788
OSM 5601	4.67	Betweenness 500m	208 445 651
Weekday	4.60	OSM 24	206 418 895
OSM 21	4.43	Walk intersections	195 218 881
Betweenness 500m	3.91	Plot 500m	133 159 434
City	2.38	Weekday	52 799 925
Plot 500m	2.20	City	21 646 961

Percentage of MSE Increase indicates how much MSE would increase if the variable in question is assigned other values.

A high Node Purity Increase indicates that splits using the variable in question leads to better grouping of the data.

The description of each OSM code is found in Table 2.4.

Starting from the Combination model, multiple models are evaluated where the first one includes all variables and each following model has one more variable removed. Which variable to remove in which order is chosen using the importance of each variable. These importances are presented in Table 9.2 in the form of Percentage of MSE Increase and Node Purity Increase. The reason why this is done using the importance measurements instead of correlation is because the importance measurements directly indicate how useful each variable is in respect to the others. The importance scores are calculated from running cross-validation on the Combination model. This removal of variables is done to keep the model as simple as possible and can also limit the possibility of the model overfits to the data.

Table 9.3: Results for combination model and variable filtering

Model	MAE	RMSE	R^2	Adjusted R^2
Spatial	760	1 725	0.666	0.663
Combination	640	1 493	0.750	0.743
-City	+3	+15	−0.005	−0.004
-Plot	−8	−9	+0.003	+0.004
-Weekday	+4	+4	−0.001	+0.000
-Bet500	+13	+27	−0.009	−0.007
-OSM21	+8	+21	−0.007	−0.004
-OSM5601	+17	+46	−0.016	−0.013
-OSM24	+19	+25	−0.008	−0.005
-Int500	+27	+37	−0.012	−0.009
-WalkIntersection	+28	+53	−0.018	−0.014
-Int5000	+39	+65	−0.022	−0.018
-Bet2500	+53	+137	−0.048	−0.043

Each of the metrics reported after the Combination model are relative scores, relative to the Combination model and the model names are based on which of the variables have been excluded in that model. E.g., in the -City model, the City variable has been excluded and in the -Plot model, the City variables and the number of plots variable have been excluded.

Table 9.4: Variables included in the Final model.

Street centrality	Attractions	Built density & Road network
Integration 5 000m	OSM 21 (Hospitals, pharmacies, ...)	Floor Space Index (FSI)
Integration 2 500m	OSM 23 (Restaurants, pubs, cafes, ...)	Walk intersections
Integration 500m	OSM 24 (Hotel, motels, ...)	
Betweenness 5 000m	OSM 25 (Supermarkets, bakeries, ...)	
Betweenness 2 500m	OSM 5601 (A larger railway station.)	
Betweenness 500m		

The results of the variable filtering are presented in Table 9.3. The -Weekday model scores the same in Adjusted R^2 as the combination model and almost as good in all the other metrics. It is difficult to know which of these models is the best choice but because the -Weekday model scores almost as well as the Combination model, it is chosen as the final model. A summary of the variables included in the Final models is presented in Table 9.4.

Table 9.5: Results for Final model evaluated using the test data

Model	MAE	RMSE	R^2	Adjusted R^2
Spatial	842	1889	0.491	0.449
Final	810	2024	0.416	0.301

9.3 Evaluation

Table 9.5 presents the metrics from evaluating with the test data. The results here seem to indicate that the Spatial model performs better than the Final model, all of the metrics except for MAE are in favor for the Spatial model. It is, however, worth noting that the test data is only approximately 10 percent of the data collection, in total 80 street segments. With that in mind, it is difficult to know how these scores represent the actual performance of the models.

The Spatial model scores an R^2 of 0.491 which can be interpreted as describing 49 percent of the variance in the data. This is noticeably lower than when the Spatial model is evaluated using cross-validation on the training data, 67 percent, those scores are presented in Table 9.3.

9.3.1 Exploring highly over- and underestimated counts

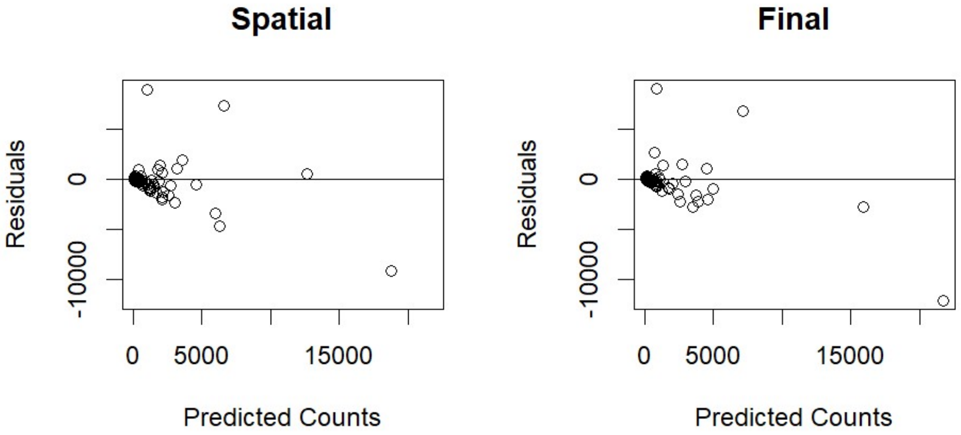

Figure 9.2: Residuals for Final and Spatial model using test data.

The fact that the Final model scores worse in all the metrics except MAE indicates that there are one or more larger errors. The residuals for each of the models are presented in Figure 9.2. There are three very noticeable residuals, two underestimated and one overestimated. These three residuals look similar for both the Spatial and the Final model but the overestimated error is a bit larger for the Final model. This difference can have a big impact when the error is squared, therefore it can be the explanation for why the Final model scores worse.

An interesting fact to note here is that two of these three noticeable residuals are street segments in Norrmalm, Stockholm which is the area that is explored in Section 5.4. So this again indicates that there is either something different with Norrmalm that is not captured in the model or that the measurements in Norrmalm are somehow not representative of the "usual" pedestrian counts in the area.

Table 9.6: Results for Final model evaluated using the test data with the largest error removed

Model	MAE	RMSE	R^2	Adjusted R^2
Spatial	738	1602	0.585	0.550
Final	668	1520	0.626	0.551

The metrics recalculated when removing the top error, the overestimated street segment, are presented in Table 9.6. The metrics are clearly better because the top error is removed but the interesting part is how much better the metrics are. The R^2 score for the Final model increases with more than 0.2. The fact that one data point affects the metrics as much as this opens up the question if the test set is representative of the general population of street segments.

9.3.2 Cross-validation using all data

Table 9.7: Results for Final model evaluated using cross-validation on all the data

Model	MAE	RMSE	R^2	Adjusted R^2
Spatial	760	1751	0.646	0.643
Final	656	1533	0.729	0.724

In order to test if the test data is not representative of the general population of street segments, a 10-fold cross-validation is run on all the data. This can be seen as 10 different runs with randomly chosen training set of 90 percent and test set of 10 percent (in this case the test set is better viewed as a validation set). The result from this cross-validation is presented in Table 9.7. From the results it is possible to see that the Final model scores higher than the Spatial model in all metrics. It is also possible to see that the decrease in the metrics is similar between both models. The Spatial model decrease from an R^2 of 0.666 to an R^2 of 0.646 and the Final model decrease from an R^2 of 0.749 to an R^2 of 0.729.

As with all other experiments in this thesis, this cross-validation is run 10 times using different random seeds. The results from each individual run is not presented here but they all indicate the same things, the Final model performs noticeably better in all of them. This result does support the fact that the test data is not representative of the general population of street segments, this could mean that it was by chance that the Spatial model performs better than the Final model using the original train and test data split.

10

Discussion

This chapter includes discussion about the result from the experiments performed in this thesis. There is also a section that explores ethical considerations.

10.1 Discussion

10.1.1 Reproducing previous work

There are three noteworthy findings from reproducing the previous work, Stavroulaki et al. [30]. The first finding is that the negative binomial performs much better when transforming FSI and betweenness. The second finding is that there are two very high counts in Norrmalm, Stockholm, and when considering the spatial layout and the shopping malls there, they seem possible. The third finding is that the Spatial model (the model that includes street centrality, built density and land division) scores higher than the Attraction model (the model that includes attractions in addition to the variables included in the Spatial model). That indicates that the attraction variables created previous to this thesis complicates the model more than it provides new information to the model when training on this data. If given a larger set of data, there is a possibility that the Attraction model could perform better than the Spatial because it is then possible for the model to learn more complex relationships.

To summarize the best performing model from this experiment, the Spatial model, it includes the variables that represent street centrality, built density and land division. More specifically the variables are:

- Angular integration within 1 000 meters
- Angular betweenness within 3 500 meters
- FSI accessible within 500 meters
- Number of accessible plot within 500 meters

10.1.2 Exploring different representations of street centrality

One noteworthy finding from exploring different representations of street centrality is that the highest scoring model using different representations of street centrality is the Three model (the model that includes street centrality measurements using three different radii). The Three model includes the following street centrality variables:

- Angular integration within 500, 2 500 and 5 000 meters
- Angular betweenness within 500, 2 500 and 5 000 meters

The Three model probably scores higher than the Spatial model because it includes the street centrality variables at more ranges, providing more information to the model about the street. This information is also included in the All model (the model that includes all the street centrality variables, i.e., street centrality measurements using ten different radii). However, the All model scores lower than the Three model. This indicates that the information gain with including all radii is less than the increase of complexity in finding a relationship to the pedestrian movement count when training on this data. If given a larger set of data, there is a possibility that the All model could perform better because it is then possible for the model to learn more complex relationships.

10.1.3 Exploring different representations of attractions

There are three noteworthy findings from exploring different representations of attractions. The first finding is that attraction variables counted on the street segment generally correlate lower with the pedestrian counts in comparison to counts within 500 meters. The second finding is that the attraction variables with counts of attractions generally correlate higher with the pedestrian counts in comparison to the variables with the distance to the nearest attraction. The third finding is that the highest scoring model for negative binomial and random forest are not the same.

When using negative binomial, the highest scoring model is the OSM Distance model (the model that includes the distance to the nearest attraction within each OSM category). When using random forest, the highest scoring model is the OSM Railway Top model (the model that includes the five highest correlated variables that count the number of attractions within 500 meters for each OSM category). The reason why the negative binomial and random forest score higher with different models is probably because of the linear nature of negative binomial. The OSM Railway Top model, which scores best using random forest, contains attraction count variables and those variables contain a considerable amount of zeroes, on average about 40 percent. Because of that, the negative binomial might not be good at finding a linear relationship. Random forest can naturally treat the zero values and the non-zero values differently, therefore, it is not affected by this fact.

The new variables included in the OSM Railway Top model, the highest scoring using random forest, are the attraction counts within 500 meters for the following categories:

- OSM-21: Hospitals, pharmacies, ...
- OSM-23: Restaurants, pubs, cafes, ...
- OSM-24: Hotel, motels, and other places to stay the night
- OSM-25: Supermarkets, bakeries, ...
- OSM-5601: A larger railway station of mainline rail services.

10.1.4 Exploring new variables based on road network

There are three noteworthy findings from exploring new variables based on road network. The first finding is that network length calculated using network reach (a way of limiting the included road network from a point by using the distance traveled) correlates higher with the pedestrian counts in comparison to using a bounding box (a way of limiting the included road network from a point within a square with each side being the same distance from the center). The second finding is that variables created from the walk network (the road network for pedestrians) correlated higher with the pedestrian movement counts in comparison to the bike, drive (the road network for cars, motorcycles, etc.) and non-motorized (the road network for non-motorized traffic) networks. The third finding is that only random forest, not negative binomial, scored noticeably higher than the Spatial model with any of the models created during this experiment. The highest scoring model using random forest is the Walk Intersection model which includes the count of walk intersections within 500 meter network reach.

The reason why only random forest scored noticeably better with road network models could be that the relation between the road network variables and the remaining variance after considering all the other variables might not be linear.

The summary of the findings in the experiment is:

- Using network reach to limit the road network seems more informative than limiting it using a bounding box.
- The walk network seems more informative than the bike, drive and non-motorized networks.
- The highest scoring model for random forest includes the count of intersections in the walk network limited by a 500 meter network reach.

10.1.5 Selection and evaluation of the final model

In this experiment, a model called Combination is constructed from combining all of the variables from the highest scoring model using random forest for each of the preceding experiments. This model is then used as the basis for creating the model called Final. The Final model is found by sequentially removing one variable at a time from the Combination model based on the importance given by random forest. The Final model is then chosen by arbitrarily finding a good balance between the performance and the number of variables included. See Figure 10.1 for the variables

Table 10.1: Variables included in the Final model.

Street centrality	Attractions	Built density & Road network
Integration 5 000m	OSM 21 (Hospitals, pharmacies, ...)	Floor Space Index (FSI)
Integration 2 500m	OSM 23 (Restaurants, pubs, cafes, ...)	Walk intersections
Integration 500m	OSM 24 (Hotel, motels, ...)	
Betweenness 5 000m	OSM 25 (Supermarkets, bakeries, ...)	
Betweenness 2 500m	OSM 5601 (A larger railway station.)	
Betweenness 500m		

that are included in the Final model. The rest of the experiment is to evaluate the Final model in comparison to the Spatial model (the best performing model when reproducing the previous work).

A summary of the initial results in this experiment is provided below:

- The Combination model scores noticeably higher than the Spatial model using cross-validation on the training data, 0.750 compared to 0.666 in R^2.
- The Final model excludes three variables from the Combination model and scores an R^2 of 0.749.
- The Final model scores worse than the Spatial model in most of the metrics when evaluating on the test data, 0.416 in comparison to 0.491 in R^2. However, the Final model scores better in MAE, 810 in comparison to 842.

The fact that the Final model scores worse on the test data in all metrics except MAE is surprising given the fact that the Final model scores considerably better than the Spatial model on the training data. There are two possible reasons for this, one is that the Final model generalizes worse than the Spatial model and the other is that the test set does not represent the general population of street segments. When the second reason is explored we find that one street segment, amongst the 80 included in the test data, affects the metrics greatly. If this one street segment is excluded, the Final model scores an R^2 of 0.626 instead of 0.416. In this case, the Final model also scores better than the Spatial model, which scores an R^2 of 0.585.

Whether the test set is representative of the general population of street segments is further explored by running a 10-fold cross-validation on all of the data, i.e., including both the training and the test set. The result of this is that Final model scores noticeably higher than the Spatial model, 0.729 in comparison to 0.643 in R^2.

A summary of the results from this further analysis is provided below:

- When excluding a street segment that is greatly over-estimated by both the Final and Spatial model, then the Final model performs better than the Spatial model, 0.626 in comparison to 0.585 in R^2.
- When running a 10-fold cross validation on all of the data (both training and test data), the Final model scores considerably better than the Spatial model, 0.729 in comparison to 0.643 in R^2. Note also that the decrease in performance

in comparison to the result when running on only the training data is similar between the Spatial and Final model.

The fact that the Final model scores higher than the Spatial when running cross-validation on all the data is to be expected. It is expected since the Final model scores higher than the Spatial model on the training data, which is the majority (90 percent) of the data used in this cross-validation. One interesting thing to note here, however, is that both models decrease similar in performance when comparing with the result from using only the training data. If the Spatial model generalizes better than Final model, then we would expect that the Final model would decrease more in performance than the Spatial model when including the test data. Since this is not the case, a reasonable assumption is that the Spatial model by pure chance scored better in the original train and test data split and not because it generalizes better.

10.1.6 Main findings in relation to this thesis' objectives

The objectives in this thesis are the listed below, with an explanation of the findings.

Evaluate random forest as an alternative algorithm to negative binomial.
Random forest seems to be more robust than the negative binomial in two ways. There is no need to transform any variables and it does not perform worse using zero-inflated variables. Random forest, in general, scores higher than negative binomial using MAE, RMSE, R^2 and Adjusted R^2. However, it is important to note that the choice of algorithm should be based on the purpose of using it. Negative binomial is a probabilistic model which means that it provides probabilities for each of the samples in the data, this is not the case for random forest. So negative binomial could be more useful if the purpose is mainly to model and analyze the data rather than performing predictions.

Explore new variables to further explain pedestrian movement counts.
Many new variables are explored in this thesis. Such as:

- polynomial estimation of integration and betweenness, see Section 6.1.1 for more details.
- attraction variables based on OSM codes (both counts within 500 meters and distance to nearest attraction), see Section 7.1.1 for more details.
- road network lengths within 500 meters and intersection counts within 500 meters (total count and categorized into three-way, four-way and more-way intersections), see Sections 8.1.1 and 8.1.2 for more details.

Both the road network lengths and the intersection counts are evaluated using four different road networks, the OSM walk, bike and drive network and also a non-motorized network created by Berghauser Pont et al. [3].

Find, from the variables available, the set of variables that best explain the pedestrian movement counts.
The best set of variables found is the one that composes the Final model, see Table

10.1. These variables include street centrality variables within three different radii, built density in the form of FSI, attraction counts for five different OSM categories and the intersection count based on the OSM walk network. See Figure 9.1 and Table 9.2 for more details of the correlation and importance for these variables.

10.2 Ethics

This thesis explores the relation between the built environment and pedestrian movement counts. If findings are used as a basis for change of the built environment, it is important to know what the potential biases in the data are. Otherwise it can lead to exclusion or down-prioritization of specific areas or groups of people.

It is possible to imagine that there are generally fewer people having their phone and Wi-Fi turned on in some specific areas. An example of those would be areas with elderly care, meditation centers or schools. The attraction variables for schools did not show any correlation to the pedestrian counts but whether or not that gives any indication that people tend to have their phones or Wi-Fi turned off in those areas is difficult to determine. It might just be that schools are placed in areas where there tend to be few pedestrians.

It is also possible to imagine that different groups of people generally have different number of phones. For example, it is common for people in some professions to have a work phone and then also a personal phone. This may mean that this group of people are likely to be counted more than the general population. On the other hand, it might be less common for elderly, children or people with lower income to have phones with Wi-Fi turned on. This may mean that these groups of people are counted less than the general population when using this technology for measurements. These biases for different groups of people support the idea that there can be a bias in some areas, if we make the assumption that different areas tend to be visited more or less by different groups of people.

There were as mentioned manual counts for some of the select street segments as well, and these counts would not be biased in the same way. These counts were then used to scale the pedestrian counts. However, since the scaling factor was the same for all street segments, it seems unlikely that it removed the possible biases discussed above.

11

Conclusion

This chapter includes conclusions from this thesis and future work.

11.1 Conclusion

This thesis had a goal to help urban designer, planners and policy-makers understand how urban environments function by extending previous work in predicting pedestrian movement counts. The final result gave a model that scored an R^2 of 0.729 compared to the highest scoring model created in Stavroulaki et al. [30] which scored 0.646 when evaluating using a 10-fold cross-validation on all the available data. This score might be high enough for the model to be used in practice, at least in the measured cities and areas.

It is, however, fair to assume that the model could perform differently on street segments not included in the data. The important question is then how much worse it would perform. This can only be tested reliably by gathering new data. It is also important to keep in mind that the street segments included in the data used in this thesis were measured for only one day. The results might differ if an average count for each street segment was used instead.

No matter if the predictive model is useful by itself or not, there are findings in this thesis that may be useful by themselves. Those findings are (all in regards to pedestrian movement counts):

- Count of attractions within the surrounding area seems more informative than the count of attractions on a street segment.
- Count of attractions seem more informative than the distance to the nearest attraction.
- The OSM walk network seems more informative than the less detailed non-motorized network.
- Using network reach to constrain the surrounding area when creating variables seems to be better in comparison to a bounding box.

There are also some methodological findings for future work in predictive models for pedestrian movement counts. Those are:

- Random forest is preferred over the negative binomial when the goal is to predict pedestrian counts.
- Splitting the training and test data using a stratified random split based on street types (presented in Table 3.2) seems like a good approach when weighing the advantages and disadvantages (presented in Section 4.1), even better if it is possible to combine with splitting on areas or cities.
- The test data set should include considerably more than 80 street segments since that amount did not seem to provide reliable test results.

11.2 Future work

A few possible extensions for this work in predictive models for pedestrian movement is discussed in this section.

The most valuable extension is to collect more data, more streets and more cities. During future measurements it would probably be useful to note if there are any events, major sales or other things that could temporarily increase or decrease the number of pedestrians. There is currently a data set available from Gothenburg[1], this data was collected by Trafikkontoret in Gothenburg, again by using the service provided by Bumbee Labs, Stockholm. However, this data is not exactly the same as the data used in this thesis. There are two important differences. The first one is that the gates in the Gothenburg data were not placed in every intersection in the area covered. It is therefore difficult to aggregate count per street segment. The second one is that only the raw data is available, that means that the data processing performed previous to this thesis, as mentioned in Section 2.2, would be needed before the data can be used in the same way.

One other possible extension of this work is to create another type of categorizations of street segments based on attractions, similar to centrality and density types developed by Berghauser Pont et al. [3]. This categorization could be similar to the ones used in Özbil et al. [35].

In the experiment in Chapter 8 the road length and intersections of the car and bike road networks were included because of a statement in Forsyth and Southworth [12]. The statement referred to here is that walkability has decreased when adding other means of transport. One way of creating variables to represent this difference could be to include counts of cars and/or bikes. These counts can be created in the same way as the pedestrian counts, by filtering on the speed of movement as discussed in Section 2.3.1. While this would not be variables that are possible to use when the making prediction on non-measured street segments, it could still give some interesting insights into the relationship between pedestrians, bikes and cars.

There are then also a few other extensions, however, smaller ones. Those are, e.g., evaluating:

[1] `https://data.goteborg.se/`

- attraction counts using zero-inflated negative binomial.
- attraction counts at different radii.
- attraction counts only for attractions at ground floor level.
- average distance to attractions using route directness (ratio between the shortest path and the straight line path).
- network length for the non-motorized network.
- OSM attraction variables created by calculating the distance in the non-motorized network instead of the OSM walk network.
- variables representing the "centrality" in terms of the public transport network.

Bibliography

[1] Vladimir Agafonkin. Leaflet. `http://leafletjs.com/`, 2019. Accessed: 2019-12-06.

[2] Richard Becker. *The new S language.* CRC Press, 2018.

[3] Meta Berghauser Pont, Gianna Stavroulaki, and Lars Marcus. Urban types, based on network centrality and built density, and their impact on pedestrian movement. In *Environment and Planning B*, 2019.

[4] Meta Yolanda Berghauser Pont and Per André Haupt. Space, density and urban form. 2009. URL `http://resolver.tudelft.nl/uuid:0e8cdd4d-80d0-4c4c-97dc-dbb9e5eee7c2`. Doctoral thesis, Delft University of Technology.

[5] M.Y. Berghauser Pont, G. Stavroulaki, J.A. Lopes Gil, L. Marcus, M Serra, B. Hausleitner, J. Olsson, E. Abshirini, and A. Dhanani. Quantitative comparison of cities: Distribution of street and building types based on density and centrality measures. *XI SSS: 11th International Space Syntax Symposium 2017*, 2017.

[6] Evgeniya Bobkova, Lars Marcus, and Meta Berghauser Pont. Multivariable measures of plot systems: describing the potential link between urban diversity and spatial form based on the spatial capacity concept. In *Proceedings of the 11th Space Syntax Symposium*, 07 2017.

[7] Evgeniya Bobkova, Lars Marcus, Meta Berghauser Pont, Ioanna Stavroulaki, and David Bolin. 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, 2019.

[8] A Colin Cameron and Frank AG Windmeijer. R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, 14(2):209–220, 1996.

[9] CARTO. Carto: Unlock the power of spatial analysis. `http://cartodb.com/attributions`, 2019. Accessed: 2019-12-06.

[10] Ruth Conroy Dalton. The secret is to follow your nose: Route path selection and angularity. *Environment and Behavior*, 35(1):107–131, 2003. doi: 10.1177/0013916502238867. URL `https://doi.org/10.1177/0013916502238867`.

[11] Peggy Edwards and Agis D Tsouros. *Promoting physical activity and active living in urban environments: the role of local governments.* WHO Regional Office Europe, 2006.

[12] Ann Forsyth and Michael Southworth. Cities afoot—pedestrians, walkability and urban design. *Journal of Urban Design*, 13(1):1–3, 2008. ISSN 1357-4809.

[13] Joseph M Hilbe. *Negative binomial regression.* Cambridge University Press, 2011.

[14] B Hillier, A Penn, J Hanson, T Grajewski, and J Xu. Natural movement: Or, configuration and attraction in urban pedestrian movement. *Environment and Planning B: Planning and Design*, 20(1):29–66, 1993. doi: 10.1068/b200029. URL `https://doi.org/10.1068/b200029`.

[15] Bill Hillier and Shinichi Iida. Network and psychological effects in urban movement. In Anthony G. Cohn and David M. Mark, editors, *Spatial Information Theory*, pages 475–490, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-32020-3.

[16] WRG Hillier, Tao Yang, and Alasdair Turner. Normalising least angle choice in depthmap-and how it opens up new perspectives on the global and local analysis of city space. *Journal of Space syntax*, 3(2):155–193, 2012.

[17] Erik Håkansson. Statistical modelling of pedestrian flows. Master's thesis, University of Gothenburg, 06 2019. URL `http://hdl.handle.net/2077/60462`.

[18] Andy Liaw. randomforest v4.6-14. `https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest`, 2020. Accessed: 2020-02-16.

[19] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R news*, 2(3):18–22, 2002.

[20] Jeannette Montufar, Jorge Arango, Michelle Porter, and Satoru Nakagawa. Pedestrians' normal walking speed and speed when crossing a street. *Transportation Research Record*, 2002(1):90–97, 2007.

[21] Anne Vernez Moudon, Paul M Hess, Mary Catherine Snyder, and Kiril Stanilov. Effects of site design on pedestrian travel in mixed-use, medium-density environments. *Transportation Research Record*, 1578(1):48–55, 1997.

[22] Vinicius M Netto, Renato Saboya, Julio Celso Vargas, Lucas Figueiredo, Cássio Freitas, and Maira Pinheiro. The convergence of patterns in the city:(isolating) the effects of architectural morphology on movement and activity. In *Proceedings of the 8th International Space Syntax Symposium*, pages 1–32. Pontificia Universidad Católica de Chile Santiago de Chile, 2012.

[23] OpenStreetMap Foundation. OpenStreetMap. `http://www.openstreetmap.org/copyright`, 2019. Accessed: 2019-12-06.

[24] A Penn, B Hillier, D Banister, and J Xu. Configurational modelling of urban

movement networks. *Environment and Planning B: Planning and Design*, 25(1): 59–84, 1998. doi: 10.1068/b250059. URL `https://doi.org/10.1068/b250059`.

[25] Meta Berghauser Pont and Lars Marcus. What can typology explain that configuration cannot. In *SSS10 Proceedings of the 10th International Space Syntax Symposium*, 2015.

[26] R-INLA. R-inla. `http://www.r-inla.org/`, 2020. Accessed: 2020-02-16.

[27] Stephen Read. Space syntax and the Dutch City. *Environment and Planning B: Planning and Design*, 26(2):251–264, 1999. doi: 10.1068/b4425. URL `https://doi.org/10.1068/b4425`.

[28] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2): 319–392, 2009.

[29] Gianna Stavroulaki, Meta Berghauser Pont, Lars Marcus, Kailun Sun, and Staffan Liljestran. Methodology and results of an international observational study on pedestrian movement tracking anonymised wi-fi signals from mobile phones. AESOP Annual Congress, Gothenburg, 2018 July 10-14, 2018. URL `http://www.trippus.se/eventus/userfiles/101941.pdf`.

[30] Gianna Stavroulaki, David Bolin, Meta Berghauser Pont, Lars Marcus, and Erik Håkansson. Statistical modelling and analysis of big data on pedestrian movement. In *12th International Space Syntax Symposium*, 2019.

[31] Gianna Stavroulaki, Daniel Koch, Ann Legeby, Lars Marcus, Alexander Ståhle, and Meta Berghauser Pont. Documentation PST 20191122, 11 2019. DOI: 10.13140/RG.2.2.25718.55364.

[32] United Nations. The world's cities in 2016. `https://www.un.org/en/development/desa/population/publications/pdf/urbanization/the_worlds_cities_in_2016_data_booklet.pdf`, 2016. Accessed: 2019-11-04.

[33] Jacob Wasserman. mplleaflet. `https://github.com/jwass/mplleaflet`, 2018. Accessed: 2019-12-06.

[34] Ayse Özbil, John Peponis, and Brian Stone. Understanding the link between street connectivity, land use and pedestrian flows. *Urban Design International*, 16 (2):125–141, Summer 2011. URL `https://search-proquest-com.ezproxy.ub.gu.se/docview/864304211?accountid=11162`. Palgrave Macmillan, a division of Macmillan Publishers Ltd 2011; Last updated - 2013-10-08.

[35] Ayse Özbil, Demet Yesiltepe, and Görsev Argin. Modeling walkability: The effects of street design, street-network configuration and land-use on pedestrian movement. *ITU J Faculty Arch*, 12(3):189–207, 2015. URL `https://dx.doi.org/`.

A

Appendix 1

A.1 Attraction calculation comparison

This appendix presents the results from both the ways of calculating attraction. The results presented in Chapter 7 are from the attraction variables calculated from the center of the street segment.

Table A.1: Variable correlation for all OSM attraction variables.

Center				Nearest node			
OSM Counts		OSM Distance		OSM Counts		OSM Distance	
OSM-20	0.160	OSM-20	-0.160	OSM-20	0.171	OSM-20	-0.199
OSM-21	0.419	OSM-21	-0.193	OSM-21	0.413	OSM-21	-0.227
OSM-22	0.379	OSM-22	-0.226	OSM-22	0.383	OSM-22	-0.251
OSM-23	0.425	OSM-23	-0.138	OSM-23	0.427	OSM-23	-0.180
OSM-24	0.409	OSM-24	-0.244	OSM-24	0.409	OSM-24	-0.238
OSM-25	0.487	OSM-25	-0.143	OSM-25	0.495	OSM-25	-0.190
OSM-26	0.280	OSM-26	-0.241	OSM-26	0.281	OSM-26	-0.256
OSM-27	0.245	OSM-27	-0.213	OSM-27	0.258	OSM-27	-0.210
OSM-56	0.343	OSM-56	-0.108	OSM-56	0.343	OSM-56	-0.133
OSM-5601	0.405	OSM-5601	-0.158	OSM-5601	0.400	OSM-5601	-0.141

Table A.2: Results for negative binomial attraction model using attraction variables calculated from the center of the street segment

Model	MAE	RMSE	R^2	Adjusted R^2
OSM Groups^T	749	1 902	0.594	0.584
OSM Railway^T	766	1 936	0.580	0.568
OSM Distance^T	713	1 727	0.666	0.656
OSM Railway Top^T	746	1 896	0.597	0.589
OSM Distance Top^T	720	1 739	0.661	0.654
OSM All Top^T	744	1 866	0.610	0.599

Table A.3: Results for negative binomial attraction model using attraction variables calculated from the nearest node

Model	MAE	RMSE	R^2	Adjusted R^2
OSM Groups^T	742	1 855	0.614	0.604
OSM Railway^T	745	1 867	0.609	0.598
OSM Distance^T	714	1 763	0.652	0.642
OSM Railway Top^T	736	1 829	0.625	0.617
OSM Distance Top^T	717	1 768	0.650	0.643
OSM All Top^T	726	1 820	0.629	0.618

Table A.4: Results for random forest attraction model using attraction variables calculated from the center of the street segment

Model	MAE	RMSE	R^2	Adjusted R^2
OSM Groups	723	1 808	0.634	0.625
OSM Railway	720	1 819	0.629	0.619
OSM Distance	758	1 798	0.637	0.628
OSM Railway Top	719	1 771	0.649	0.642
OSM Distance Top	729	1 783	0.652	0.646
OSM All Top	729	1 783	0.644	0.634

Table A.5: Results for random forest attraction model using attraction variables calculated from the nearest node

Model	MAE	RMSE	R^2	Adjusted R^2
OSM Groups	725	1 782	0.644	0.635
OSM Railway	715	1 766	0.650	0.641
OSM Distance	776	1 843	0.619	0.609
OSM Railway Top	708	1 679	0.684	0.678
OSM Distance Top	760	1 828	0.625	0.619
OSM All Top	729	1 751	0.656	0.647