



# Analysis of Urban Space Networks for Recreational Purposes based on Mobile Sports Tracking Application Data

Rusnė Šilerytė

Technische Universiteit Delft



# Analysis of Urban Space Networks for Recreational Purposes based on Mobile Sports Tracking Application Data

by

**Rusnė Šilerytė**

Student number: 4329244

E-mail: r.sileryte@student.tudelft.nl

in partial fulfillment of the requirements for the degree of

**Master of Science**  
in Geomatics

at the Delft University of Technology,  
to be defended publicly on Friday June 26, 2015 at 10:45 AM.

Supervisors: Pirouz Nourian  
Prof. Dr. Stefan van der Spek  
Dr. Hugo Ledoux

An electronic version of this thesis is available at: <http://repository.tudelft.nl/>

All developed scripts are available at: <https://github.com/rusne/RunabilityIndex>



## ABSTRACT

---

Even though studies of the built environment's impact on citizens' physical activity have become an aspiring topic in the recent years, up to now the most investigated topic is transport-related walking or cycling. However, little is known about the patterns of leisure-related physically active travels. One of the reasons behind this lack of research is the matter of collecting ground truth data for validation. Yet the data suitable for this kind of research is voluntary produced by people using sports tracking applications.

Thus the aim of this research is to develop a method to acquire, manage and process the data from sports tracking applications in such a way that it would serve as a ground truth not only for examining urban recreational travel patterns but also for modelling the phenomena. In other words, the goal is being able to define where recreational activities happen, where they do not and finally, use this knowledge to give an indication to every space of how likely it is that the space *is* or *will be* used for recreation.

The Master Thesis report describes methods used for mobile sports tracking application data acquisition and processing in tandem with OpenStreetMap and Eurostat Urban Atlas datasets. The processed data is used as a ground truth in order to calibrate and validate the developed Runability Index, which has been introduced as an indication of space potential to be used for recreation, based on the well-known measure of walkability.

The developed method for acquiring ground truth urban recreational travel data has proved to be suitable for investigation of related matters in European cities with sufficient application users. Even though, the Runability Index has not provided enough correlation with the collected data to be validated, it has ascertained that transport-based active travels have different characteristics than leisure-based ones and therefore need to be explored separately.

**Keywords:** GPS tracking, sports tracking applications, urban recreational travels, urban space network, OpenStreetMap, Urban Atlas, Runability Index.



## ACKNOWLEDGEMENTS

---

Sincere thanks to my supervisors Pirouz Nourian, Stefan van der Spek and Hugo Ledoux for all advise, attention, motivation and inspiration.

Moreover, I would like to thank Nikita Barsukov for helping with data acquisition inquiries. Wilko Quak for helping with database issues and providing access to the TU Delft server. Carl Chen for help with C++. Justinas Liubinskas for running every day and that way helping me to track workout ids. Martijn Meijers and Ravi Peters for centerline advices. *endomondo* for not objecting using self-collected data from their resources for the research purposes. And, finally, Aiste Eidukeviciute, Iris Theunisse and Dimitris Zervakis for all kinds of support.

## ABBREVIATIONS

---

BMI – Body Mass Index

COTS - Commercial off-the-shelf

CRS – Coordinate Reference System

ESRI - Environmental Systems Research Institute

GIN - Generalized Inverted Index

GIS – Geographic Information System

GNSS - Global Navigation Satellite System

GPS – Global Positioning System

HTML (html) – Hyper Text Markup Language

HTTP – Hyper Text Transfer Protocol

JSON – Java Script Object Notation

LUM – Land Use Mix

NACH – Normalised Angular Choice

NDVI - Normalized Difference Vegetation Index

OGC – Open Geospatial Consortium

OSM – Open Street Map

RMSE - Root Mean Squared Error

SQL – Structured Query Language

URL - Uniform Resource Locator

VGI – Volunteered Geographic Information

GiST - Generalized Search Trees

USN – Urban Space Network



# TABLE OF CONTENTS

---

Abstract .....	3
Acknowledgements .....	4
Abbreviations.....	5
List of used datasets .....	8
1 . Introduction.....	11
1.1 Problem Statement and Relevance.....	12
1.2 Goal and Scope of the Research.....	15
1.3 Research Question .....	16
1.4 Methodology .....	17
1.5 Method description.....	19
2 . Related research.....	23
2.1 Urban Analysis Based on GPS Data .....	23
2.2 Physical Activity in Urban Environments.....	24
2.3 Urban Network Analysis.....	26
3 . Data Acquisition.....	27
3.1 Choice of Application .....	27
3.2 Scheme of Data Acquisition .....	29
3.3 Data Limitations .....	33
3.3.1 Application Users .....	33
3.3.2 Incorrect or irrelevant attributes.....	33
3.3.3 Incorrect GPS tracks.....	34
3.4 Case Study Cities .....	36
3.4.1 Vilnius.....	38
3.4.2 Valencia.....	39
3.4.3 Gothenburg.....	40
4 . Urban Space Network.....	41
4.1 Network Definition .....	41
4.2 Data Sets .....	42
4.2.1 OpenStreetMap .....	42
4.2.2 European Urban Atlas Road Land-use Data .....	43
4.3 Network Generation Framework .....	45
4.3.1 Inconsistencies of Datasets.....	45
4.3.2 Space Centerline .....	46
4.3.3 Integration of Datasets and Generalisation .....	48
4.3.4 Post-processing.....	53

4.4	Method Limitations .....	56
5 .	Actual Recreational Usage .....	57
5.1	Data Management .....	58
5.2	Filtering GPS trajectories .....	59
5.3	GPS Track Snapping on an Urban Space Network .....	61
5.4	Value of Actual Recreational Usage.....	65
5.5	Validation and Verification .....	68
6 .	Runability Index .....	71
6.1	Value of Greenness .....	72
6.2	Value of Land Use Mix .....	74
6.3	Value of Network Centrality .....	78
6.4	Value of Residential Density .....	82
6.5	Unification of Measures .....	83
6.6	Comparison.....	84
6.7	Limitations .....	91
7 .	Future Work and Conclusions .....	93
7.1	Conclusions .....	93
7.2	Discussion and Recommendations .....	95
	References.....	97
	Appendix A. Comparison Values between different variations of potential recreational usage and the actual recreational usage .....	105
	Appendix B. Scatter plots .....	107
	Reflection .....	109



## LIST OF USED DATASETS

### Endomondo workout GPS tracks

<i>Dataset Purpose</i>	Actual data of active urban recreational travels
<i>Source</i>	Endomondo.com public workouts
<i>Temporal Coverage</i>	Mar 2014 – Apr 2015
<i>Entity type</i>	Vector: JSON polyline
<i>Resolution/ Precision</i>	0.1m
<i>Geographic coverage</i>	Europe
<i>CRS</i>	WGS84
<i>Rights</i>	The database extraction right does not apply for copying parts of the database since the extracted content cannot be considered substantial (<1% of the database). Moreover, the main purpose of <i>endomondo</i> is not data collection therefore the copied data can only be considered a by-product of the application.

### European Urban Atlas land-use data

<i>Dataset Purpose</i>	Land use data
<i>Source</i>	Directorate-General Enterprise and Industry
<i>Temporal Coverage</i>	2005-2007
<i>Entity type</i>	Vector: ESRI Polygon
<i>Resolution/ Precision</i>	2.5m
<i>Geographic coverage</i>	LUZ (Larger Urban Zones) of European Union countries
<i>CRS</i>	ETRS89
<i>Rights</i>	EEA standard re-use policy: unless otherwise indicated, re-use of content on the EEA website for commercial or non-commercial purposes is permitted free of charge, provided that the source is acknowledged ( <a href="http://www.eea.europa.eu/legal/copyright">http://www.eea.europa.eu/legal/copyright</a> ). Copyright holder: Directorate-General Enterprise and Industry (DG-ENTR), Directorate-General for Regional Policy.

### Open Street Map

<i>Dataset Purpose</i>	Up-to-date city street, road and footway network
<i>Source</i>	© OpenStreetMap contributors
<i>Temporal Coverage</i>	Mar 2015
<i>Entity type</i>	Vector: ESRI Polyline
<i>Resolution/ Precision</i>	0.01m

<i>Geographic coverage</i>	Extent of chosen cities: Gothenburg, Valencia, Vilnius
<i>CRS</i>	WGS84
<i>Rights</i>	OpenStreetMap is open data, licensed under the Open Data Commons Open Database License (ODbL). User is free to copy, distribute, transmit and adapt the data, as long as it credits OpenStreetMap and its contributors. If altered or built upon our data, one may distribute the result only under the same licence.

#### **Landsat 8 satellite imagery**

<i>Dataset Purpose</i>	Identification of green areas within a city
<i>Source</i>	USGS (U.S. Geological Survey)
<i>Temporal Coverage</i>	Gothenburg: 2014/09/03; Valencia: 2014/09/01; Vilnius: 2014/09/08
<i>Entity type</i>	Raster: GeoTIFF
<i>Resolution/ Precision</i>	30m
<i>Geographic coverage</i>	Extent of chosen Landsat 8 scenes
<i>CRS</i>	WGS84
<i>Rights</i>	There are no restrictions on the use of data received from the U.S. Geological Survey's Earth Resources Observation and Science (EROS) Center or NASA's Land Processes Distributed Active Archive Center (LP DAAC), unless expressly identified prior to or at the time of receipt.

#### **European Urban Audit city boundary and population data**

<i>Dataset Purpose</i>	City boundaries and population
<i>Source</i>	GISCO - Eurostat (European Commission)
<i>Temporal Coverage</i>	2004 (boundaries), 2012 (population)
<i>Entity type</i>	Vector: ESRI Polygon
<i>Resolution/ Precision</i>	1:3 Million
<i>Geographic coverage</i>	LUZ (Larger Urban Zones) of European Union countries
<i>CRS</i>	WGS84
<i>Rights</i>	<p>The GISCO Urban Audit 2004 geographical dataset contains geometry derived from EuroBoundary Map 2004 (EBM2004) developed by EuroGeographics. Where polygon geometry from the GISCO Urban Audit 2004 geographical dataset is used in any printed or electronic publication, the copyright on the source data set should be acknowledged in the legend of the map and in the introductory page of the publication, following the European Commission data source acknowledgement.</p> <p>The copyright notice on the legend of the map and the introductory part of the publication is as follows: © EuroGeographics for the administrative boundaries. For publications in languages other than English, the translation of the appropriate source and copyright notices in the language of the publication shall be used.</p>





# 1 . INTRODUCTION

---

The problem of discovering and understanding various scopes of city dynamics has been of high interest throughout the history of urbanism. Yet the advances over the last decades in data collection and distribution methods, open GIS and spatial analysis technologies and open geospatial standards has allowed to analyse, monitor, evaluate and interpret many of the urban phenomena. With the help of recent technologies, it is possible to build urban data models that are detailed, large scale, comparative and reproducible.

A powerful tool enabled by the recent technologies is crowdsourcing defined by Estellés-Arolas (2012) as a type of participative online activity in which the crowd should participate in the undertaking of the task, bringing their work, money, knowledge or experience, always entailing mutual benefit. The user receives the satisfaction of economic need, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer obtains and utilizes to their advantage what the user has brought to the venture.

An interesting part of crowd-sourced GPS data can be obtained from mobile sports tracking applications such as Nokia Sports Tracker, Endomondo, Strava, Garmin, MapMyRun, etc. These data are constantly generated worldwide by tracking various smart device users who are willing to record their spatio-temporal activity while doing sports activities involving overcoming of long distances in outdoor environment. The later has an advantage towards the other data collection methods as the available data is extremely big, provided voluntarily by a large number of users, public and thus does not raise privacy issues, always available up-to-date, can be easily used for comparison since the method of collection is the same world-wide, and finally, it is constantly growing.

As these GPS tracks closely relate to the urban space network used solely for recreational purposes, data can be properly used not only in order to analyse and monitor such network layer but also act as a ground truth for explaining and predicting the phenomena. Generally, two distinct cases need to be concerned while delving deeper into the matter: where within a city recreational activities happen and where they do not. If an appropriate data model was developed to predict these cases, respective actions could be taken in order to encourage or discourage one or the other.

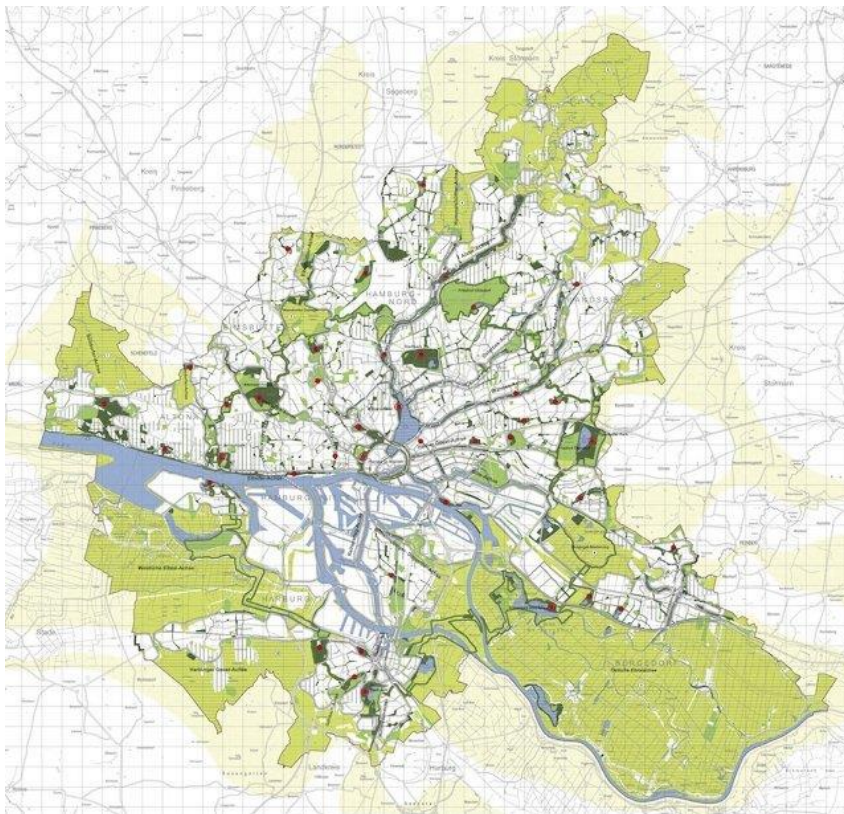
The research aims to develop a comprehensible method to observe, analyse and possibly predict the actual situation in various European cities based on the mobile sports tracking application data. Correspondingly, it aims to find out whether correlation exists between theoretically estimated potential of a space to be used for recreation and the actual observed values. The Runability Index, introduced by this research, intends to be used for creating knowledge-based models of cities recreational systems, developing existing and creating new strategies for urban recreation systems and their regeneration and assessment of structural elements, investigating the relationship between various characteristics of the built environment and the recreational physical activity, etc.

The Master Thesis Report is organised as following: the initial chapters explain problem statement and relevance, goal and scope of the research and the followed research methodology. The third chapter delineates related research and qualifies how this particular research distinguishes from the others. The next one explains the framework for acquiring mobile sports tracking application data and overviews data's characteristics, which is followed by the informed choice of case study cities. The following chapter reports on the construction of an urban space network. Fifth chapter explains the determination of an actual recreational values and the sixth one explains how the Runability Index is formed. Finally, the conclusions are drawn, which are followed by the discussion and recommendations.



## 1.1 PROBLEM STATEMENT AND RELEVANCE

The general goal of urban design and planning disciplines is to make urban areas functional, attractive and sustainable. This includes a wide scope of urban issues among which supporting physical activity, sunlight, clean air, safety and social connectedness. The later mentioned values define what is recently being addressed as the “livability of a city”. Architects, urbanists and legislators have already begun creating guidelines and regulations to approach the goals of livability. E.g. New York City has announced “Active Design Guidelines: Promoting Physical Activity and Health in Design” (NYC DDC, 2010). Similarly, Hamburg has decided to unite its green areas into the “Green Network Plan” (Figure 1), which aims to diminish the need for vehicles over the next few decades at the same time providing sufficient amount and quality of easily accessible recreational spaces. London City has held High Line for London competition, which asked for green infrastructure proposals (newlondonlandscape.org, 2012). Thus, it is safe to assume that many other cities will follow this precedent (Vinnitskaya, 2013).

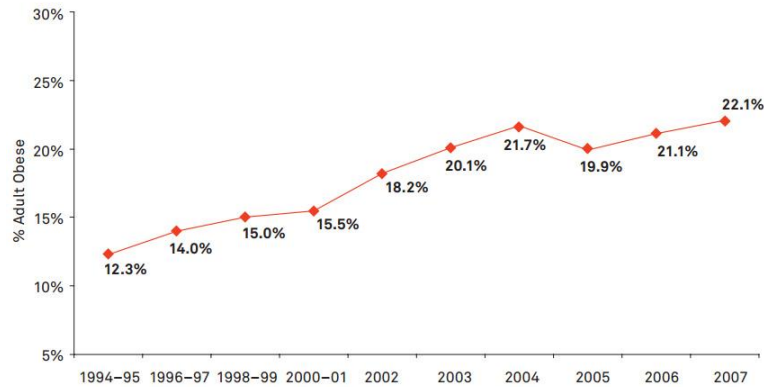


**Figure 1. Suppositional master plan of the Hamburg's "Green Network Plan" (Rasuli et al., 2010)**

It is considered that in annual balance of time dedicated for resting, daily relaxation makes  $\frac{1}{3}$  of total time amount (NRPA, 2014), which may be spent at home or near home at the local recreational area. Therefore, qualitative recreational areas throughout the city not only increases satisfaction levels of citizens but also the value of nearby real estate because proximity of an attractive area influences desirability of residential units. Hamburg Institute of International Economics has noted that “even if the green network occupies space that is needed for housing and businesses, on the other hand, it brings economic advantages because it attracts highly educated and competent people to the city” (Braw, 2013).

Furthermore, is it also a matter of sustainability, as ensuring qualitative recreational network accessible for all citizens, can become one of the means to limit the urban sprawl. Tratsaert (1998) indicates the lack of accessible public green space as the main reason for householders moving away from the cities in Belgium. Similarly, the lack of recreational urban spaces attracts more people to the indoor activity centres, where sustainability is heavily violated by people using cars to reach fitness halls where they run on a treadmill.

Lastly and most importantly, implementing plans for green recreational networks addresses another tender issue of the modern cities: today in Europe, more than 50% of people are overweight, in US - 2 out of 3, and 1 being obese with the situation constantly getting worse throughout the last decades (Figure 2). Obesity as well as other chronic diseases such as heart disease, diabetes and some cancers all have roots in the sedentary lifestyles partly caused by the earlier design trends that have contributed to declining physical activity (Vinnitskaya, 2012).



**Figure 2. Adults with self-reported obesity in USA, 2004-2007 (NYC DDC, 2010)**

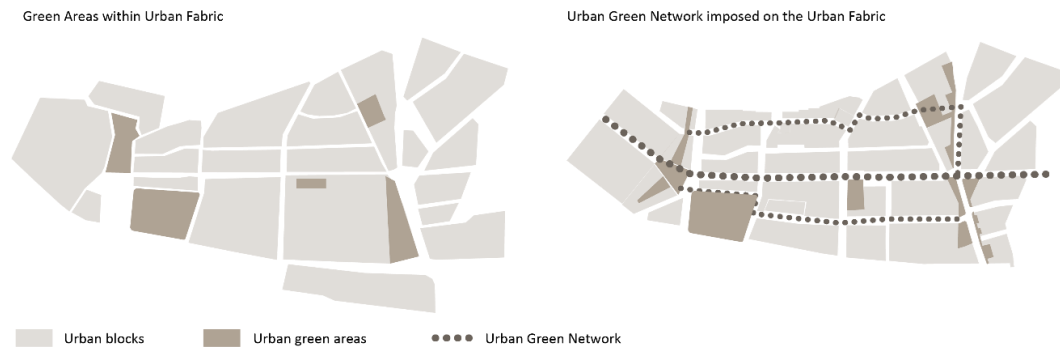
Obesity exacts a toll not only on people's health but also on country's economy, in the form of rising health care and disability costs and declining productivity and workforce availability. More far-reaching economic consequences include fuel expenses and costs from insurance, disability, absenteeism, and decreased productivity for the business sector. This economic burden is only anticipated to grow: if the current rate of increase in obesity continues, the total health care costs attributable to obesity are anticipated to double every decade (Wang et al., 2008).

As noted by Lindsay (2010) built environments give cues as to how to inhabit them and have tremendous effects, sometimes subconscious, on people's lifestyles: "If someone dies walking on a sunny day next to an intensive traffic road before making it home, the official cause of death would be "heat stroke," and not a lack of sidewalks and shade trees. If one were to be hit by a truck, it would be considered an auto fatality, not the victim of a lack of transportation route alternatives." Rather than just telling people to go the gym public health sectors must work with architects, urban designers, and planners to create opportunities and encouraging environments for exercise in daily life (NYC DDC, 2010). Thus it must be learned to think of space not as the background to human activity, but as an intrinsic aspect of everything human beings do (Hillier, 1996).

This conjunction between urban complexity, cognition, planning and design indicates a potential for the emergence of a new field of study in which urbanism is not an external intervention in a spontaneous and complex urban process, but rather integral element in its dynamics (Portugali, 2011). The importance of scientific knowledge in these matters is that when trying to improve a situation in urban planning, there should be evident knowledge of both present situation and consequences of any action. Specifically, a better understanding of how leisure-based physical activity is incorporated into people's daily activities enables development of effective policy interventions to promote physically active lifestyles in different built-environment contexts. In addition, because recreational travels make up a substantial part of mobility, studies of recreational activity patterns contribute to activity-based travel demand modelling (Lin et al., 2015).

On the other hand, the green spaces within a city are also considered luxurious as they take up expensive land estate and require constant maintenance costs from the city. That is where the notion of "urban green network" comes in (Figure 3) – it defines a network throughout the city and its outskirts which

connects recreational spaces such as sports facilities, gardens, parks and squares with routes designed primarily for cyclists and pedestrians (Quirk, 2014). As with any other type of infrastructure, to be most effective green infrastructure needs to be part of a shared vision that is planned, designed and managed. When it is, it creates multifunctional landscapes capable of delivering social, economic and environmental benefits simultaneously (newlondonlandscape.org, 2012).



**Figure 3. Concept scheme of imposing Urban Green Network on the existing Urban Fabric.**

Although there is a consensus on the benefits of such shifts, there is little scientific knowledge on how cities are being used for recreational purposes and even harder it is to monitor any changes in people behaviour after changes are made (Sener and Bhat 2012). Previous studies of recreational activities within a city rely on manual data collection by directly observing chosen locations during certain short periods and registering observed physical activities on registration sheets (Floyd, 2008), asking adult residents in surrounding areas of a park to complete 7-day physical activity logs that include the location of their activities (Kaczynski, 2008), comparing recipients places of residence with their physical activity registered by accelerometers (Cohen, 2006) or even using an annual telephone survey (Lopez, 2004).

All the previously mentioned methods are performed by intensive human labour and can only be applied on relatively small-scale measurements. Therefore neither constant re-monitoring, nor data collection of the complete urban area become impossible. This is where the crowd-sourced data from mobile sports tracking applications has an enormous potential for solving the problem. GPS tracking offers to urban dynamics studies a new layer, which provides insight in processes and actual movement of people doing sports in outdoor environment.

However, even though appropriate data is being constantly collected and even published online, it is still a challenge to make it usable for the desired purpose. Mobile sports tracking application data can be considered as Volunteered Geographic Information (VGI). Similarly, to VGI, data is collected by a worldwide user community gratuitously and for personal motives such as social reward, enhanced personal reputation, competition or simply personal interest. Since the motivation of application provider is not collection of geographical features but personal satisfaction of a user, emphasis of application design lies on an individual workout rather than on data collection.

In fact, nor the data is available in a single click; neither application providers are willing to prepare an interface for ready-made free access due to privacy issues. Furthermore, the data is so big that it cannot be analysed as it is and special selection, filtering and aggregation procedures need to be applied. Finally, the GPS data is just a sequence of points in Euclidian space, which need to be mapped and analysed in a network space, while the appropriate space network for any of the cities is a non-trivial issue itself.

All of the previously mentioned issues form the strong basis for the scope and goal of this research, which are stated in the following chapter.

## 1.2 GOAL AND SCOPE OF THE RESEARCH

The general goal of the graduation thesis research is to develop a reusable framework for mobile sports tracking application data acquisition, management, processing and analysis, as basis for using it as a ground truth for the validation and calibration of data model to estimate or predict the potential recreational usage of an urban network space. In other words, the goal is being able to define where recreational activities happen, where they do not and finally, use this knowledge to give an indication to every space of how likely it is that the space *is* or *will be* used for recreation.

The research focuses on the European cities where sufficient amount of data is obtainable and where this kind of research is currently relevant. It aims to achieve the following goals:

- Develop an efficient method for obtaining crowd-sourced GPS data from a chosen mobile sports tracking application;
- Choose an efficient method for big data management, storage, querying and spatial operations;
- Choose and apply proper methods for the processing of GPS tracks – removing of blundering points and snapping to the actual urban space network;
- Choose and process other relevant datasets needed for the analysis and ensure that the chosen datasets are available and comparable for the various locations throughout Europe;
- Develop a method for constructing an urban space network, which would be relevant for the analysis of human recreational travels, relying on datasets, which are equally available for all European cities and enable further analysis and calculations based on the network.
- Develop a model of the potential recreational usage based on related research and available data sources.
- Develop a method of using the acquired workout data as a ground truth for the calibration and validation of developed data model to estimate the potential recreational usage of an urban network space;



### 1.3 RESEARCH QUESTION

The main question to be answered by this research is:

*How can GPS data, generated by mobile sports tracking applications, be used to assess, analyse and model the recreational usage of an urban space network?*

With the sub questions of:

- *How can GPS data of a mobile sports tracking application be used for the research?*
  - *Which mobile sports tracking application is the most suitable for the experiment, considering its popularity in Europe, ease of data access and amount of available data?*
  - *Which cities are the most suitable as case studies for the research?*
  - *What method should be used for obtaining and storing and querying huge amounts of public GPS tracks of running and walking workouts?*
  - *How does the data need to be processed by means of filtering and snapping?*
- *How should an urban space network be constructed to be relevant and facilitating for further analysis?*
  - *What are considered the boundaries of various cities?*
  - *Which datasets should be used as a base for an urban space network and how can these datasets be combined into a single one?*
  - *How detailed or generalised the network needs to be in order to serve the analysis purpose?*
- *How can a potential recreational usage of an urban space be modelled?*
  - *What theoretical framework can be used for modelling potential recreational usage?*
  - *Which other datasets are relevant for the analysis and are they equally available for various countries in Europe?*
  - *How can various measures and quality indicators from different datasets be aggregated into a single value?*
- *How can GPS data of a mobile sports tracking application be used to assess the data model for estimating potential of recreational usage?*
  - *How can the GPS data be aggregated and converted into a single measure of actual recreational usage?*
  - *How can the actual and potential values be compared and what are the results of this comparison?*
  - *What kind of knowledge can be provided from the comparison of the two?*

## 1.4 METHODOLOGY

Complex systems theory has been developed in the early eighties and ever since has been stimulating urban dynamic's modelling. The main initiative has come from mathematicians and physicists applying their models of emerging properties to biological and even social sciences. As stated by Sanders (2008), "complexity arises in situations where an increasing number of independent variables begin interacting in interdependent and unpredictable ways". As suggested by Klaasen (2003), reality can be conceived as a complex of systems, which is simplified by each individual into a single and unique system that needs to be explored. However, the data and tools that are used for the exploration induce the reduction even further, making it only possible to explore a certain system by its subsystem. The subsystem is further simplified based on intentional considerations and scientific assumptions and then represented as a model (Figure 4). "A simplified representation of the real world conceived as system is a model of reality only if it has a certain structural kinship with that reality, and only if the model is the result of a conscious interpretation of that reality" (Klaasen, 2003).

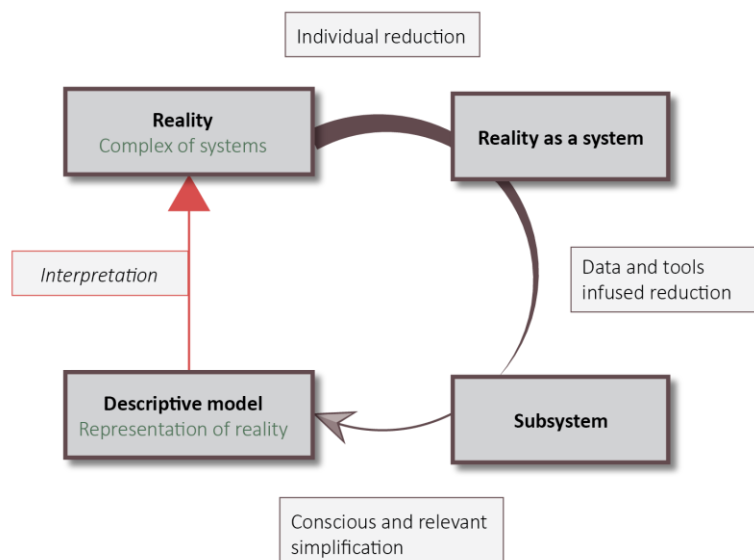


Figure 4. Concept scheme of relation between reality and model (based on Klaasen (2003)).

The new complex systems theory focuses on "the emergence of properties at a macro-level as resulting from the interactions between individual behaviour at a micro-level" (Bretagnolle et al., 2006). There are two different approaches to understand the individual behaviour at a micro level – one is through multi agent-based simulation, another – through monitoring the actual behaviour and dealing with real-world data. However, given the limits of simplification that every model has, no assertions can be made based on a model other than within these limitations: what is put into a model determines what comes out and assertions based on a model may be made only within the field of applicability. The same counts for the interpretation of the model - predictions or other statements drawn from such a formal system depict the real world only insofar as the model is valid.

The system that the research project is aiming to explore is an intersection of urban space network system and an urban recreational system. Where urban space network system means city as a physical network of paths and routes, where spaces are nodes and intersections are links. Urban recreational systems stand for places, objects and processes, which are characterized by recreational functions and activities. Structural components of urban recreational system consist of network of public urban spaces used for physical, social, intellectual and spiritual recreation and connections between them.

The extracted subsystem comprises solely from the urban path (or space) network used for recreational activities, which, in particular, are “physically active recreational travels” (Bhat & Lockwood, 2004) – voluntarily tracked and published running and walking activities. (Cycling activities are not explored due to the assumption based on personal experience that cycling for transportation is rarely tracked by a mobile application and cycling for sports commonly happens outside urbanised environment). The subsystem is dictated by the data available for the research. The reduction of reality is contracted into one-year span and network space. Furthermore, the available data comprises only the users of a single sports tracking application.

Simplification of the subsystem into a correct, representative and comprehensible descriptive model is exactly the scope of the research carried on. Urban studies are mostly empirical and try to explain “what will probably be the case”, however the practical approach aims to explain, “what is or can be the case” building descriptive and exploratory models (Klaasen, 2003). Since the goal of this research is merely to provide a method to look into the data and use for both purposes – practical and empirical, the research itself is more of a practical science. Broadly speaking, the application of the research is a societal one, but the result is a technically developed scientific instrument, which can eventually contribute to the advance of empirical scientific knowledge.

This research does not supply the justification of hypotheses and theories as much as supply feedback to a heuristic approach in the context of data available for the research. Thus, the interpretation of provided descriptive model must rely on all the characteristics infused by all different steps of the reality simplification process. The following chapter explains the particular research method in detail.

## 1.5 METHOD DESCRIPTION

The research follows the framework shown in Figure 5 and consists of 8 main levels:

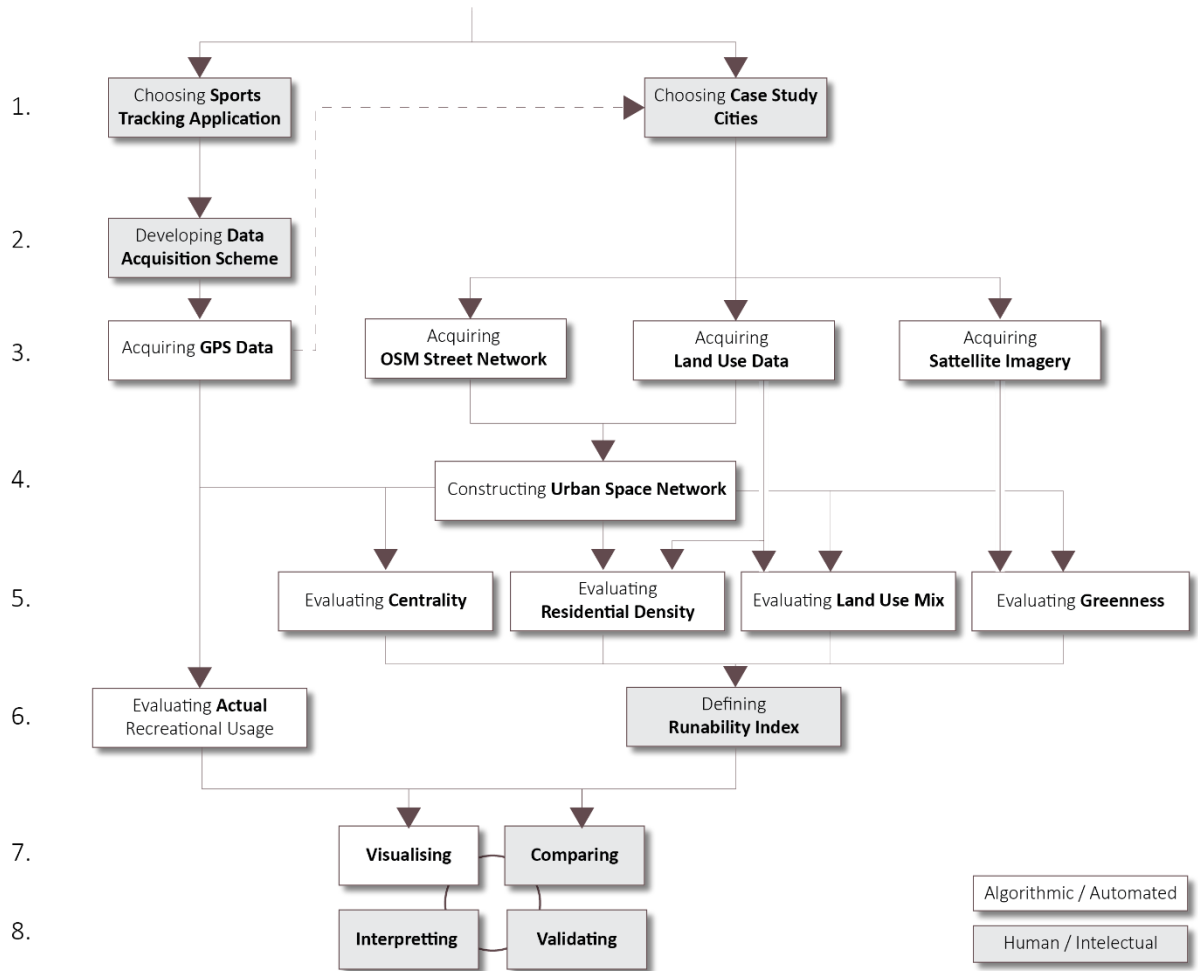


Figure 5. Scheme of the followed research method.

### 0 Preparation for the research

Preparation for the research includes specification of the research problems, goals and questions and collection of relevant information.

Literature research is also part of the preparation stage. Once focused on a formulated research question and its sub questions, it aims to identify, appraise, select and synthesize all relevant high quality research evidence and arguments. Literature research is augmented during the whole research process. It mainly covers topics of spatial qualities related to physical activity, spatial activity analysis based on GPS tracking and urban network analysis.

All software and plugins used during the research are free or open source: PostgreSQL with Postgis, Qgis, Boost library, etc. Scripts are written in Ruby and Python programming languages. All scripts are available at <https://github.com/rusne/RunabilityIndex>.

## **1 Choosing case studies**

Single sports tracking application is chosen as a case study based on a number of criteria such as ease of data access, popularity in Europe, etc. (as explained in chapter '3.1 Choice of Application'). When a sample part of data is collected, the case study cities can be chosen based on their population, number of chosen sports tracking application users and differences of geographical position (chapter '3.4 Case Study Cities'). All case study cities need to be of a similar size in population, however distinguished by their urban structures, climate and even cultural characteristics. This requirement is set in order to ensure that the developed method is equally suitable throughout Europe and not only valid in certain region with particular properties. This requirement also eliminates the possibility of getting good results accidentally.

## **2 Developing acquisition scheme for mobile sports tracking application data**

Afterwards the data acquisition framework is developed in order to make use of the sports tracking application data. A special framework of data acquisition has to be prepared considering aspects such as efficiency, process automating and relevance to the project. Data acquisition framework includes testing characteristics of a chosen application and developing data extraction algorithm from the application website, which contains public GPS tracks of running and walking activities. Also data loading into the database, determining necessary time intervals and automating the process until the highest extent possible.

## **3 Acquiring data**

Due to the extremely extensive data download time from the sports tracking application website (2 months). Algorithms are constructed using sample data and later collected data is fed into the algorithm at the final stage of the project. For assurance purposes, occasional tests have been made with newly collected data to ensure that developed algorithms are scalable.

Supplementing datasets are acquired for two purposes – first, they are necessary as a base to construct the underlying urban space network, second, additional datasets are needed to model the potential recreational usage of the network. These datasets are chosen based on literature research and availability of datasets for all three case study cities. Open Street Map data is used as a constituent of an urban space network, supplemented by the Urban Audit Land Use data, which is also used to determine the land use mix and residential density values. Satellite imagery is used in order to evaluate greenness of urban spaces. Detailed description of the used datasets can be found in 'List of used datasets'.

## **4 Constructing urban space network**

After the GPS trajectories are obtained and cleaned from any blundering points, it is already possible to visualise them as 2-dimensional lines. However, the visualisation is still crumbly and does not allow a comprehensible recognition of spatial patterns. Generally, there are two ways of dealing with GPS trajectories. One is using locational and geometric similarity measures to replace similar moving patterns by the representative ones (Shaw et al., 2008; Kobayashi and Miller, 2012). The other one is clustering trajectories on a network, i.e. snapping them to an existing network space and assigning as attributes to the network edges (Abraham and Lal, 2010; Sadahiro et al., 2013).

The later method is more applicable in case of this research, since the collected GPS trajectories were obtained from numerous different GPS devices and thus have different accuracy and precision characteristics obstructing a reliable clustering. Furthermore, the aggregated trajectories are later to be used for analysis in comparison with the estimated potential recreational usage in order to define the Runability Index. Therefore, trajectories need to be matched with a base network, which would later act as an intermediary between different datasets. Construction of such urban space network is a bottleneck of the whole project since no further process can be done before satisfying result is achieved.

Urban space network has a number of requirements, such as relevance to the analysed phenomena, low level of detail, simplicity, topological validity, time-compliance, etc. The network is constructed by utilising Open Street Map data complemented with the land use data from Urban Audit as explained in chapter '4 Urban Space Network'

## **5 Defining indicators of potential recreational usage**

The indicators of the data model for estimating the potential recreational usage are chosen based on the well-known measure of walkability (explained and referenced in chapter '2 Related research') which consists of multiple measures, among which most common – greenness, land use mix, street connectivity and residential density. Even though Walkability Index aims to explain physically active travels for transportation purpose, it is used as a base theory for, what is defined by this research, as a Runability Index. It is an index of space potential to be used for active recreational travels.

The Runability Index is constructed based on the greenness, land use mix, residential density and normalised angular choice, which is a measure of network centrality. Measures are taken to adapt the characteristics of active travels for recreational purposes, rather than transportation purposes and the changes are tested towards the actual data. The procedure of model construction is explicitly explained in chapter '6 Runability Index'.

## **6 Evaluating actual recreational usage and defining the Runability Index**

Evaluating actual potential usage involves GPS data processing (i.e. managing, filtering and snapping to an urban space network), assigning it as attributes to urban space network edges and finally mapping the values into what can be called the likeliness of space to be used for active recreational travels, based on the collected data. The procedure is explicitly explained in chapter '5 Actual Recreational Usage'.

Potential recreational usage, once has its indicators, needs to be aggregated into a single value which would correspond to the actual recreational usage, i.e. likeliness of a space to be used for recreation. Different combinations of diverse runability indicators' variants are tested in order to find out which of them have the best correlation with the actual usage and can define the Runability Index. It is explicitly explained in the chapter '6 Runability Index'

## **7 Visualisation and comparison**

There are two levels of visualising project results – one is through visualising resultant maps for each of the case study cities to enable visual inspection and analysis, and the other is through graphs, which explain statistics of data collection, characteristics of the network and correlation values between the actual and estimated potential values. The comparison between the actual and estimated potential recreational usage is done using statistical analysis methods to assess if correlation between the two datasets exists and until what extent the defined model of Runability Index is valid. Research results should be visualised in graphically clear, comprehensible and appealing way which can be as well interpreted for urban design and planning related matters.

## **8 Result validation and interpretation**

Part of the validation is finding out what differences exist between the mismatch of the actual and potential values, since this knowledge can be used for the recommendations for future research. It is also assessed if any correspondence exists between the actual recreational usage of the network generated by this research and by different sports tracking applications. The analysis and validation results are finally interpreted and used for drawing the conclusion, discussing the results and preparing recommendations for future research.





## 2. RELATED RESEARCH

The place of this research in a wider research context is shown in Figure 6. In particular, this research focuses on utilising crowd-sourced data (both from mobile sports tracking applications and Open Street Map) and GIS-based urban network analysis methods in order to explore the patterns of physically active urban travels. More precisely, its emphasis lies on recreational travel purposes as in contrast to much more commonly explored transport-based travel purposes.

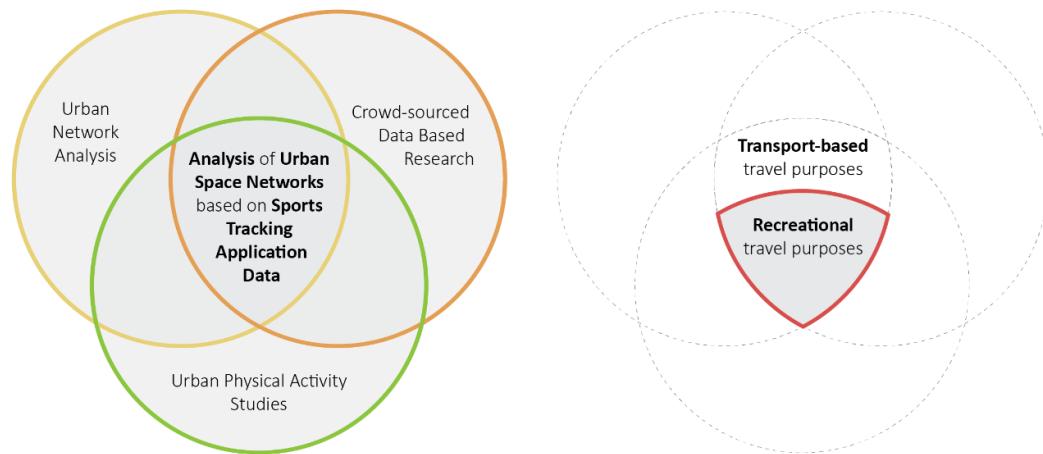


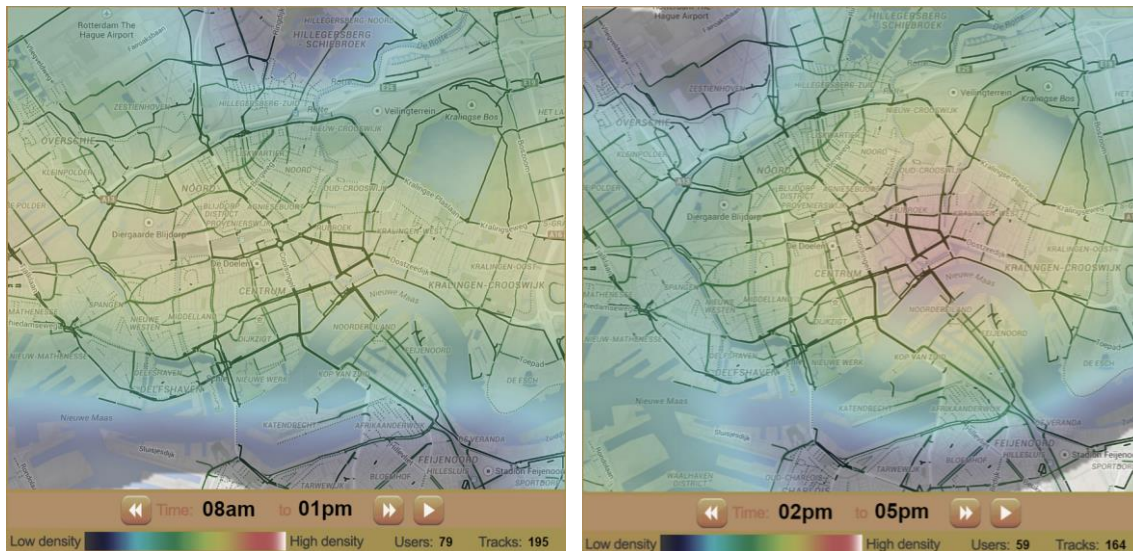
Figure 6. Position of this research in wider research context.

### 2.1 URBAN ANALYSIS BASED ON GPS DATA

GPS is a GNSS system, which allows spatial position of a special receiver to be calculated, based on the reception of satellite signals. While originally a military project, GPS has become a widely deployed and useful tool for cartography, navigation, geotagging, surveying and many others. Strong points of GPS include its ability to deliver results in any weather conditions, it is free to use for anybody with a GPS receiver, easy and cheap to integrate into other technologies and available anywhere on the Earth where there is an unobstructed line of sight to four or more GPS satellites. Furthermore, saving a travelled route into a track log makes the technology useful to collect spatiotemporal data, observe and measure activities of people (Shoval, 2008). One of its major weaknesses is the capability for positioning in dense urban environment and indoors. Thus as every other method, the use of GPS devices for urban analysis has both advantages and challenges (Van der Spek et al., 2009).

Van der Spek et al. (2009; 2013; 2014) have carried out a research called ‘Spatial Metro’ which aims to explain pedestrians’ behaviour in various cities by deploying GPS tracking system supplemented with questionnaires. During the data acquisition phase, different user groups were tracked by distributing GPS devices and asking respondents to carry them for a specific period time ranging from several hours to a full week. The study has found various spatiotemporal patterns of both residents and tourists behaviour in a city of question. The experiment has been positively evaluated and GPS has been evaluated as a promising research instrument for urban studies. The noted shortages of the experiment were as following: short tracking time, complicated repeating of the experiment, too small sample size and all deficiencies associated with GPS inaccuracies and malfunction in urban environments. While the later remains a question of technology, the first mentioned issues are well tackled by using crowd-sourced GPS tracks from various mobile applications.

Piorkowski (2009) has pioneered in using mobile sports tracking application data for analytic purposes. However, he rather aimed on enhancing location privacy and designing better context-aware services than on analysing patterns of urban mobility. Ferrari & Mamei (2011) have used GPS data from Nokia Sports Tracker application to identify both the areas of a city most used for a given sports activity and the temporal routines of people using these areas (Figure 7). These results were displayed via mobile applications to offer a wealth of context-aware services.



**Figure 7. Areas of a city most used for running during a given period of day. Comparison between morning and afternoon in Rotterdam (Ferrari & Mamei, 2011).**

In another work, Ferrari & Mamei (2013) use this data to highlight cultural and climate-related differences among cities, describe the human-centred geography of the city with regard to sports activities and show that such areas can be partitioned among groups of users to highlight differences in the routine behaviour of various demographic/social communities. The most recent relevant research performed by Oksanen et al. (2013) aims to extract frequently used routes from massive public workout data by developing scalable algorithms for spatiotemporal clustering of the workout trajectories in order to define the most popular routes as a suggestion for application users. The main difference between their work and this research is that in case of this research analysis is performed in a network space in contrast to Euclidian space. In addition, all of these works have been using data from Nokia Sports Tracking application, which is exceptionally popular in Finland, however, has less users in the other countries.

All the previously mentioned studies form a relevant background for integrating network analysis methods with crowd-sourced GPS data in order to achieve previously explained goals.

## 2.2 PHYSICAL ACTIVITY IN URBAN ENVIRONMENTS

The concept of Urban Green Network, also called Urban Green Infrastructure, has been introduced to upgrade urban green space systems in a coherent manner of planning. Originally, it is considered to comprise of all natural, semi-natural and artificial networks of ecological systems within urban areas, at all spatial scales. However, later this concept has merged with the emerging need for the network of ecological, sustainable and physical-activity-enabled urban space network, which would be used for both transportation and recreation (Sandstrom, 2002).

Recently there has been an increased interest for human behaviour models in health promotion research. The theoretical background of such models emphasize the fact that a human behaviour is not only affected by one's individual characteristics, but also by the surrounding environment (Sallis et al., 2008). The relationship between urban development and physical activity has become the most heavily

investigated subject in urban planning, generating numerous studies in the past decades (Gebel et al., 2007). Likewise, a number of space features have been indicated which enhance likelihood of physical activity in that space, which will be most probably - but not necessary, - performed by the residents of surrounding urban blocks:

- according to that Brown et al. (2009), van Dyck et al. (2010), Yamada et al. (2012) and a number of other studies, increased balance of land-use mix is the most important attribute of neighbourhoods that encourage walking;
- Pucher et al. (2008) have found that obesity rates are lower among the first-world-countries in such countries as Germany, Denmark, and the Netherlands, where transport infrastructure gives priority to cycling and walking than to motorized transport;
- numerous studies have linked proximity of parks and other recreational facilities to higher levels of physical activity (Henderson, 2005; Cohen et al., 2006, 2007; Floyd et al., 2008; Maroko et al., 2009). More stringently, it is recommended that “people living in towns and cities should have an accessible natural green space less than 300m from home” (English Nature, 2005; Harrison et al., 1995; Handley et al., 2003; Wray et al., 2005).

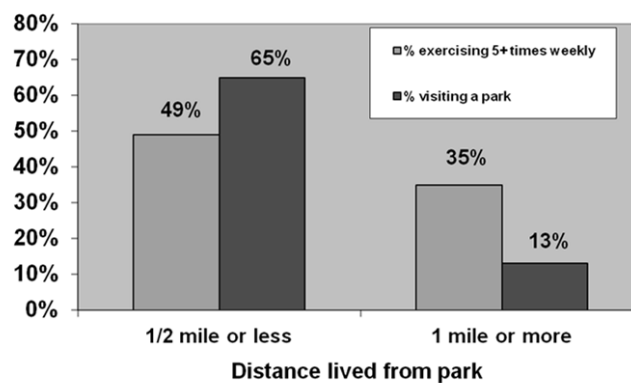


Figure 8. Relation between distances lived from park and number of respondents involved in physical activity (Cohen, 2007).

- Powell et al. (2006) and Gordon-Larsen et al. (2006) have found correlation between neighbourhood's socioeconomic status and physical activity, since residents of higher status regions were more likely be engaged in physical activities than those of low-income;
- Cervero & Kockelman (1997) and Ewing (2001; 2009) have identified the five “D” variables that influence relationship between urban design and physical activity: density, diversity, design, destination accessibility and distance to transit. Density describes concentration of both jobs and people in an urban zone. Diversity indicates number, variety, and balance of land uses in the area. Design includes the characteristics of a street network and streetscape. Destination accessibility reflects the ease of travel to recreational zone. Distance to transit measures the average distance from a place of residence/work to the nearest rail station or bus stop.
- Another range of studies (Gauvim et al., 2005; Giles-Corti et al., 2005; Troped et al., 2010; Koohsari et al., 2013c; etc.) have examined street-specific features which influence people's preferences for route related recreational activities. Their findings are namely as following:
  - Presence of shadow, absence of windiness and other microclimatic characteristics;
  - Low traffic and air pollution;
  - Presence of resting places (benches, potable water fountains, etc.);
  - Non-concrete surface cover;
  - Accessibility from many streets/public spaces;
  - Sufficient width of a sidewalk;
  - Safety (considering crime and traffic accidents).

To highlight the importance of the built environment effects on physical activity, the term Neighbourhood Walkability has emerged which may be conceptualized as “the degree to which the attributes of the built environment may promote/inhibit walking and physical activity behaviour” (Leslie et al., 2007). Walkability studies have provided evidence through statistical analysis that certain features of built environment correlate with individuals’ walking behaviour. However, even though Cutumisu (2011) has defined walkability as “a property of the built environment that measures the conduciveness to walking, running, biking, rollerblading, or other activities that involve non-motorized movement”, Choi (2013) has noticed that walking for pleasure or recreational reasons shows observably different behaviour from transportation base active travels. Recreational walking trips are generally conducted with less purposeful attitude, at a slower speed, with more flexibility between moving and sojourning and not always directed by the shortest distance route, as in the case of utilitarian trips.

Furthermore, as noticed by Troped et al. (2010) previously mentioned studies have mostly and almost only measured physical activity using an analytic approach that assumes such activity to occur within a designated area in residential neighbourhoods. The area of focus with respect to where physical activity occurred may contribute to the dilution of the observed associations, resulting in blindness for true associations or an underestimation of the bias.

## **2.3 URBAN NETWORK ANALYSIS**

A number of distinct network analysis methods have been already developed in order to explore city dynamics and mobility patterns. Lynch (1960) has been the first to define that urban fabric is perceived as a combination of five main elements: paths, edges (boundaries), districts, nodes (intersections) and landmarks. Euler most probably has been the first one to apply graph theory on a spatial network in his famous paper on the seven bridges of Königsberg published in 1736, where each land mass is considered as a node of a planar graph and the bridges connecting them are the edges. Later the theory of Space Syntax (Hillier & Hanson, 1984) proposed that the continuous urban space is configured in to a set of discrete interconnected units (sub-spaces) and the topological connectivity and accessibility of these units steer human movement within them.

Space Syntax as the methodology of urban network analysis shows that human movement is predictable for individual streets; that is, well-connected streets tend to attract more traffic than less-connected streets. Until now it raises growing evidence of the correlation between the centrality indicators of urban spaces and phenomena as diverse as crime rates, pollution, pedestrian and vehicular flows, commerce vitality and way-finding (Nagar & Tawfik, 2007) capacity.

As mentioned in the previous chapter, a number of researches have explored that there is definite relation between the configuration of urban street network and street walkability. Moreover, a number of researches integrate the urban network configuration with additional attributes to explain better the desired phenomena. Since in some cases the effect of configuration is weaker than in others and such factors as transport nodes, land use, infrastructural elements, major attractors or generators, aesthetic features and etc. are needed to improve the prediction of walkability (Hillier, 2005; Gauvin et al., 2005; Gebel et al., 2007; etc.) Hillier & Stutz (2005) suggest that Walkability Index should be based on the multiple regression analysis which is a statistical methodology to analyse data empirically and that way determine the impact of each separate factor as an input to the movement model.

Besides the network centrality measures (Hillier & Hanson, 1984), walkability has been explained through such methods as weighted Page Rank algorithm (Jiang, 2009), random walk or eigenvector centrality (Blanchard & Volchenkov, 2008), etc.

## 3. DATA ACQUISITION

### 3.1 CHOICE OF APPLICATION

Seven mobile sports tracking applications have been compared according to the number of available tracks, popularity and data access possibilities in order to choose the most appropriate one for the project (Table 1). Considering also the other aspects, data access possibility was the most important one. Since the sports tracking applications cannot provide direct access to their databases due to privacy issues, some of them provide a service of buying anonymised data for analysis purposes while others display public workouts on dedicated websites. In that case, users are aware and content with publicly displayed (but not distributed) personal data.

After considering a number of applications, Endomondo was chosen due to its popularity rate and relatively convenient data access.

Application name	Tracks available	Most popular in	Data access
Endomondo	>450 000 000	Scandinavia/Europe	Through website /reading html
Strava	> 2 500 000 000	USA	Paid access
MapMyRun	-	USA/North America	Only user suggested routes
RunMeter	-	Only iPhone users	Only own tracks
(Nokia) Sports Tracker	>5 000 000	Finland	Through website /reading html
Runtastic	-	Germany/Europe	Only user suggested routes
RunKeeper	1 500 000	USA	Through website /reading html

Table 1. Comparison of mobile sports tracking applications.

Endomondo is a mobile sports tracking application available for a big range of mobile devices, including Android, iPhone, Windows Phone, Blackberry and Garmin watches. The application allows tracking individual movements' trajectories of a chosen sports activity, e.g. running, walking, cycling, kayaking and many others, which can be measured by distance. Tracking is based on GPS receiver of a phone and therefore is based on the characteristics of each individual device.

Each workout is tracked locally on a device and afterwards uploaded online. Another option of adding a workout is manually drawing a route on a map and specifying the duration of training. Every workout is publicly available to be viewed on "[www.endomondo.com/workouts/](http://www.endomondo.com/workouts/) + workout ID" unless specified to be private by a user (Figure 9).

Every workout is a JSON object embedded into an HTML code of a page. Additional to the GPS trajectory, there is number of other attributes available: type of the workout (running, walking, etc.), date and time, user name (id), distance, duration, average speed, maximum speed, burnt calories, hydration, altitude and weather information. Some of these attributes are derived directly and only from the GPS trajectory (distance, duration, altitude and speed) while the others require personal information added by the user (user name, burnt calories and hydration). Date and time are obtained from the mobile device and dependant on its indications and weather conditions are taken from online services. A user can choose to make any of these attributes private, edit the values or delete the workout permanently at any time. All data is saved on a server and later on synchronised with the application (endomondo.com, 2015).

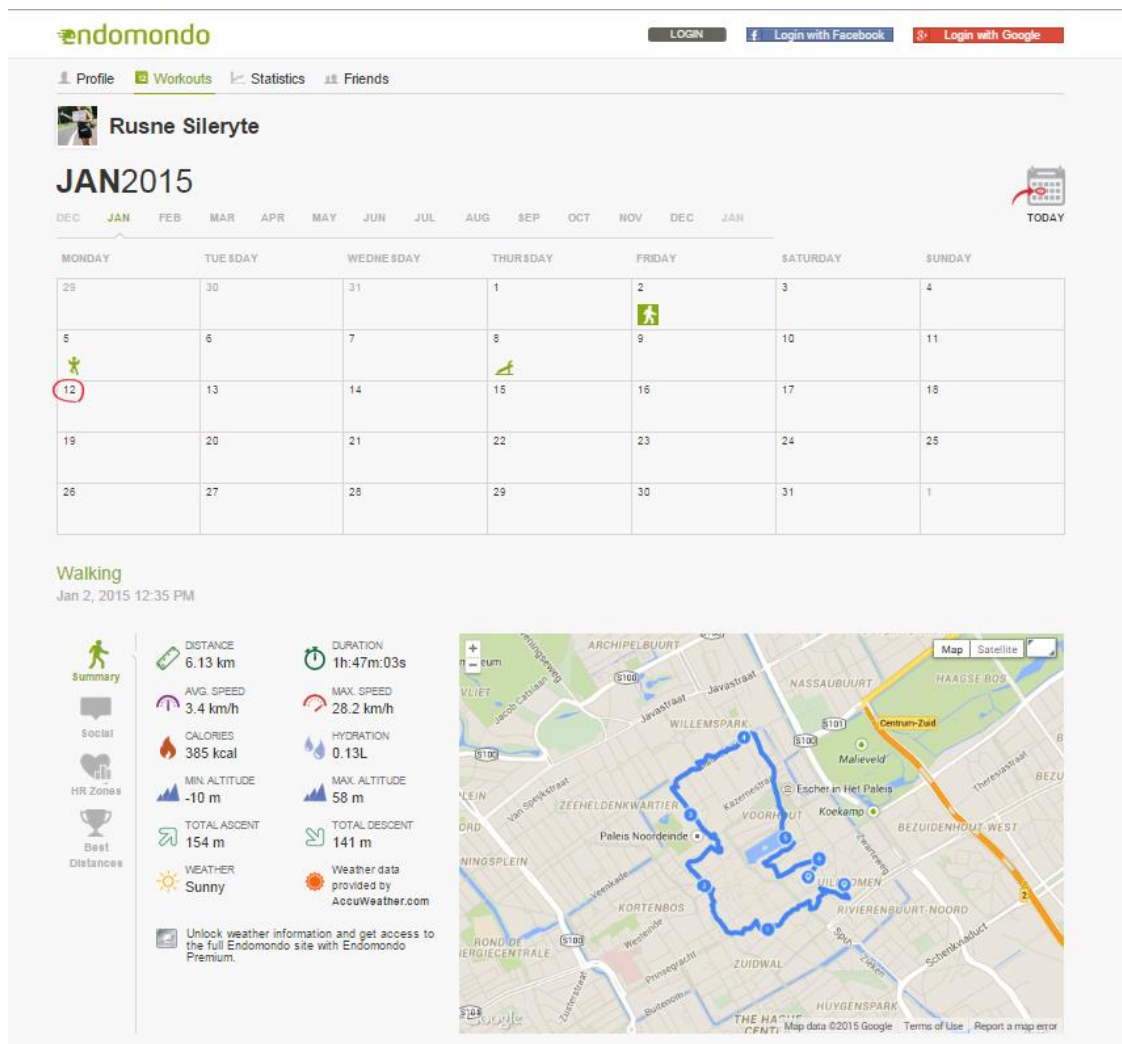


Figure 9. Public workout and its attributes accessible at <https://www.endomondo.com/workouts/453360586>



### 3.2 SCHEME OF DATA ACQUISITION

Data samples are timed every 8 days starting from the May of 2014 until the May of 2015, aiming to have sufficient data throughout the full year and a variety of weekdays as well as occasional public holidays (Figure 10). The reason for collecting data in chunks per day is the time needed for data download since one day's data requires one day and night of downloading time as explained later.

2014												2015					
May	June	July	August	September	October	November	December	January	February	March	April						
1	1	1	1	1	1	1	1	1	1	1	1						INTENDED
2	2	2	2	2	2	2	2	2	2	2	2						COLLECTED
3	3	3	3	3	3	3	3	3	3	3	3						IN PROGRESS
4	4	4	4	4	4	4	4	4	4	4	4						
5	5	5	5	5	5	5	5	5	5	5	5						
6	6	6	6	6	6	6	6	6	6	6	6						
7	7	7	7	7	7	7	7	7	7	7	7						
8	8	8	8	8	8	8	8	8	8	8	8						
9	9	9	9	9	9	9	9	9	9	9	9						
10	10	10	10	10	10	10	10	10	10	10	10						
11	11	11	11	11	11	11	11	11	11	11	11						
12	12	12	12	12	12	12	12	12	12	12	12						
13	13	13	13	13	13	13	13	13	13	13	13						
14	14	14	14	14	14	14	14	14	14	14	14						
15	15	15	15	15	15	15	15	15	15	15	15						
16	16	16	16	16	16	16	16	16	16	16	16						
17	17	17	17	17	17	17	17	17	17	17	17						
18	18	18	18	18	18	18	18	18	18	18	18						
19	19	19	19	19	19	19	19	19	19	19	19						
20	20	20	20	20	20	20	20	20	20	20	20						
21	21	21	21	21	21	21	21	21	21	21	21						
22	22	22	22	22	22	22	22	22	22	22	22						
23	23	23	23	23	23	23	23	23	23	23	23						
24	24	24	24	24	24	24	24	24	24	24	24						
25	25	25	25	25	25	25	25	25	25	25	25						
26	26	26	26	26	26	26	26	26	26	26	26						
27	27	27	27	27	27	27	27	27	27	27	27						
28	28	28	28	28	28	28	28	28	28	28	28						
29	29	29	29	29	29	29	29	29	29	29	29						
30	30	30	30	30	30	30	30	30	30	30	30						
31			31			31		31			31						

Figure 10. Data collection log on Google Spreadsheets.

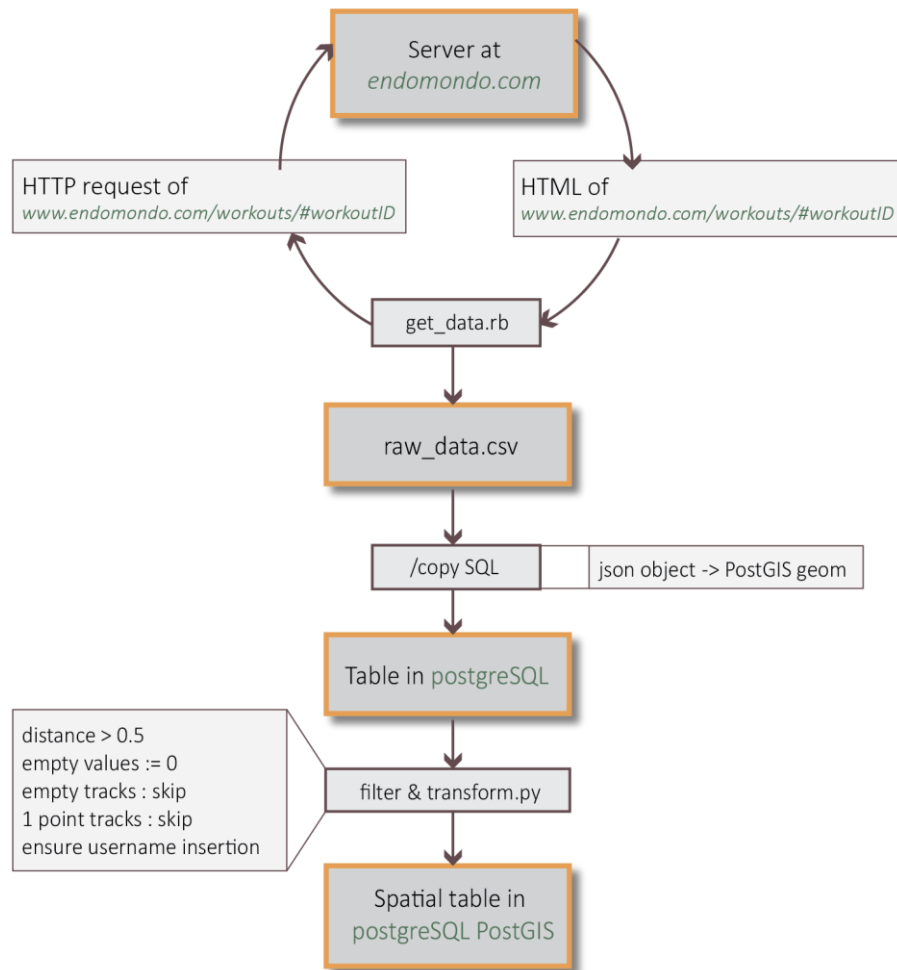
The desired data must include only those entries, which represent running and walking activities within the bounding box of Europe in period between the 05-2014 and actual time of data acquisition. They must have a valid GPS track (covered distance is longer than at least 0,5 km and not longer than 42km (length of a marathon is taken as reference maximum length), duration does not exceed 12h, there is more than 1 GPS point available and points have coordinates). The extracted attributes are track ID, username, distance and duration of the workout.

However, the data available online cannot be queried directly and can be accessed only workout by workout by changing the id number in a particular URL. Fortunately, the ids are sorted by the workout upload time starting from 0 and currently reaching almost half a billion. Therefore it can be expected, that workouts which ids are similar, will have similar time of execution. The only mismatches in this case happen firstly, because of time differences between countries, secondly because of delayed uploads, e.g. from Garmin watches or other devices.

There have been a few attempts to design efficient and consistent data acquisition framework. A sample tutorial in Barsukov's (2014) blog has been used as a basis for the framework design. However, due to different project purposes, the proposed schema could not be used and had to be moderately adapted.

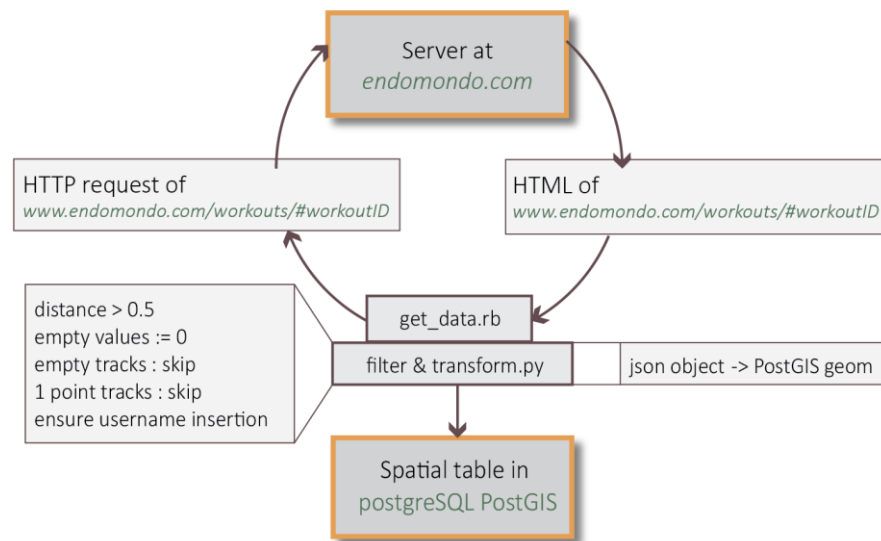
The primary designed and tested scheme of data acquisition is shown in Figure 11. A Ruby script *get\_data.rb* sends an HTTP request to the server for a workout with a chosen ID and either gets a negative response (in case the workout is listed as private or it has been deleted) or a positive response and an HTML code of a page. If HTML code is received, the algorithm continues exploring the data. If GPS trajectory is available and listed as 'Running' or 'Walking', the required fields are read into a string and output into a *raw\_data.csv* file, which is later read from command line with \copy command into PostgreSQL database transforming the GPS track from a json object into a PostGIS geometry linestring. After the data is loaded into a database, another Python code selects only the tracks which have distance longer than 0,5 km, filters out invalid data points based on the difference between timestamps and

distance between GPS points, assigns 0 values to missing fields and ensures that username is read and inserted correctly.



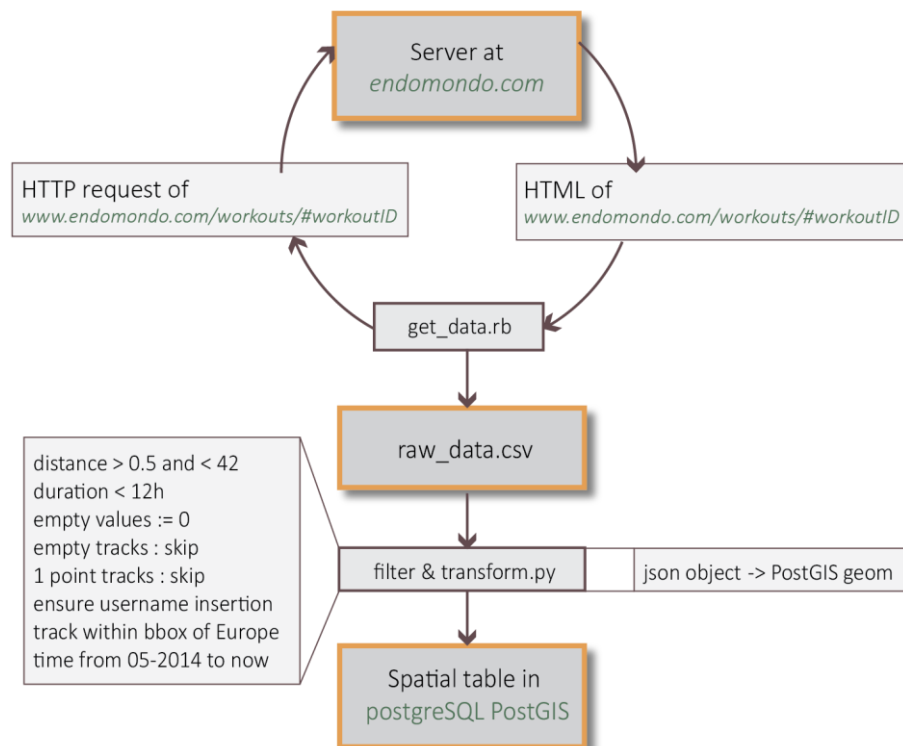
**Figure 11. Initial scheme of data acquisition framework.**

Even if the data acquisition is successful, there are still a few points of improvement that need to be made due to inefficiency, which happens because of redundant data copying into the database. A second and more proper approach is shown in Figure 13, where ruby and python codes are bounded to perform data collection consecutively, filtering, transformation and insertion into the database.



**Figure 13. Alternative scheme of data acquisition framework where ruby and python codes are bounded consecutively to perform data collection, filtering, transformation and insertion into the database.**

However, the major limitation of this approach is extended runtime since both HTTP response and data insertion into the database take up a significant amount of time, which becomes a crucial issue when dealing with huge amounts of data. Therefore another framework has been designed to tackle the time of execution issue (Figure 12). The final approach first collects raw data into a .csv file which is then read and processed by python code in parallel to the data collection of another chunk of data. Furthermore, the filtering of outliers is omitted in order to save execution time and performed only for the chosen workouts in case study cities as explained in chapter ‘5.2 Filtering GPS trajectories’



**Figure 12. Final scheme of data acquisition framework, where data collection can be run parallel to the data filtering, transformation and insertion into a database.**

The visualisation of data acquisition result from a single day (January 1<sup>st</sup>, 2015) can be seen in Figure 14. It can be clearly seen that the chosen sports tracking application is indeed the most suitable for conducting analysis in Europe and that data entries cover all of its urbanised part. The statistics of all successfully acquired data can be seen in Table 2. The unexpected jumps of GPS coordinates, which appear as straight lines in the Figure 14, are explained in chapter '3.3.3 Incorrect GPS tracks'

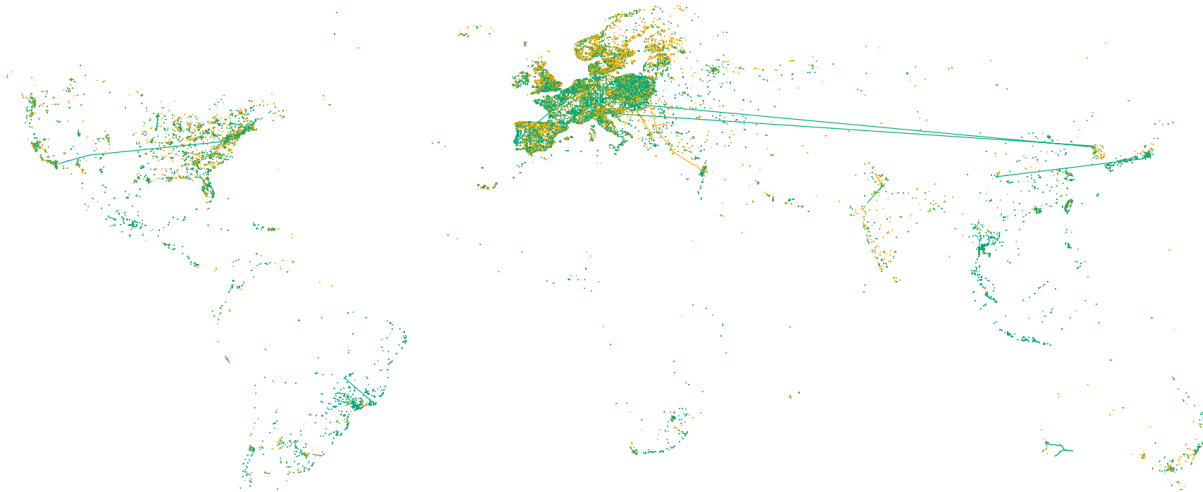


Figure 14. Visualisation of GPS tracks collected on a single day (January 1<sup>st</sup>, 2015), walking in orange; running in green.

<b>Time needed for requesting, reading and writing data in .csv files:</b>	1248h (2 months)
<b>Number of http requests sent:</b>	15 600 000
<b>Number of lines written in a .csv file:</b>	5 964 008 (38%)
<b>Number of data entries that passed filter:</b>	3 610 735 (23%)
<b>Number of distinct users registered:</b>	911 588
<b>Average number of tracks per user:</b>	4
<b>Total timespan covered</b>	01 04 2014 – 01 04 2015
<b>Number of tracks with unidentified user</b>	383 402

Table 2. Statistics of data acquired from a mobile sports tracking application (endomondo).

The efficiency of data acquisition process can be verified through distribution of collected data in a target timespan of one year. As it can be seen from Figure 15, the acquired data is evenly distributed throughout the year with occasional peaks and troughs, however, without any apparent kinks. This verifies data's time-wise validity.

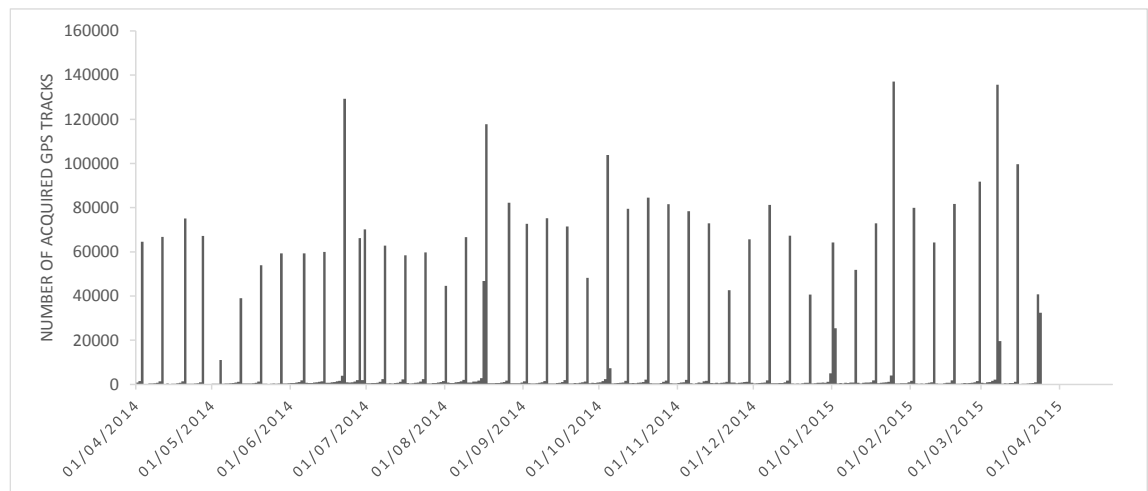


Figure 15. Distribution of acquired data in a timespan of the target year.

### 3.3 DATA LIMITATIONS

During the process of data acquisition, a number of data defects and limitations have been noticed. Some of them have been removed while filtering the data; others require more advanced data processing algorithms, while the general limitations cannot be addressed and have to be taken into account while assessing data validity.

#### 3.3.1 APPLICATION USERS

Generally, the collected data is limited to only one sports tracking application and therefore only its users are tracked which limits the set of tracked individuals to those who have certain characteristics such as knowledge of a foreign language, possession of a smart phone, ability to use the application and, of course, a consent to be tracked.

Moreover, the general statistics of application users address another important issue of the ratio between male and female users being 75 and 25 % respectively.

Therefore, it must be acknowledged that acquired data represents only a certain subset of all recreational travels conducted in a city, which may cause some related bias to the research results.

#### 3.3.2 INCORRECT OR IRRELEVANT ATTRIBUTES

One of the most common incorrect attributes is a timestamp of a GPS track. As explained previously, the timestamp in obtained data entry does not refer to the data upload time, however, at the time when the track was recorded by the device where Endomondo application is installed. This means, that sometimes workouts are uploaded a few hours or even days later than their actual time of execution. Another reason for the timestamp inconsistency is a possibility to import workouts from another application into Endomondo diary (e.g. from Strava, MapMyRun, etc.). In that case, the workouts still receive an id according to the upload time. That explains why some of the workouts date back as far as to 2009.

Another problematic field is the username. Username in this research is used as code of user identification and its content does not have influence for the project, however problems appear due to the syntax issues.

Since there are no regulations about creating a username, some users create names that confuse algorithms and therefore have to be treated in a special way. A few examples of such names are:

**'Name O'Name'** – quote mark in a name.

**'Ramon/Name'** – n/ sign which is a notation used for the newline.

**'Name Surname <http://www.facebook.com/namesurname>'** - double slash used for database escape.

**'Name Surname'** – use of tab instead of space between the names (confuses reading tab delimited files).

**'เปรี้ยว เยี่ยวราด'** - symbols unrecognised by UTF-8 encoding.

**'Anonymous'** – commonly used name which indicates different persons, however is recognised as the same user. This accounts for 10% of all GPS tracks which have passed the filter.

**''** – empty name string.

### 3.3.3 INCORRECT GPS TRACKS

Even after removing 0 length, single point or missing geometry tracks, the remaining tracks need to be checked for the “jumping point” which cause the trajectories to become unrealistically long as shown in Figure 16.



Figure 16. GPS track jump (in yellow); left: jump due to user error; right: incorrect GPS positioning.

There are multiple reasons for a mobile GPS receiver to get wrong coordinates. One of them is lack of satellites in sight. In order to calculate location to within about 2 meters GPS devices typically need to receive signals from at least 7 or 8 satellites. With fewer than 4 satellites, many GPS receivers are unable to produce any location estimates, and will report either lost signal or a very rough estimate of location which can differ from the real location even in thousands of kilometres (Modsching et al., 2006).

Buildings, trees, tunnels, mountains, and even the human body can prevent GPS signals from the satellites reaching the receiver. Second, when GPS receivers initiate with what is called 'cold start' (no priori data

saved) they need time to acquire first signals from satellites and might produce wrong indications at the beginning. Finally, when signals from the GPS satellites bounce off buildings, the GPS receiver can be confused by the extra time the signal took to reach it. In these cases, sudden large errors in position can be observed (Figure 17). Alternatively, such error can occur due to the user fault if a user disables GPS in one's device without switching off the application and continues using it later from a different location, the application understands that as a continuation of a previous workout with a huge jump (Figure 16). Filtering of such cases is explicitly explained in chapter '5.2 Filtering GPS trajectories'.

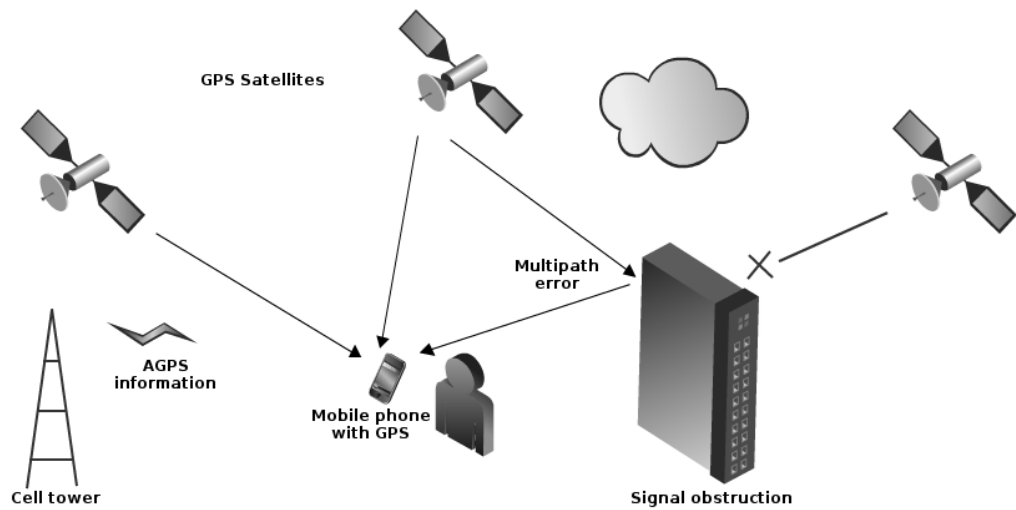


Figure 17. Reasons for incorrect GPS positioning (Anderson, 2012).



### 3.4 CASE STUDY CITIES

Three cities have been chosen as case studies for the further research based on the following requirements:

- A city must belong to one of the countries in the European Union. This requirement is set because of necessary availability of datasets provided by the Eurostat.
- A city must be categorised as an “extra-large” by the Eurostat Urban Audit project (EUA, 2007) (i.e. its population at the city core must be 0,5 -1 million inhabitants).
- All the chosen cities must have a similar rate of the chosen sports tracking application users.
- The three cities should have distinct characteristics in terms of climate, urban development, amount and distribution of urban recreational spaces.

In order to facilitate the choice of case study cities Query 1 has been developed. The query is making use of the database table with already collected GPS workout data and the Urban Audit dataset tables, which hold the geometry of cities’ administrative boundaries and their population data.

```
SELECT ua.city_name, ua.nmb AS users, ua_population.population,
ua.nmb/ua_population.population::float AS ratio
FROM (
    SELECT COUNT(DISTINCT(routes_eu.name)) AS nmb, ua_codes.city_name
    FROM routes_eu, ua_codes, cities_eu
    WHERE TRIM( trailing '1' from ua_codes.city_id ) = cities_eu.urau_city_
    AND ST_Intersects(routes_eu.route, cities_eu.geom)
    GROUP BY ua_codes.city_name )AS ua
LEFT JOIN ua_population ON (ua.city_name = ua_population.name)
WHERE ua_population.population BETWEEN 500000 AND 1000000
ORDER BY ratio DESC;
```

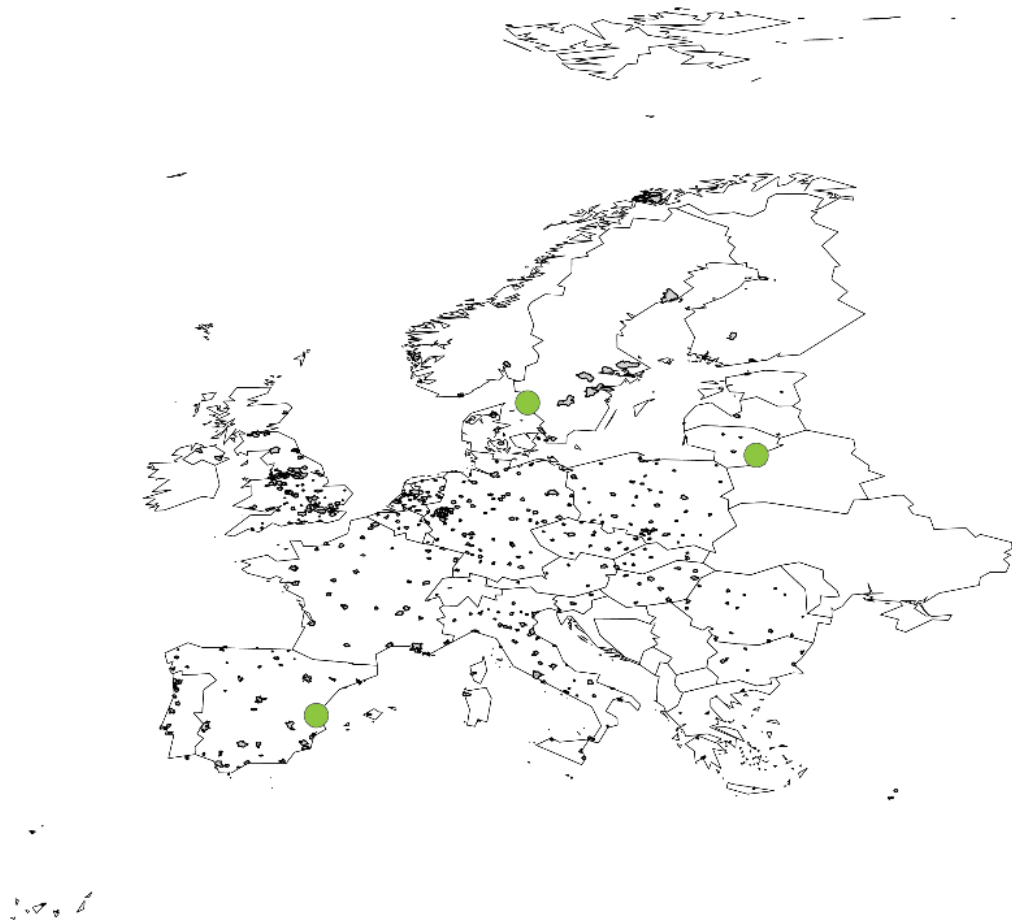
**Query 1. Selection of extra-large European cities ordered by the ratio between a number of distinct application users already found in the database and city’s population.**

The purpose of the query is to retrieve these European cities which population ranges between 500 000 and 1 000 000 inhabitants and find a ratio between the number of distinct users, which have at least one workout trajectory lying inside the boundary of a city and the population of the city. The first 20 results of the query (ordered descending by the ratio) can be seen in Figure 18. At the time of the querying database contains 1 214 014 valid GPS tracks acquired from both summer and wintertime.

	city_name character varying	users bigint	population integer	ratio double precision
1	København	9204	559440	0.016452166452166451
2	Poznan	3737	570778	0.0065472039917446011
3	Wrocław	3458	636268	0.0054348167753210915
4	Oslo	3201	623966	0.0051300872162906314
5	Kraków	3150	759137	0.0041494486502436315
6	Rīga	2555	735241	0.0034750510376869623
7	Vilnius	1721	558165	0.0030833176569652343
8	Valencia	1829	814208	0.0022463547398207831
9	Göteborg	1110	520374	0.0021330812069780582
10	Zaragoza	1394	679624	0.0020511341565336125
11	Stockholm	1666	864324	0.0019275179215201706
12	Sevilla	1288	704414	0.001828470189405662
13	Málaga	882	568507	0.0015514320843894622
14	Lisboa	587	549998	0.0010672766082785755
15	's-Gravenhage	403	502055	0.00080270089930386114
16	Rotterdam	467	616260	0.00075779703371953394
17	Amsterdam	592	790110	0.0007492627608813963
18	Glasgow	420	594100	0.00070695169163440495
19	Manchester	351	506800	0.00069258089976322024
20	Leeds	527	787700	0.00066903643519106257

**Figure 18. List of extra-large European cities ordered by the ratio between spotted application users and city’s population. The chosen case study cities are marked in green.**

Copenhagen has been rejected as a suitable case study due to its extremely high ratio, which is not comparable to any other European city. Oslo was rejected as a city outside the European Union. Poznan, Wroclaw, Krakow, Riga and Vilnius all share similar geographical location as well as cultural background; therefore, only one of them was chosen based on the personal preferences of the author of the research. The chosen one, Vilnius (Lithuania), has a similar ratio (2-3 spotted users per 1000 inhabitants) with Valencia (Spain) and Gothenburg (Sweden), which makes these a suitable trio for the case study. All three cities are distinct for their geographical location (Figure 19), thus climate, urban development and state of recreational network.



**Figure 19. Location of the chosen case study cities marked in green: Vilnius (Lithuania), Valencia (Spain) and Gothenburg (Sweden).**

It must be noted that at the end data acquisition process the ratios between a number of distinct users and city's population have changed due to additionally collected data and the ratio did not stay as similar as at the time of decision, however the ratios are still reasonable for continuing with these case studies. The final ratios can be found in cities' descriptions.

### 3.4.1 VILNIUS

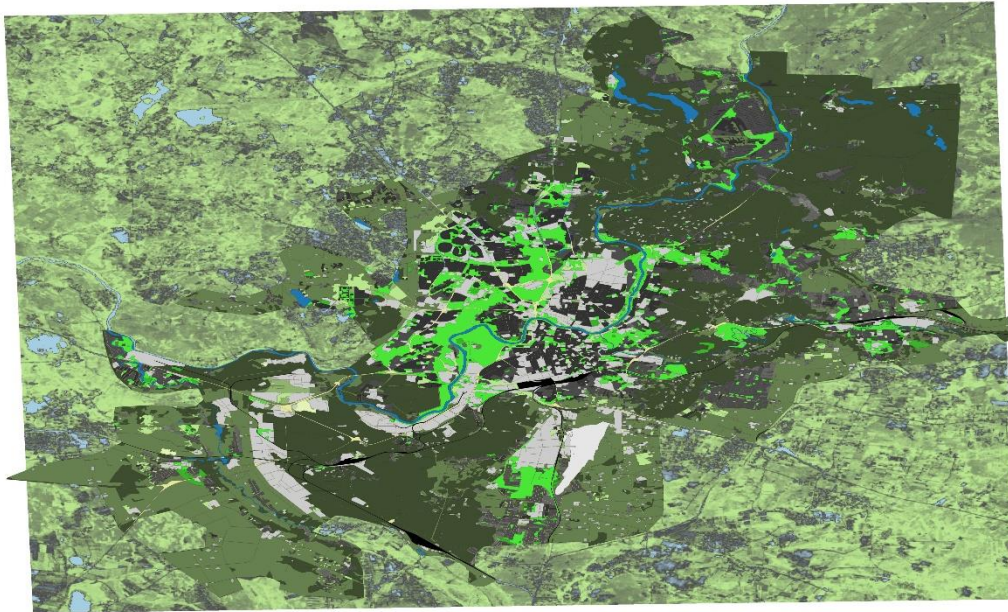


Figure 20. Urban Audit land-use data of Vilnius city boundary placed on top of the Landsat 8 image NDVI band.

Vilnius is the capital of Lithuania and the second biggest city of the Baltic States. According to Köppen climate classification Dfb (2006) the climate of Vilnius is humid continental with temperatures ranging from  $-30^{\circ}\text{C}$  in wintertime to  $+30^{\circ}\text{C}$  in summer time, making its outside public spaces seasonally operational. Approximately 20.2% of the official Vilnius area is developed with the remaining 43,9% of green space and 2,1% of water (Vilnius.lt, 2015)

According to the study of transformations in spatial expression of urban recreational functions in post-soviet spaces conducted by Urbonaitė (2013) Vilnius, as well as most of the other soviet cities, have developed extensively during the post-war period, yet the density of accommodation is half of other Western European cities of a comparable size. It is full of public greenery, which nevertheless is not properly used since the recreational efficiency depends not only on quantity of recreational resources. Rapid urban transformations in post-war Vilnius caused disproportional spread of recreational functions and generally ineffective recreational system.

Field research (Urbonaitė, 2013) showed that “the recreational potential of Vilnius city is heterogenic: it consists not only of regulated open recreational spaces, but also includes urban public places, territories of common use, recreational connections, and indoor recreational facilities. Recreational mobility in Vilnius city is very fragmented. Although the suburbs are rich in recreational resources, the accessibility and means of communication to the inner city are not properly distributed”. Research results have revealed that informal recreational possibilities have the biggest potential to bring an even distribution of the recreational functions, especially those of active outdoor recreational activities.

Population	No. of application users	No. of GPS tracks	Ratio users/population
558 165	3950	10 165	0.007

### 3.4.2 VALENCIA

Valencia is the capital of the autonomous community of Valencia and the third largest city in Spain. It stands on the banks of the Turia river, located on the western part of the Mediterranean Sea. Valencia has a subtropical Mediterranean climate (Köppen climate classification Dfb, 2006) with very mild winters and long warm to hot summers which makes city's outdoor spaces widely used throughout the year.

Gomez et al. (2010) have addressed a lack of green zones within a city as a shortcoming that Valencia must overcome in upcoming years. The World Health Organisation has for some years been proposing the need for this provision not to be under 9 m<sup>2</sup> per inhabitant, or more recently even from 10–15 m<sup>2</sup> of green zone per inhabitant, however in case of Valencia the current provision of reachable green zones is 5 to 34 m<sup>2</sup> per inhabitant (Valencia City Council, 2011). Moreover, one of the current municipality urban design and planning policies includes city's ambition to provide a high standard sport-related outdoor spaces approachable and accessible to everyone, as well as gradually spread throughout the city (Valle, 2013).



Figure 21. Urban Audit land-use data of Valencia city boundary placed on top of the Landsat 8 image NDVI band.

A particular of feature of Valencia's urban recreational network is the 'Garden of the Turia' (Jardí del Túria/Jardín del Turia) which runs through the very heart of the city on the diverted river bed. After a catastrophic flood in 1957, Turia river was divided in two at the western city limits, so that the water has been diverted southwards along a new course that skirts the city, before running into the Mediterranean sea. The old riverbed is now a vivid green park that enables cyclists and pedestrians to traverse much of the city without using roads while many bridges carry the heavy traffic overhead across the park. The Turia river makes up around  $\frac{2}{3}$  of the city's total green area (Valencia City Council, 2011).

Population	No. of application users	No. of GPS tracks	Ratio users/population
814 218	3583	9 443	0.005



### 3.4.3 GOTHENBURG

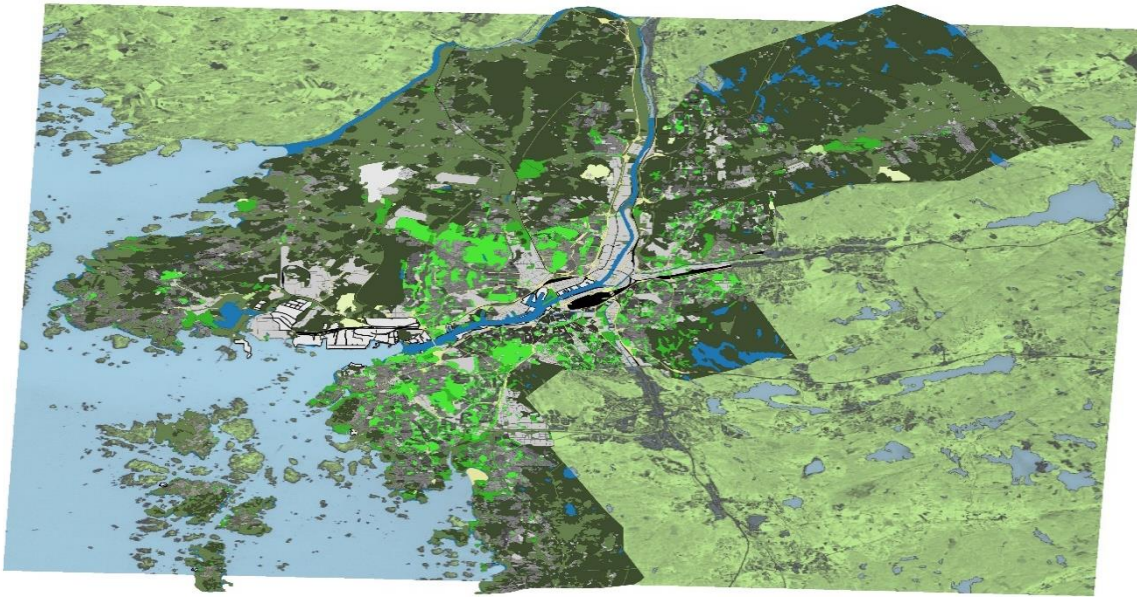


Figure 22. Urban Audit land-use data of Gothenburg city boundary placed on top of the Landsat 8 image NDVI band.

Gothenburg (Goteborg) is the second largest city in Sweden situated by the Kattegat bay of the Baltic and North seas, at the mouth of the river Göta. Due to the Gulf Stream the city has a mild climate with a lot of rain. Despite its northern location, temperatures are quite mild throughout the year and much warmer compared to places in similar latitude. The climate is categorised as oceanic by the Köppen climate classification Dfb (2006). Frequent precipitation and strong winds diminishes the use of outdoor spaces, however, seasonal use is not of a big influence.

The Göta river valley divides Gothenburg into distinctive eastern and a western parts. To the east is a forest area with elevations ranging between 50 to 150 m above sea level while the western part is rather dominated by flat and open low areas, interspersed with higher areas at an average of 60 m above sea level. The river valley landscape is responsible for many natural green spaces within the city. According to Alm et al. (2002), Gothenburg has no shortage of green areas, regarding both quantity and quality, since all central, almost all residential and even most of the industrial zones lie within 5min walk from a green open space. These green areas vary from local parks and squares with well-kept lawns to extensive botanical gardens, from tree covered granite hillocks to utter wilderness towards the city boundaries.

Even if the quantity and quality of outdoor recreational areas are in a rather great condition, the baseline study of urban challenges in Gothenburg conducted by Cullberg et al. (2014) highlights a number of related issues. First of all, Gothenburg is a rather socially segregated city featuring great differences between southern and western against northern-eastern districts. Moreover, the transportation is very sensitive to disruptions and has insufficient carrying capacity, therefore a high interest lies into increasing both public transport use and the share of travel by bike which is low compared to similar neighboring cities, e.g. Stockholm and Oslo. Finally, a number of brownfield port areas in central Gothenburg is aimed to be redeveloped into sustainable central districts.

Population	No. of application users	No. of GPS tracks	Ratio users/population
520 374	2 195	6 720	0.004

## 4 . URBAN SPACE NETWORK

---

### 4.1 NETWORK DEFINITION

Urban space network is an underlying network of navigable urban spaces, which differs from the urban street or road network in that it includes spaces, which are navigable for humans but not necessarily for vehicles. In one of the most recent overviews of research concerning the influence of built environment on walking behaviour, Lin et al. (2015) has noticed that most accessibility studies use only street and road, thus vehicle-oriented, networks in their analyses. This may result in inadequacies in the description and prediction of non-motorised travel and induce discussions about the reliability of research results. In order to trace the paths chosen by pedestrians, a base network should incorporate formal and informal paths, including sidewalks, laneways, plazas, pedestrian bridges and park paths that are frequently used for pedestrian transit. The missing pedestrian paths in the street network can greatly increase connectivity of separate locations in the real world (Chin et al., 2008).

So far, there is no consensus on how the urban space network needs to look like and no state of the art method for its generation. In case of Space Syntax (Hillier & Hanson, 1984) spaces are represented by axial lines - the visibility lines across open spaces. As has been pointed out by Jiang et al. (2009) "axial lines are used to represent directions of uninterrupted movement and visibility, so they represent the longest visibility lines in two-dimensional urban spaces". While for a long time axial maps used to be drawn by hand, later Liu & Jiang (2010) developed a set of algorithms, which calculate axial lines automatically given the polygons of spaces as an input. The drawback of this approach is that the boundaries of spaces need to be mapped beforehand and such data is rarely available on a large scale.

Lately the road network provided by the OpenStreetMap is often being chosen to form the backbone of urban networks because of its universal coverage and standard defined for all modes of transport. Moreover, due to its open access nature and volunteered contribution, OSM can have a very good level of completeness and in case of change, gets updated or refreshed extremely fast (Mooney, 2015). It also includes representation of paths for non-motorised means of transport, which is essential for the analysis of the jogging and walking movement patterns. However, data quality and completeness studies by Haklay (2010) assessed that it varies by country and then within each country, while the coverage of a given area is strongly linked to the accessibility of technologies to record geographic data and education related to internet technologies and mapping. The suitability of OSM street network for this kind of research is explicitly discussed in chapter '4.2.1 OpenStreetMap'.

Karimi & Kasemsuppakorn (2013) suggest a method, which most probably gets the closest to the desired outcome of the urban space network. They focus entirely on the generation of pedestrian-only networks based on the existing road networks, collaborative mapping, using GPS traces collected by volunteers, and high-resolution satellite and laser image processing.

Generally, a street network is defined as a system of interconnecting lines and points that represent a system of roads and streets for a given area (Mora & Squillero, 2015). In case of this research, the urban space network can be defined as a network which edges represent a single human-navigable space (i.e. street, footpath, parkway, square, etc.), and its vertices are intersections of such spaces, in which there are more than 2 choices of moving direction. Thus, the conventional urban street network is merely a subset of an urban space network. A number of requirements have been set for constructing such network in order to serve the purposes of this research:

- the used base datasets must be freely available for all cities in the same data acquisition and distribution method and format;

- the urban space network must be up to date, and time-compliant with the collected GPS workout data;
- the network must include paths for both motorised and non-motorised means of transport, including paths running through parks and squares;
- the network must form a single connected network (except the case of actual islands), have no duplicates, pseudo-nodes or other kind of invalid geometries;
- dangling edges of the network should not exceed 100m in order to avoid dead-ends and private property entrances;
- the network must be simple, i.e. contain a single edge for single street space and a single node of intersection for the crossroads. It has to be noticed that people, differently than while using any kind of vehicle, do not need to use designated paths. For example, a piece of road between 2 crossings together with all its sidewalks, bicycle and car lanes, green areas, parking spaces and other associated land is considered one space if there are no possibilities to navigate from it to another street without reaching one of the crossings.

A number of different approaches have been considered and attempted in order to find the optimal solution, which would satisfy all the requirements. These have been explicitly explained in the following chapters.

## 4.2 DATA SETS

### 4.2.1 OPENSTREETMAP

OSM data has been recently used by numerous researches as an underlying street or road network for routing, accessibility modelling and other kinds of related research (Mooney, 2015). However, due to its open nature, numerous issues need to be resolved before making use of the dataset for the required purpose. As noted by Gil (2014) “one has to identify which problems are present, to what extent, how critical they are for the intended use of the data set, and ultimately decide the degree of error that is acceptable, making corrections accordingly”.

The OSM data set is composed of points and polylines, attributed with an open set of tags, made up of keys with one or more values. A single polyline usually describes a single path or road, however, in some cases it can also form a boundary polygon and represent an area, which is usually tagged as “pedestrian” and stands for various parks, squares, plazas and even wider boulevards within which no further paths are drawn. Thus, the inconsistency of entity types also needs further attention while processing the dataset. The elimination of one of the entity types, i.e. the polygons, would result into missing connection between network edges, which would result into misleading snapping of GPS tracks and wrong evaluation of network configuration.

The OSM tags may be used to describe various aspects of data, ranging from the name of a street, its accessibility, surface material, allowed speed, etc., combining all the different features based on their ids. This data model is compact and very flexible, however, it also makes data querying and manipulation a non-trivial task. Moreover, the open tagging system also adds complexity through using regionalisms, different languages or simple misspells. Due to these problems, no reliable filtering method can be used in order to simplify and generalise the network.

Furthermore, Girres and Touya (2010) have listed a number of possible problems regarding the OSM street segment geometry and topology. The following ones also may result into confusion of GPS track snapping and network configuration algorithms, and therefore need to be resolved:

- duplicate segments, where geometry is exactly the same;
- overlapping segments, where geometry partially coincides;



- missing segments;
- closed segments, representing areas;
- orphans, unconnected segments from the rest of the network;
- segments without segmentation, where intersection nodes exist, or contiguous segments, separated where no intersection nodes exist;
- missing intersection nodes.

In addition to these, another major problem has been noted during this research. In case of OSM, the high level of detail presented in the street network is redundant and even confusing for the later applied analysis algorithms. While a single street in OSM can be represented by multiple lines which stand for different car lanes, bicycle lanes, footpaths and sidewalks, all these lines are still perceived as a single space by a person engaged into an active recreational activity. Figure 23 shows the difference between the general street networks as represented in OSM and urban space network as needed for the active recreational travel analysis.

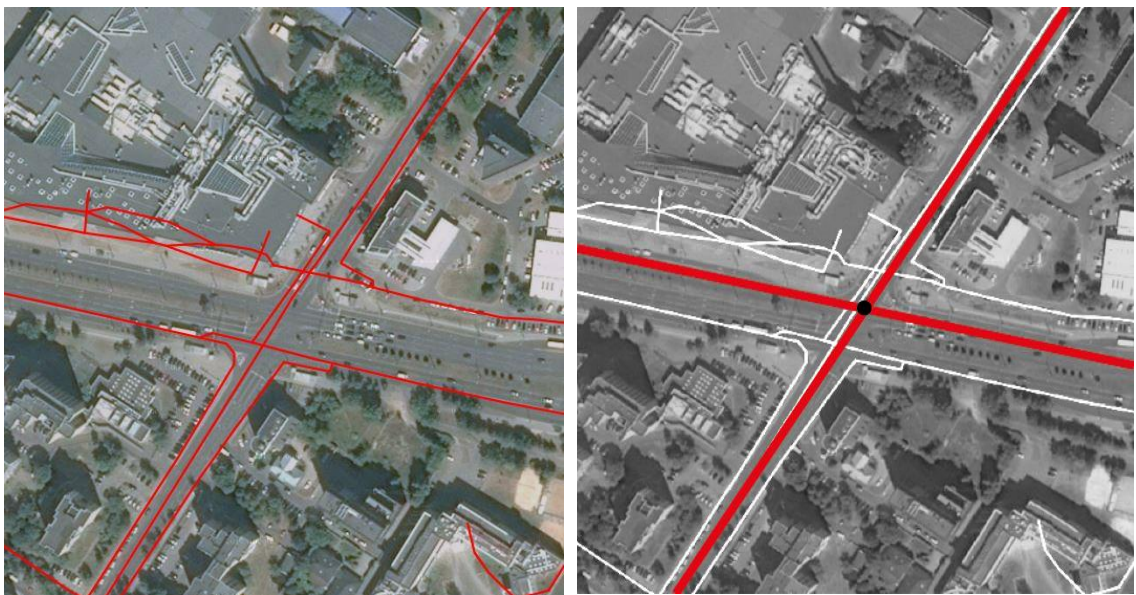


Figure 23. Original OSM street network expressed in polylines (left) and actual perceived urban space needed for the active recreational travel analysis (right).

Consequently, the OSM street network has been later processed and coupled with additional datasets in order to overcome the identified problems.

#### 4.2.2 EUROPEAN URBAN ATLAS ROAD LAND-USE DATA

The Urban Atlas is a joint initiative of the European Commission Directorate-General for Regional Policy and the Directorate-General for Enterprise and Industry with the support of the European Space Agency and the European Environment Agency. Its aim is to provide pan-European comparable and freely accessible land use and land cover data for Large Urban Zones with more than 100 000 inhabitants. The land use data is produced while combining multispectral or pan-sharpened Earth Observation data with 2.5 m spatial resolution, topographic maps at a scale of 1: 50 000 or larger and COTS (Commercial Off-The-Shelf) navigation data for the road network. In addition, ancillary data for certain classes includes local zoning data (e.g. cadastral data), field check (on-site visit) and very high-resolution imagery (better than 1 m ground resolution, e.g. aerial photographs). The resulting vector maps provide land-use classification for 21 different land-use classes with minimum overall accuracy of 85% and positional accuracy of  $\pm 5$ m (European Union, 2011).

The “Roads and associated land class” are represented by a single (multiple in case of islands or any other kind of water separation) polygon (Figure 24) which comprises fully interconnected city road network extending through the whole Large Urban Zone and thus containing city boundaries within. The associated lands are: slopes of embankments or cut sections; areas enclosed by roads, without direct access and without agricultural land use; fenced areas along roads (e.g. as for protection against wild animals); areas enclosed by motorways, exits or service roads with no detectable access; noise barriers (fences, walls, earth walls); rest areas, service stations and parking areas only accessible from the fast transit roads; railway facilities including stations, cargo stations and service areas; foot- or bicycle paths parallel to the traffic line; green strips and alleys (with trees or bushes).



Figure 24. "Roads and Associated Lands" polygon (dark grey) of the Larger Urban Zone overlaid with the boundary of Valencia city (in green).

The advantages of this data are as following:

- the provided polygon covers all the official roads for motorised transport (as of year 2006);
- the polygon covers the road zones fully, that way enabling easier determination of a single street space, i.e. a number of OSM road polylines, e.g. road lanes, cycle lanes, complicated crossroad lanes and street crossings are all covered by a single polygon part;
- the polygon is geometrically and topologically valid.

However, a number of drawbacks are also present:

- in order to use the polygon as a network, it has to be converted into a polyline feature;
- the polygon does not contain paths meant for non-motorised means of transport;
- the polygon does not contain most of the bridges, since the dataset does not allow overlapping geometries and the under-bridge land use is chosen as a priority;
- the polygon might have temporal inconsistencies with the collected GPS trajectories (8 years difference of acquiring time).

Due to all the reason mentioned above, the Urban Atlas road-land-use polygon cannot be used as it is and needs to be both processed and combined with OSM data to satisfy research needs.

## 4.3 NETWORK GENERATION FRAMEWORK

### 4.3.1 INCONSISTENCIES OF DATASETS

While Urban Atlas and OSM datasets may finely complement each other in different aspect, a number of inconsistencies need to be solved while integrating them. The biggest mismatch between the datasets is different type of entities (polygon in case of Urban Atlas and polyline and polygon in case of OSM). Another mismatch between the datasets is geometrical, when the OSM polyline of the same street does not fall into the Urban Atlas polygon. There are also semantical mismatches, when e.g. a street in OSM dataset has a wrong tag and therefore only its sidewalks and associated lanes appear in the dataset (Figure 25). In order to overcome these issues and correct the topological errors apparent in the OSM dataset, an innovative approach has been developed as explained further. The general framework of the integration method is shown in Figure 37.



Figure 25. Geometrical, semantical and entity type match and mismatch between the OSM (red) and Urban Atlas (yellow) datasets.

The first step of dataset integration is ensuring that all of them belong to the same coordinate system. For this research, along with the default WGS 84, Europe Albers Equal Area Conic (ESRI:102013) has been chosen for visualisation and calculations since it is adapted to fit Europe and uses metric unit system, thus no further recalculation from degrees to meters is needed.

Another inconsistency between the datasets is their scale, or level of detail – the Urban Atlas city boundary data is much rougher than finely detailed land-use data (Figure 26). In case, where the city boundary is rather administrative, mismatch is not an important issue; however, it becomes important when city boundary is determined by the geographical feature, such as coastline. In that case, coarse boundary line can cause clips of important (recreational) zones of a city. In order to avoid this, the coarser

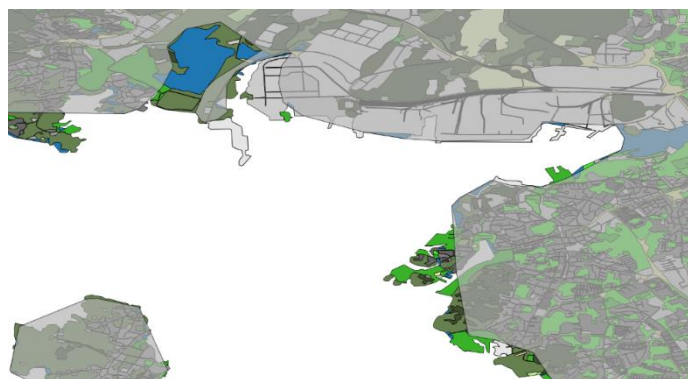


Figure 26. Scale related mismatch between Urban Atlas datasets; Gothenburg city boundary (transparent grey) is placed on top of the land use data.

boundary has to be adjusted by the finer one. After the adjustments, all remaining datasets are clipped using the city boundary polygon.

The different type of entities can be unified in two ways – one is through extracting centreline of the UA polygon and regarding polygons of the OSM pedestrian areas as their boundary lines. That way all entities become of polyline type and can be generalised by snapping vertices within a desired threshold. The second way is through buffering OSM polylines, dissolving all three kinds of polygons (UA, OSM buffered polylines and OSM pedestrian areas) into a single polygon and then extracting its centreline. In any case, the final entity type needs to be line and not area based. After trying both ways, the latter one has been chosen as more appropriate as explained in Integration of Datasets.

#### 4.3.2 SPACE CENTERLINE

The OSM street network processing framework in combination with the Urban Atlas road land use polygon requires an algorithm for the polygon centreline extraction, which would be both reliable (provide a single segment for every street) and optimised for a large dataset (should run in realistic execution time). Various algorithms have been considered and attempted as explained further.

##### 4.3.2.1 Rasterising

One of the considered algorithms is based on rasterising the vector polygon into a raster layer of very fine resolution and then using GRASS tool *r.thin*, which thins non-zero cells that denote linear features into linear features having a single cell width using the algorithm, which is explained in Jang et al. (1990).

The rasterising method for centreline extraction has been tried on a small polygon subset (1224 vertices) with raster resolution of 50cm. The thinning took 125s. The result is shown in Figure 27.

As it can be seen from the image, the result is not yet satisfying and needs further processing, moreover, the execution time is also too long to be applied on a large dataset. For these reasons, rasterising has been discarded as a possible centreline extraction method.

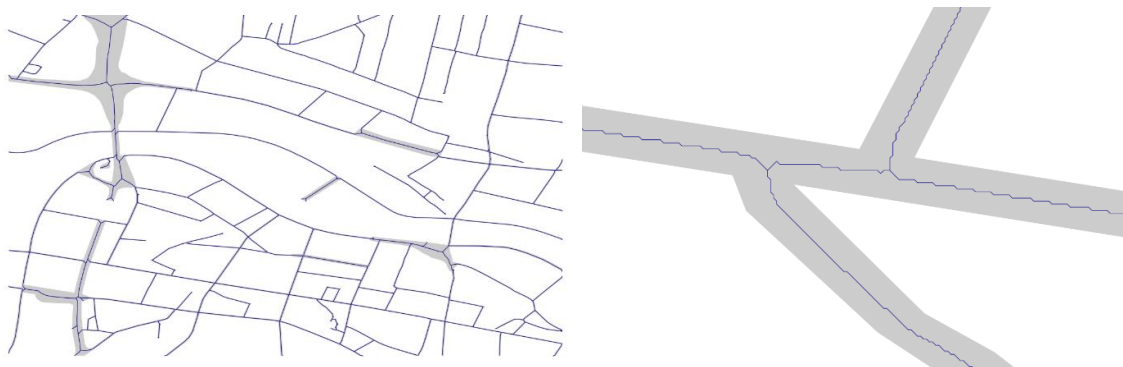


Figure 27. The resulting centerline of raster thinning method: the original polygon in grey and extracted centerline in blue.

##### 4.3.2.2 Convex polygon partitioning

Another algorithm, which was attempted to retrieve street network from a polygon, is convex polygon partitioning as explained in Miranda Carranza and Koch (2013). The essence of this algorithm is partitioning street space polygon into convex spaces which are perceived by humans as a single space. This kind of partitioning would allow construction of the street network topology using connectedness of convex polygons of street segments.



Even though the algorithm produces a proper outcome, which is well suited for street network construction and even allows street space definition from the perspective of human perception, the actual implementation requires many library dependencies, which makes the process too complicated to use and also not necessarily suitable for large datasets.

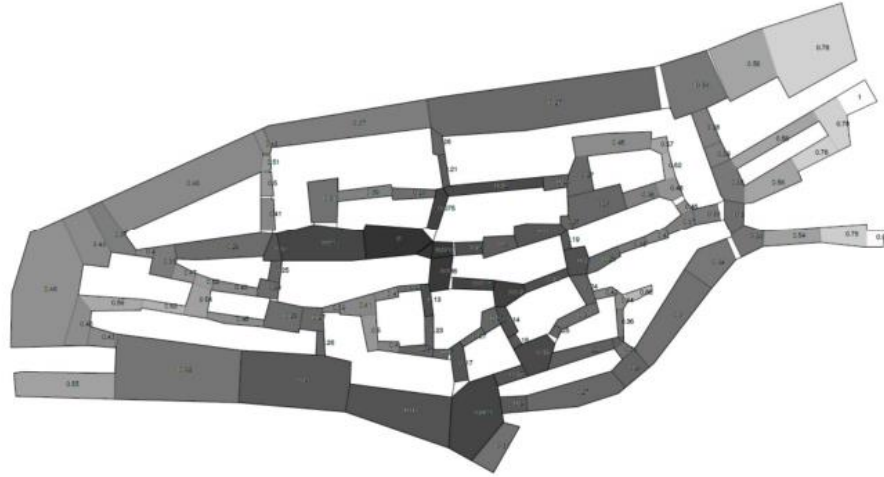


Figure 28. A street space polygon partitioned into convex spaces (Miranda Carranza and Koch, 2013).

#### 4.3.2.3 Medial axis

The medial axis of a planar shape as described by Blum (1967) is the set of the centres of the maximal inscribed disks. Ideally, the medial axis of a simple polygon is a tree-like planar graph composed of straight-line segments and portions of parabolic curves. However, the output of any computational algorithm would be precisely the medial axis of the polygon only if the computer had arithmetic with infinite precision. In any other cases, the representation of computed medial axis is sampled down to a certain precision. Later, Ogniewicz (1992) has described skeleton as a discrete medial axis of a figure, which can be computed using well-known Voronoi diagram.

There are generally two approaches to generate Voronoi based skeleton of a polygon: one is using boundary points as an input, and the other one is taking non-overlapping segments and generating segmented Voronoi (Delaunay) graph.

The first approach has been implemented using Python coded algorithm as explained by Brandt and Algazi (1992) and testing it on the final urban space network polygon of Vilnius city. Polygon's edges have been densified in order to have a vertex every 5,5m, which resulted into the polygon extending from 76 422 vertices to 1 239 547 vertices. No spatial index was implemented to speed up the process. As a result, the script was running for 8 days straight and finished with the result in Figure 29(left).

The other approach used *Boost*, which provides free peer-reviewed portable C++ source libraries. The *Boost.Polygon.Voronoi* library provides a computation of a Segment Voronoi (Delaunay) Graph, which takes line segments as an input; therefore, no geometry densification is needed. Internal segments of the generated graph are extracted in database level using PostGIS tool *ST\_ContainsProperly* (Figure 30) and resemble the centreline of the given polygon.

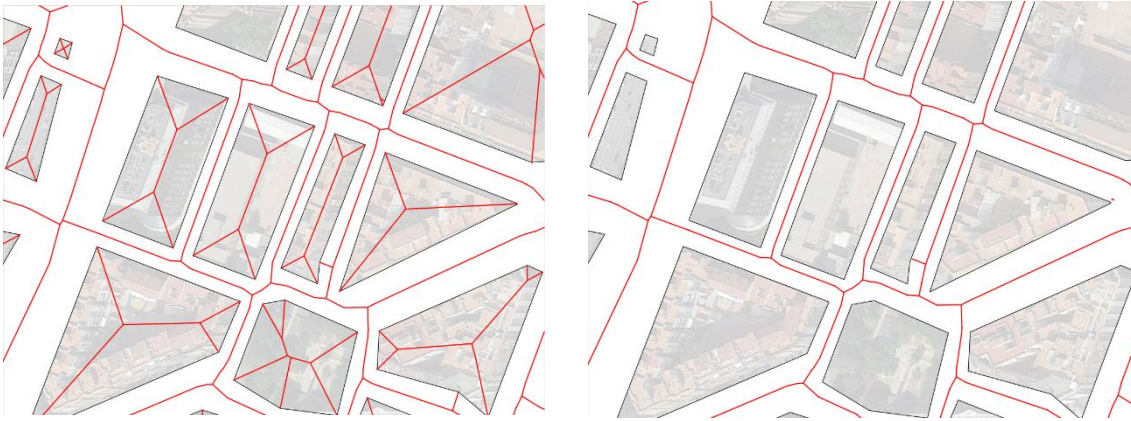


Figure 30. Left: Segmented Voronoi (Delaunay) Graph (red) of the urban space polygon (white) ; right: internal edges which resemble space centerline (red).

The same Vilnius city polygon of 76 422 vertices and accordingly, line segments, took 30s to complete with a far better result than the initial implementation of Voronoi diagram as can be seen in Figure 29. Among the minor disadvantages of the library implementation, it is noted that the algorithm provides relatively low geometrical precision as the output is straight-line segments, which approximate the arcs. However, this issue is not important in case of this research since topological relations between segments play bigger role than the geometrical precision.

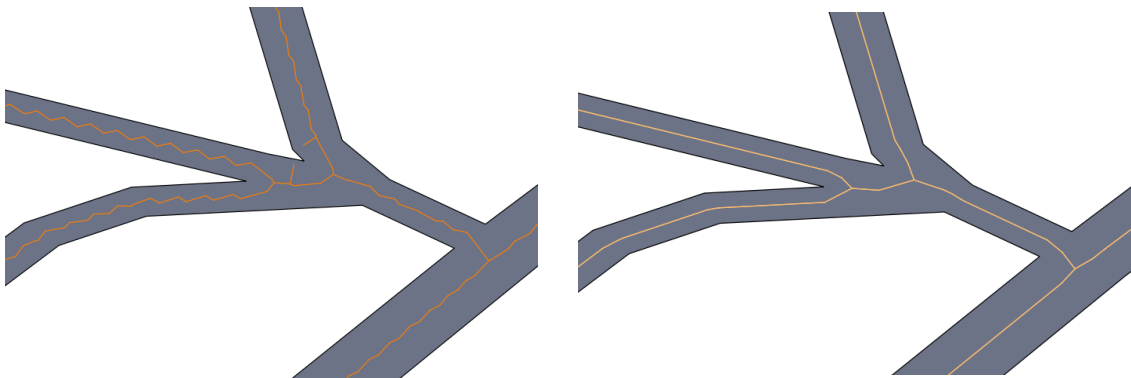


Figure 29. Left: Space centreline extracted by using standard Voronoi diagram; right: space centreline extracted by using Segmented Voronoi (Delaunay) Graph.

#### 4.3.3 INTEGRATION OF DATASETS AND GENERALISATION

As already mentioned previously, there are two ways of integrating the datasets – one is through having all entities of the polygon type and the other is having them all of the polyline type. In both cases, integration of datasets must be followed by the network generalisation and reduction of the level of detail.

Cartographic generalisation is constraint-based process used to reduce the complexity of a map. Automated generalisation has long been a research effort of cartographers (Jiang & Claramunt, 2004). Savino (2011) suggests that generalisation of road network means simply removing redundant segments which create meaningless links in a network graph. His generalisation method is based on finding cycles in a network graph below a certain radius threshold and collapsing them into a single point. Another research by Qiuping et al. (2014) generalise OSM data using a polygon-based method that relies on shape analysis and Gestalt theory, which treats polygons surrounded by roads as operating elements. The method aims to find multilane roads and generalise them into a single-line representation.

While the previously mentioned researches mainly treat road networks from a single dataset and generalisation for scaling purposes, in case of this research an additional challenge is created by using multiple datasets and pedestrian routes, which do not follow such strict patterns as road lanes.

The general level of detail set for the generalisation is 30m. This means that the smallest segments should not exceed 30m, geometric distortions are allowed up to 30m and paths lying less than 30m apart are considered to belong to the same space. Resolution is determined by the resolution of used satellite images (30m). Moreover, Modsching et al. (2006) have explored that commercial GPS receivers (which are later used for the smart devices) have to assume a positional error along the street of 28 meters for 95% of the time mostly due to the obstructions of the built environment. Finally, various errors in map digitizing can cause the road centreline disposition of a similar bias.

#### **4.3.3.1 Polygon-based Integration and Generalisation**

In order to unify the type of entities the OSM polylines are buffered by 15m. When both datasets have the same type of entity, they can be combined into one (Figure 31). However, beforehand they are simplified using Douglas-Peucker algorithm (Douglas and Peucker, 1973) with the threshold of ( $\approx 1\text{m}$ ) in order to reduce computation time.



Figure 31. Buffered OSM road polylines in grey, UA road polygon in green and OSM pedestrian area in white



All polygons are dissolved based on geometry into a single polygon and cleaned from topological errors using GRASS v.dissolve tool. Afterwards, the polygon is cleaned deleting disconnected residual parts, which mostly appear close to the boundary edges, and due to the orphan polylines in OSM dataset. Furthermore, a number of holes, which do not form a substantial gap between the paths, appear during polygon dissolving phase. They are cleaned using Python script, which removes polygon rings smaller than 90m<sup>2</sup> (10% of project resolution). The resultant polygon can be seen in Figure 32.



Figure 32. Merged, dissolved and simplified polygon of urban spaces.

After the datasets are united into a single polygon, its centreline needs to be extracted in order to return to the polyline type of entity. The centreline of a polygon is also an approximation of all the neighbouring paths into a single network edge (Figure 33). The algorithm, which is used for the space centreline extraction and simplification, is explicitly explained in Space Centerline.

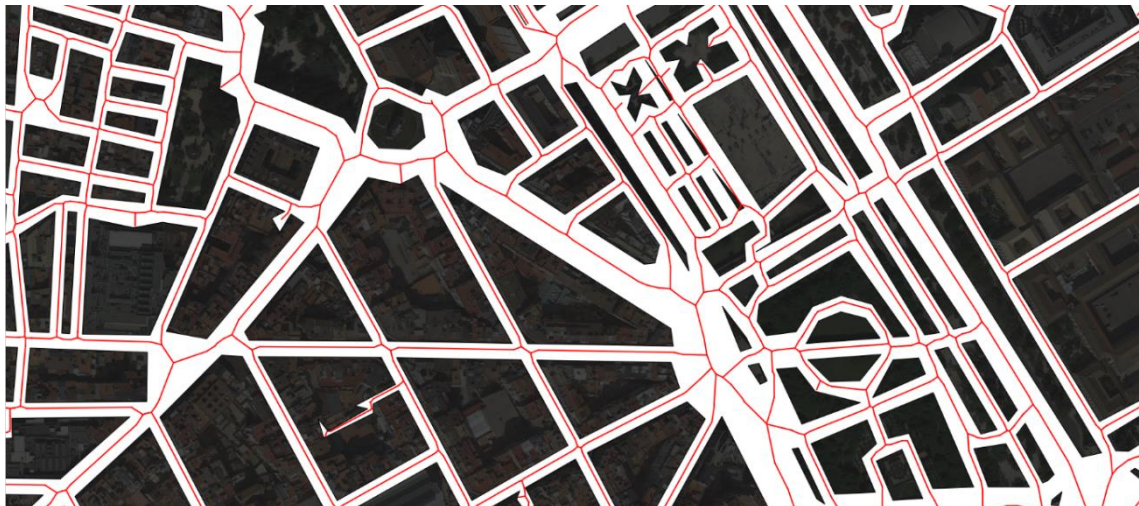


Figure 33. Centreline of urban space polygon.



#### 4.3.3.2 Polyline-based Integration and Generalisation

In order to unify all entities into the polyline type, centreline is extracted from the UA road polygon using previously explained Segmented Voronoi (Delaunay) Graph. OSM pedestrian areas are converted from polygons to polylines and their boundaries are considered as representation of navigable spaces (Figure 34).



Figure 34. OSM roads in white, centreline of UA road polygon in red, OSM pedestrian area boundaries in orange.

After all datasets are merged into a single shapefile, the further generalisation is done using GRASS tools `v.clean.break`, which breaks lines at intersections that way fixing topological overlaps, and `v.clean.snap`, which snaps vertices to another vertex not farther away than the defined threshold of 30m (Figure 35).

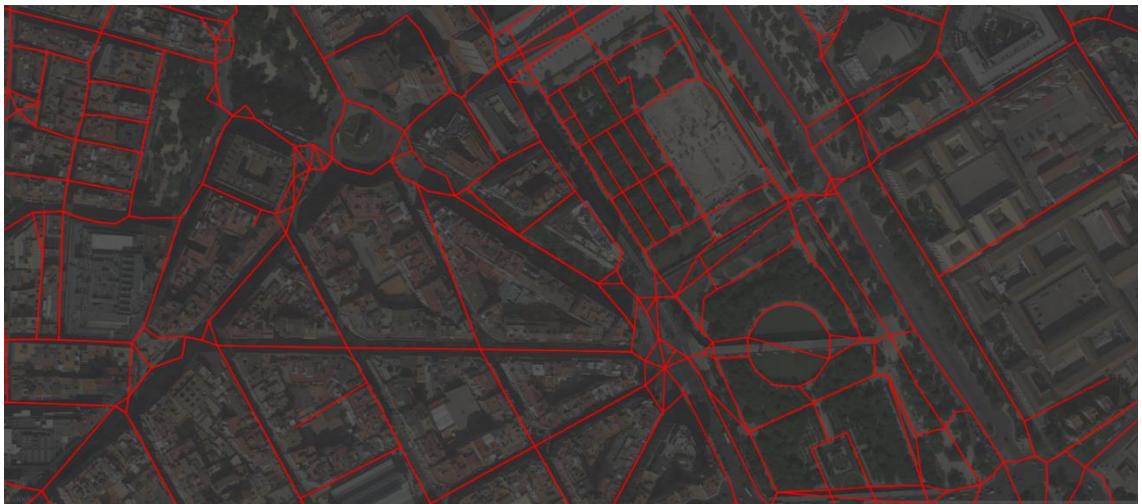








Figure 35. Urban space network after snapping vertices of different datasets within a 30m threshold.

#### 4.3.3.3 Comparison

While both of the previously explained dataset integration and network generalisation methods provide reasonable results, the methods have been compared based on the multiple criteria as shown in Table 3.

	Polygon-based	Polyline-based
<i>Better topological validity</i>	The outcome network is always topologically valid	The outcome needs to be cleaned from topological errors: mainly overlaps, pseudonodes and duplicate geometries.
<i>Higher junction simplicity</i>	Junctions need to be further processed by collapsing short segments at all intersections 	Small (4way) junctions are clean and simple, however bigger ones tend to create artefacts 
<i>Less redundant segments</i>	Centerline extraction algorithm creates redundant dangles 	Snapping algorithm results into redundant connections 
<i>Less geometric distortions</i>	Geometric distortions do not exceed 30m threshold, since the centreline always lies within the buffer zone	Polyline geometry can get severely distorted while moving all vertices of the polyline into different directions
<i>Better treatment of pedestrian areas</i>	Treats pedestrian areas as a single network edge connected with passing by spaces 	Treats pedestrian areas as a set of network edges which are connected with the passing by spaces 
<i>Preservation of attributes</i>	No attributes preserved	Attributes are preserved, however, get a random value while removing duplicates instead of keeping all possible attributes for a certain space
<i>Shorter execution time</i>	The crucial time needed for both methods is the extraction of centreline, i.e. the extraction of Segmented Voronoi (Delaunay) Graph edge, which lie inside the polygon. The buffering time in polygon-based method is comparable with the snapping time in polyline-based method, thus the final execution time stays similar.	

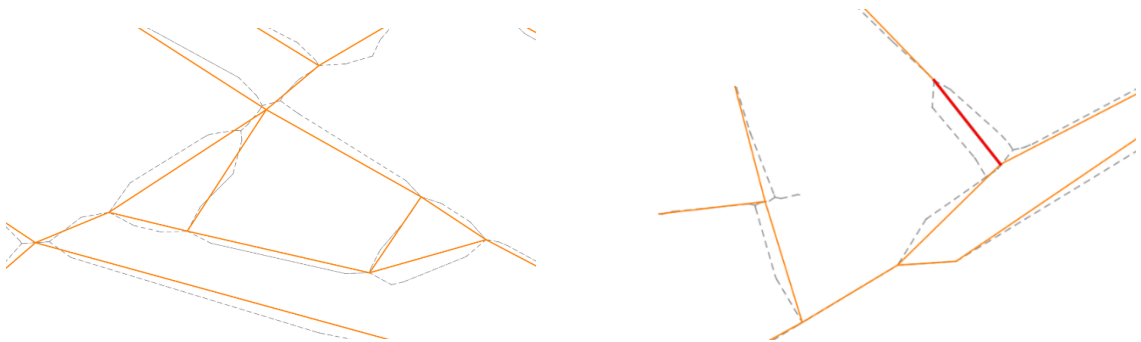
**Table 3. Comparison between polygon-based and polyline-based dataset integration and network generalisation methods.**

While both methods have their own strengths and drawbacks, the polygon-based method fits the purpose of this research better, since it provides a simpler outcome with less redundant connections, easily removable artefacts and a single network edge per single space. Furthermore, it results in less geometric distortions, unless they are yielded by the base data.

#### 4.3.4 POST-PROCESSING

After the Segmented Voronoi (Delaunay) Graph is built, later post processing is required, which includes removal of dangling residuals, collapsing short segments and line simplification, prior to which topology needs to be built to ensure network validity (Figure 36). The post processing is needed in order to decrease the complexity of the network and that way save computation time. Since all the later proceeded analysis rely either on the catchment radius around the network edge or network configuration or overlay with raster images, geometrical validity is less important than topological. It is only important to ensure that geometrical distortions do not exceed the set threshold of 30m.

Afterwards the topology is built using self-developed Python script, which also finds and removes invalid geometries (0 length segments), deletes dangling residuals and collapses short segments, which are less than 30m length, which are all the artefacts of Segmented Voronoi (Delaunay) Graph. The final simplification of line segments is done using Douglas-Peucker (Douglas and Peucker, 1973) line simplification algorithm with the displacement threshold of 15m (half of the project resolution). After the post-processing, topology is validated against the duplicate geometries, pseudo nodes and line intersections.



**Figure 36. Centerline after the post-processing: removing of dangles, collapsing short segments and applying Douglas-Peucker simplification on the line segments (orange) and the initial centreline (grey). Red line marks duplicate geometries, which appear after the post-processing.**

The overall framework used for urban space network generation can be seen in Figure 37 while the example of resultant urban space network in Valencia is shown in Figure 38.

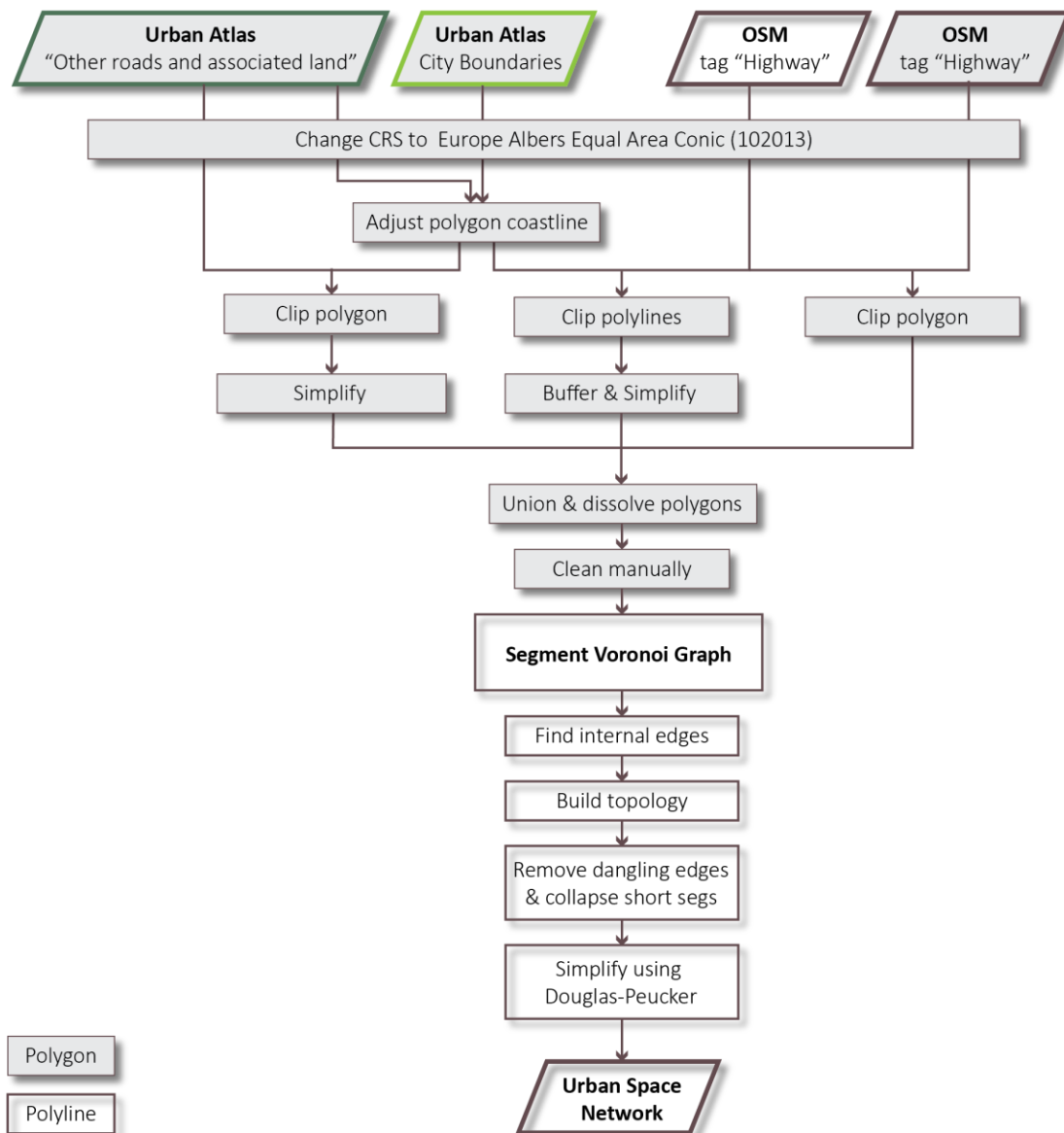
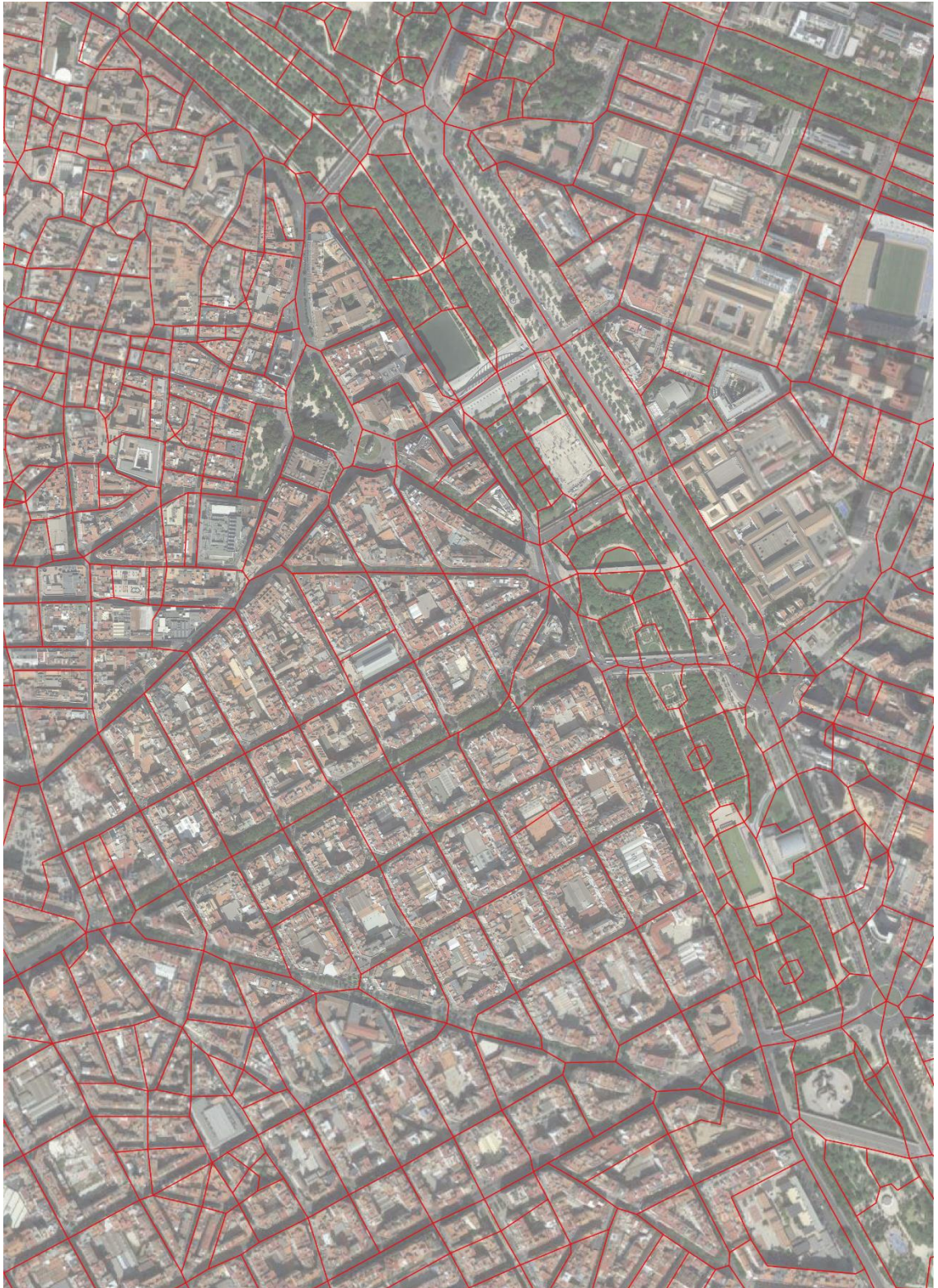


Figure 37. Integration framework of Urban Atlas and OSM datasets in order to obtain the urban space network.





**Figure 38.** Part of resultant urban space network of Valencia overlaid with Google Earth image.



## 4.4 METHOD LIMITATIONS

Even though it is true that the developed method satisfies established requirements, i.e. used base datasets are time-compliant, freely available and comparable for all cities, include paths for both motorised and non-motorised means of transport, including paths running through parks and squares. It also forms a single connected graph (except the case of actual islands), has no duplicates, pseudo-nodes or other kind of invalid geometries. Further research process has also proved that network generation method is able to provide topologically valid, clean and simple network, which is easy to handle and use for various algorithms.

However, it is the requirement for the network to be generalised and contain a single edge for single street space and a single node of intersection for the crossroads that needs to be validated. The only way to estimate this is to compare an actual image of a place (e.g. Google Earth) with the resultant network. While doing so, some limitations have been identified, namely: lack of paths and connections due to either missing data or simplification process; under-simplified or over-simplified network edges; wrongly simplified complicated intersections or too big geometric distortions (Figure 39).

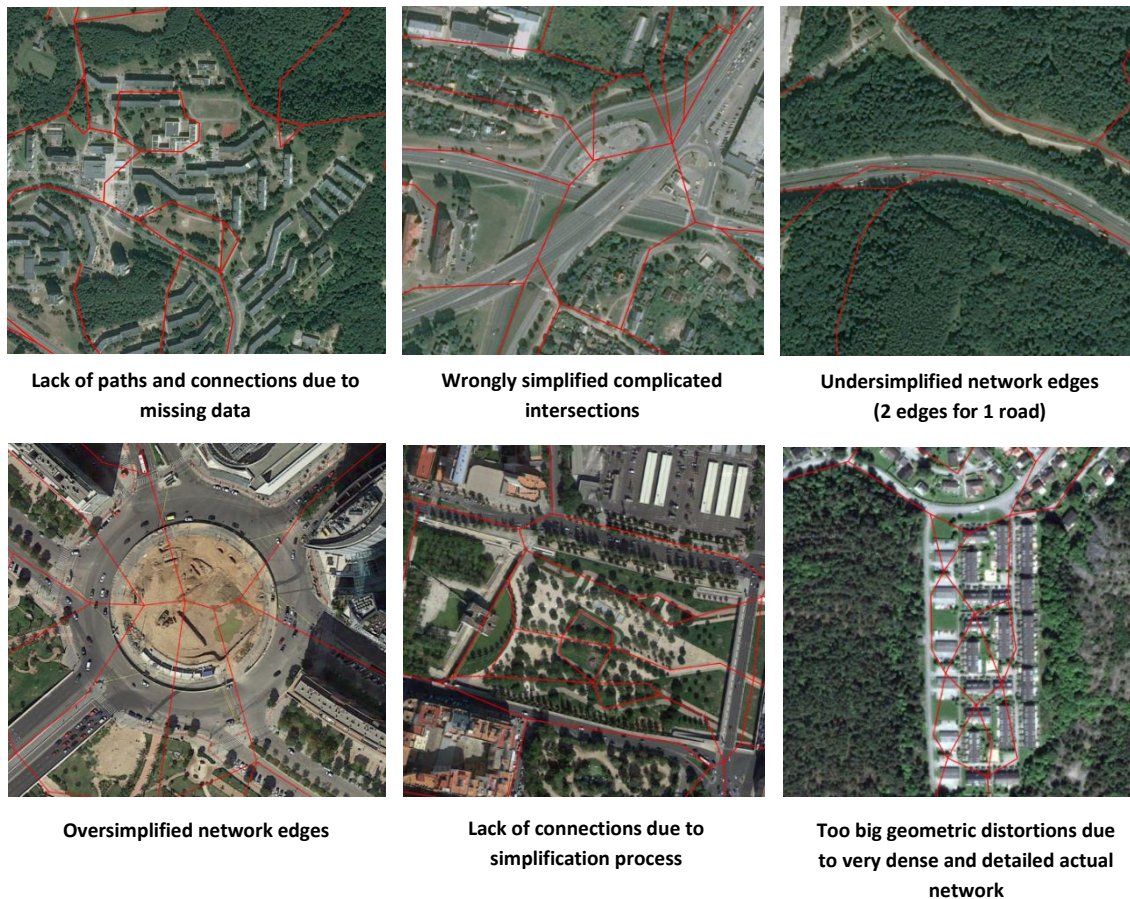


Figure 39. Identification of urban space network limitations based on visual comparison with Google Earth images.

While most of these limitations are caused rather by the base datasets than by the dataset integration and generalisation process, the process of generating urban space network can still be improved by introducing more constraints, such as not allowing the network edges to run over the buildings or by introducing variant buffer zone. Moreover, introducing the topography might unlink such cases as intersections of bridges and tunnels. However, on the other hand, since the research is aiming to investigate non-motorised means of transport some heavy traffic roads should also be considered as barriers as well as rivers or ditches, so that only certain connections through them would be possible.

## 5 . ACTUAL RECREATIONAL USAGE

While analysing actual recreational activity, a necessary attribute of a network edge is its usage for recreational purposes, which is initially defined as a number of distinct users spotted in the space represented by that network edge. However, in the initial state acquired GPS tracks are rather a set of coordinates, which are not in any way related to the urban space network (Figure 40), therefore in order to define the usage measure, GPS tracks need to be processed – effectively managed within the database, filtered and snapped to the underlying urban space network.

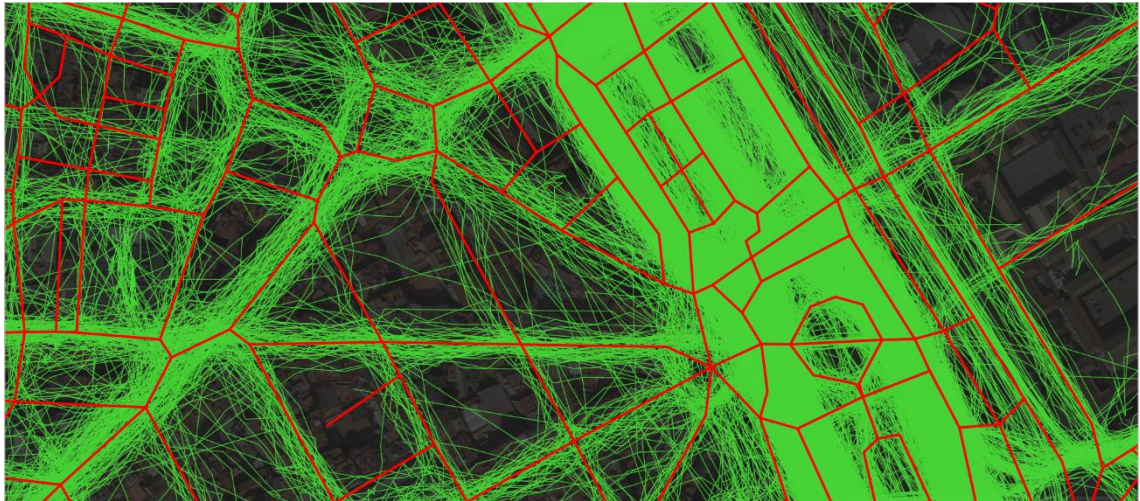


Figure 40. GPS track (green) on an urban space network (red) prior to snapping.

The overall procedure of attributing recreational usage to the edges of urban space network is shown in Figure 41. Most parts of this procedure are done using simple SQL queries except for filtering and snapping of GPS points for which separate Python scripts have been developed as explained hereafter.

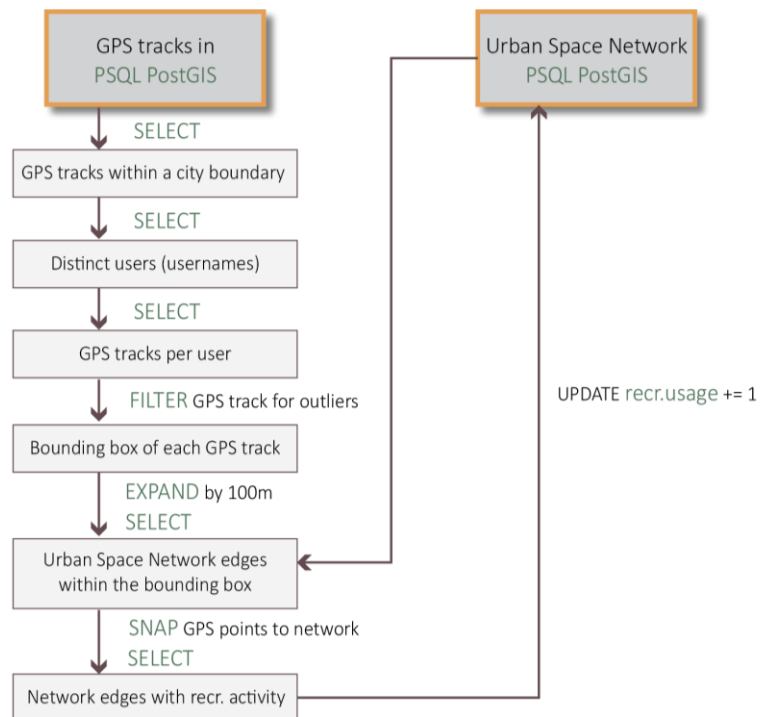


Figure 41. Attributing recreational usage to the edges of urban space network.

## 5.1 DATA MANAGEMENT

All GPS data is kept as PostGIS geometry in a PostgreSQL database. PostGIS is a spatial database extender for PostgreSQL object-relational database. It follows the Simple Features for SQL specification from the OGC and adds support for geographic objects, allowing geospatial queries to be run in SQL. There are multiple reasons to use spatial database instead of a usual PostgreSQL implementation. The reasons are namely:

- Ability to store GPS tracks as spatial objects;
- Ability to use simple SQL expressions to determine spatial relationships between objects (distance, adjacency, containment, etc.);
- Ability to spatial operations on a database level (find area and length of an object, intersection or union between objects, buffer and etc.);
- Ability to create spatial indices;
- Integration with other software (Qgis, Geoserver)

Furthermore, since the size of data is vast, its storage must be optimised for spatial queries. Spatial indices enable using a spatial database for large datasets by organizing data into a search tree, which can be quickly traversed to find a particular record. Without indexing, any search command performs a sequential scan through every record in the database.

Based on the availability of indexes supported by the PostgreSQL database and their characteristics, the data columns have been indexed as following:

- **Hash** indexes handle simple equality comparisons. They have been used for such data columns as workout type or username, where natural ordering is not possible.
- **B-Trees** are used for data, which can be sorted along one axis; for example, numbers, letters, dates, etc. The index is used whenever an indexed column is involved in a comparison or in retrieving data in sorted order.
- **GiST** indexes are not a single kind of index, but rather an infrastructure for implementing many different indexing strategies. This index breaks up data into "things to one side", "things which overlap", "things which are inside", and can be used on a wide range of data-types, including spatial data, thus geometry columns. PostGIS uses an R-Tree index implemented on top of GiST to index spatial data.
- **Clustering** is in fact not indexing the table but physically sorting its values on a disk according to the chosen index so that values that are more similar are kept closer to each other. If some data is intended to be accessed more than others are, and an index groups them together, clustering is beneficial. Once the index identifies the table page for the first row that matches, all other rows that match will be already on the same table page, and disk accesses are saved and the query is speeded up. In case of this research, clustering is applied on geometry columns.



## 5.2 FILTERING GPS TRAJECTORIES

Filtering of GPS points is needed in order to remove blundering values which appear in GPS trajectories due to various reasons as explained in chapter '3.3.3 Incorrect GPS tracks'. Removal of such values from GPS trajectories is a well-known problem, successfully tackled by various researches (Schüssler & Axhausen, 2009; Biljecki, 2010; etc.). Filtering outliers has been detached from the initial filter that takes place while writing data into the database in order to reduce total filtering time and be able to process only the relevant GPS tracks (i.e. those that fall into the territory of case study cities). However, this move appeared to be a trade-off between filtering time and loss of individual GPS fix attributes.

Consequently, such methods as proposed by Schüssler & Axhausen (2009) and Auld et al. (2008), which suggest removing the outliers from GPS data based on the unrealistic altitude or speed between two GPS points become unavailable. The methods explored by Biljecki (2010) rely on detecting sudden speed and acceleration jumps, sudden changes in heading and introducing error buffer. The detection of sudden speed and acceleration jumps is not possible due to the lack of data attributes. Furthermore, error buffer also cannot be introduced based on the characteristics of a GPS device, since data is collected from a very wide range of GPS devices, which may all have quite distinct characteristics. Finally, detecting sudden azimuth changes causes a significant amount of valid points to be marked as outliers as well (Biljecki, 2010).

In case of this research the under-filtering is the lesser evil than over-filtering due to the snapping algorithm which relies on a sequence of points rather than every single fix as explained in chapter '5.3 GPS Track Snapping on an Urban Space Network'. In addition, scarce data should not be lost during the outlier filtering (Figure 42). Finally, noisy points, which lie relatively close to the other points, have little or no influence on the snapping algorithm and the problem is caused only by outliers, which lie far away from the other points. It increases the bounding box of the GPS track geometry and consequently the search space of the network.

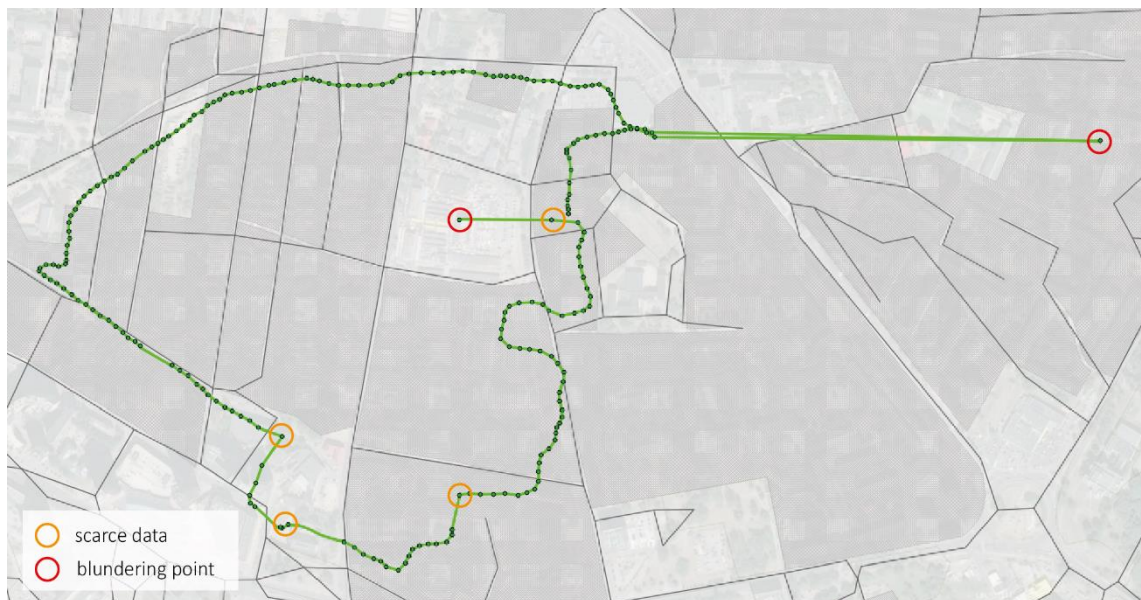


Figure 42. Geometrical difference between scarce and blundering GPS fixes.

Therefore, the definition of an outlier has been formulated as following: it is a point that lies from both of its neighbours further than three times the median while the distance between the neighbours is less than the smaller distance between the point and each of its neighbours (Figure 43). Median refers to the median distance between two consecutive points calculated for each GPS track individually.

Median value is used instead of the average because some blundering points happen to be thousands of kilometres away from the actual user location, therefore such value would highly influence the average while the median is not affected by the blundering values and most appropriately represents the “normal” distance between two consecutive points. The heuristics of using three medians comes from the evaluation of a sample set of 100 randomly chosen GPS tracks from different cities, which can be visually confirmed as not having outliers. The calculation is as explained in Pseudocode 2.

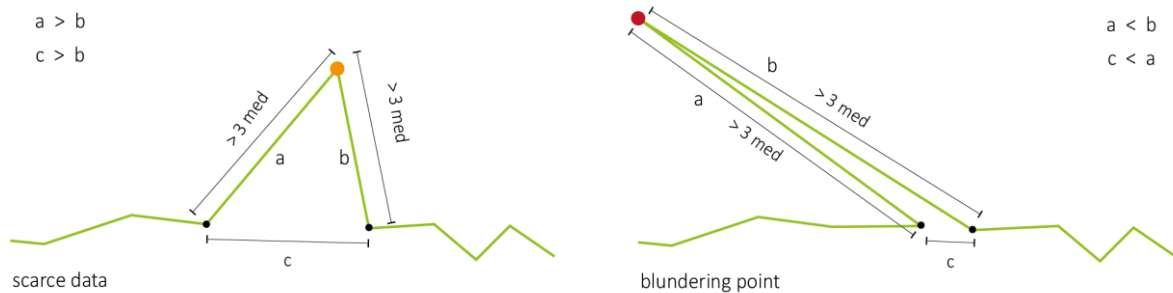
```
for each GPS_track:
    median = median distance between the 2 consecutive points
    ratio = absolute deviation from the median / median
mean = mean (ratios)
standard_deviation = standard_deviation (mean, ratios)
```

**Pseudocode 2. Determination of ratio between a point and its neighbours, which can be considered as suspicious.**

The results are:

mean = 2.2215; standard\_deviation = 0.8429.

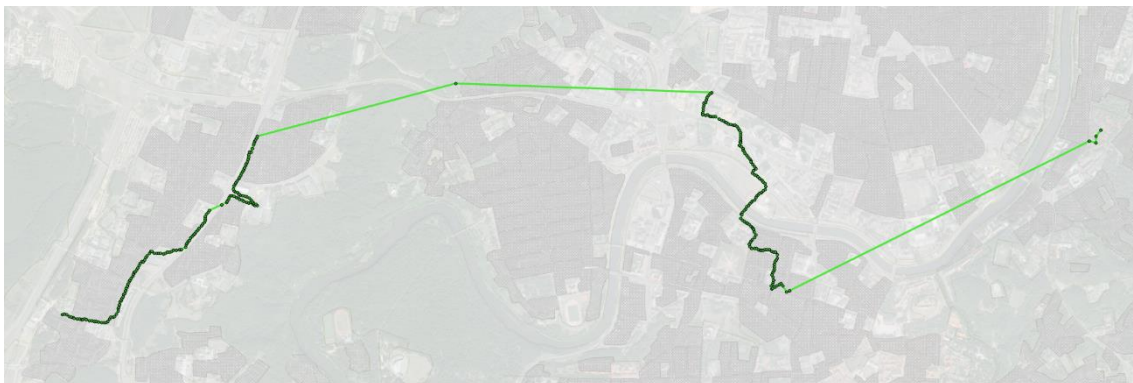
Thus, if the ratio between a point and its neighbours is higher than the mean ratio plus the standard deviation, the point can be considered as suspicious.



**Figure 43. Determination of a blundering point within a GPS track.**

Additional check is set to eliminate blunders, which happen between the scarce GPS fixes, which makes sure that the point does not lie further than half a kilometre away from its neighbours, no matter how distant the neighbours are between themselves. If the first or last point of a track is more than three medians away from the next (previous) point, it is also considered an outlier.

The limitation of the developed filtering algorithm is that it cannot cope with a sequence of blunders as depicted in Figure 44. However, in such case it is also doubtful whether the group of points is actually a result of wrong positioning or an extremely long time of missing GPS fixes.



**Figure 44. A sequence of blundering points, which confuses the outlier filtering algorithm.**

### 5.3 GPS TRACK SNAPPING ON AN URBAN SPACE NETWORK

The technique of associating coordinate values obtained from on-field GPS devices with the digital network in GIS science is called map matching. The reason for the geometric mismatching between the GPS tracks and the space network lies both in the inaccuracy of the GPS measures and the missing data in the datasets which were used to construct the urban space network. A number of map matching algorithms have been analysed and considered for this research as explained further.

According to Quddus et al. (2015), map-matching algorithms can be classified into three groups:

- Geometric – the ones that use only geometric information of the road network. They can be further classified into point-to-point matching, point-to-curve matching and curve-to-curve matching algorithms. The required input data include only position fixes (coordinates) and a base network map therefore it can be applied to any frequency positioning data. However, the accuracy of such algorithms usually ranges around 80-85%.
- Topological – these algorithms make use of topological analysis of a road network, GPS fixes, turn restrictions at junctions, road curvature, road segment connectivity, etc. The performance is much better than geometric algorithms with accuracy ranging from 80% to 96%. However, they are only effective in for matching high frequency positioning data.
- Advanced algorithms use advanced statistical, mathematical and artificial intelligence techniques. They utilise the additional information such as the positioning data quality, vehicle heading error and speed and therefore, the performance is even better than topological map-matching algorithms. However, most of them are also designed for use with high frequency positioning data.

However, most of the map-matching algorithms tend to deal with the GPS tracks of vehicle movements, which are in many aspects different from the running and walking workout data. E.g. runners as in contrast to vehicles, do not necessarily stay on a path, can also change moving direction at junctions as well as in the middle of an edge, do not have any movement direction restrictions or predictable moving speed. Moreover, very little is known about the characteristics of the GPS device, positioning data quality or satellites in range and the frequency of GPS fixes is variant. Due to these reasons, most of the advanced algorithms cannot be implemented and therefore only geometrical and topological data is used for workout snapping.

The algorithm, which is used for the GPS snapping in this research, has been developed based on the algorithms proposed by Marchal et al. (2004), Yang et al. (2005), and Quddus et al. (2015). It is a topological algorithm, which relies on the multiple hypothesis' technique. It allows to keep track of several positions or paths at once and to select eventually which candidate is the best. The first point is snapped to the two closest segments of the extracted piece of the whole network. Later, the best-fit edge is decided by checking the following points and choosing the best matching one. The path is augmented through topological connections of the best fitting edge, always choosing 2 of them based on a single point and deciding the better one based on a sequence of points up until the last GPS point is reached.

The pseudo-code for the map-matching algorithm is shown in Pseudocode 3.

```

INPUT: sequence of GPS coordinates to be snapped;
       list of candidate edges (all, when initializing or neighbours of previous
snapping candidates).
OUTPUT: first and second best fit edges for snapping.

Function snapGPSpoints:
best_fit_edge = closest edge to the first point
alternative_edge = second closest edge to the first point

while points_snapped_to_best_fit_edge < all_GPS_points:
for each of GPS points:
    for each in (edge + its direct connections):
        find deviation from point
    choose smallest deviation
    points_snapped (best_fit) += 1
    total_deviation (best_fit) += deviation(point, edge)
    if edge of smallest deviation <> best_fit_edge:
        break

while points_snapped_to_alternative_edge < points_snapped_to_best_fit_edge:
for each of GPS points:
    for each in (edge + its direct connections - best_fit_edge):
        find deviation from point
    choose smallest deviation
    points_snapped (alternative) += 1
    total_deviation (alternative) += deviation(point, edge)

if points_snapped_to_best_fit_edge < 3:
    extra_check:
        deviation(best_fit_edge) = sum (deviation (snapped_points +
2_following_points))
        deviation(alternative) = sum (deviation (snapped_points +
2_following_points))
        if deviation(best_fit_edge) < deviation(alternative):
            extra_check = True
        else:
            extra_check = False

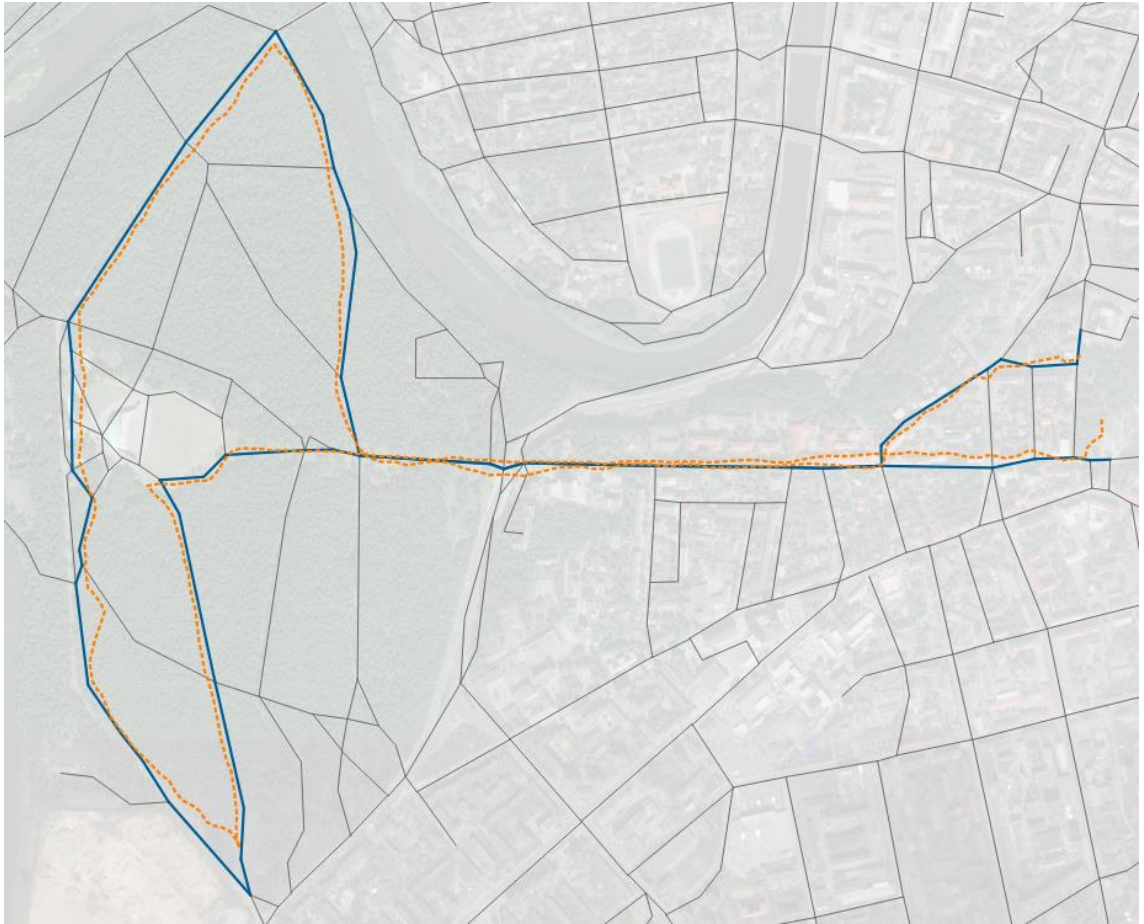
if total_deviation (best_fit) < total_deviation (alternative) and extra_check:
    if snapped_points == all_points:
        return best_fit_edge
    else:
        new_best_fit_edge = closest edge from best_fit_edge_connections to the first not
snapped point
        new_alternative_edge = second closest edge from best_fit_edge_connections to the
first not snapped point
        return snapGPSpoints (new_best_fit, new_alternative)

else:
    while points_snapped_to_alternative_edge < all_points:
    for each in (edge + its direct connections):
        find deviation from point
    choose smallest deviation
    points_snapped (best_fit) += 1
    total_deviation (best_fit) += deviation(point, edge)
    if edge of smallest deviation <> best_fit_edge:
        break
    if snapped_points = all_points:
        return alternative_edge
    else:
        new_best_fit_edge = closest edge from alternative_edge_connections to the first
not snapped point
        new_alternative edge = second closest edge from alternative_edge_connections to
the first not snapped point
        return snapGPSpoints (new_best_fit, new_alternative)

```

**Pseudocode 3. GPS trajectory snapping on an urban space network.**

The sample results of map matching algorithm can be seen in Figure 45.



**Figure 45. Urban space network attributed with the value of recreational usage: thick blue line indicates presence of the recreational activity; thin - absence. GPS track is shown in orange.**

The computational time is first reduced by considering only the part of the network, which falls into the bounding box of the GPS track under investigation, expanded by 90m to all sides. The size of expansion is based both on the level of detail of the whole network and theoretical positioning accuracy of GPS devices multiplied by 3 which means that since details smaller than 30m are not significant, the geometrical accuracy of the space centreline should not exceed 60m (2x30m) plus the error of GPS.

The accuracy of the map-matching algorithm has been computed visually comparing the GPS track with the assigned urban space network edges of 25 randomly selected samples in Vilnius city, which all together make up almost 5000 GPS points. Mapping accuracy has been computed as a number of correctly assigned network edges over the number of all edges considered (assigned, over-assigned and under-assigned) and results into 85% of overall mapping accuracy, which is reasonable for a geometrical/topological, map matching algorithm. The average deviation of GPS points to the network edge they are snapped to is 76.587m, which is mostly influenced by a few tracks with an extremely high average point-edge deviation. The more representative median deviation value is 15.859m, which is twice higher than the expected positioning accuracy of a GPS device.



<i>Route ID</i>	GPS points	Correctly assigned network edges	Over-assigned network edges	Under-assigned network edges	Mapping accuracy (%)	Average point-edge deviation (m)
352935628	58	10	0	0	100	12.006
352659805	215	0	1	0	-	52.644
352655104	246	16	0	0	100	11.319
352652635	224	32	0	0	100	15.861
365243312	264	68	0	0	100	21.373
365237419	146	6	0	0	100	12.367
365235329	149	27	1	1	93.1	135.655
365226343	295	7	2	0	77.78	9.453
365225677	285	44	0	0	100	11.436
365223658	278	40	13	5	68.97	224.394
365223653	102	9	1	3	69.23	44.158
365221775	279	10	0	0	100	8.119
392592229	301	33	10	9	63.46	521.543
392591927	292	26	0	0	100	15.858
392590430	272	14	1	0	93.33	10.028
392589996	203	34	0	0	100	10.395
392588224	273	29	1	0	96.67	12.386
392587104	82	11	2	1	78.57	9.717
392584966	87	0	3	0	-	29.332
392581830	34	9	0	0	100	9.013
392576596	302	65	3	0	95.59	38.136
392575938	248	22	3	0	88	17.484
392573791	189	6	2	15	26.09	467.64
392571659	150	7	3	8	38.89	137.772
<i>Total</i>	<b>4974</b>	<b>525</b>	<b>46</b> commission <b>8.11%</b>	<b>42</b> omission <b>7.4%</b>	<b>85.64437</b>	<b>avg. 76.587</b> <b>med. 15.859</b>

Table 4. Accuracy of the developed map-matching algorithm for GPS track snapping to the urban space network

As it can be seen from Table 4, network edge over-assigning is more frequent than under-assignment. This, as well as mostly all other inaccuracies happen mostly due to the lack of edges in the network, i.e. recreational activities happening in spaces which are not represented by any edge in the network. This can happen because of two reasons – either the lack of an existing path in the OSM or Urban Atlas data or the absence of a path as such, e.g. running in outdoor stadium, in meadows or private lands (Figure 46).

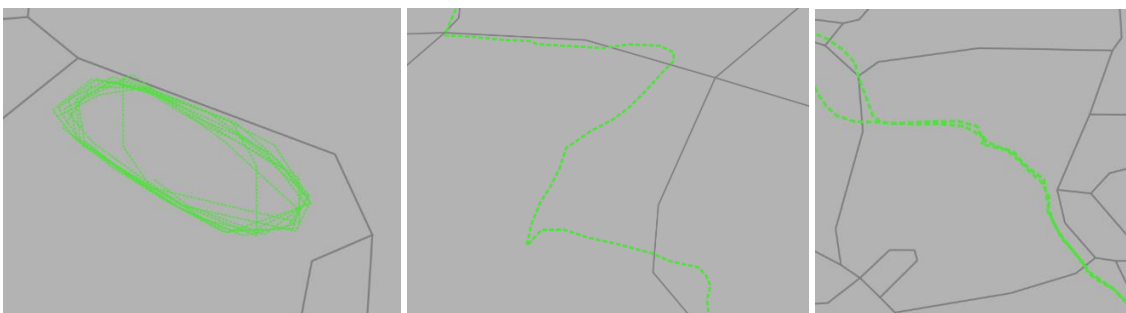


Figure 46. Cases, which confuse the map-matching algorithm: running in stadiums, running far away from the designated path, absence of edges in urban space network based on absence of paths in OSM and Urban Atlas datasets.

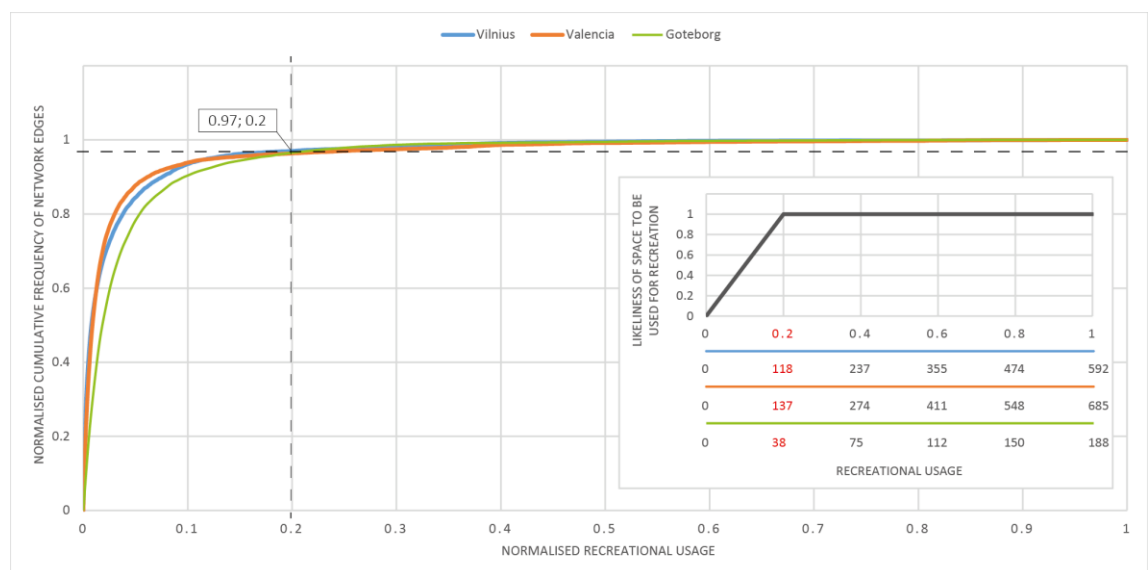
## 5.4 VALUE OF ACTUAL RECREATIONAL USAGE

The overall goal of using GPS data from mobile sports tracking applications is being able to define where recreational activities happen, where they do not and finally, use this knowledge to model the potential recreational usage of urban spaces, i.e. give an indication to every space of how likely it is that the space is or will be used for recreation. In theory, if defined correctly, the values of potential recreational usage should correlate with the actual recreational usage, i.e. there should be correspondance between the calculated and observed values. However, in case of recreational usage, literal quantification is impossible, since it is a rather qualitative measure.

Furthermore, current value of recreational usage – which is literally a number of distinct sports tracking application users spotted in an urban space – does not fully account for how much the space is used for recreation, since twice more spotted users does not mean that a space is twice more used. This is both because the usage value in this case is not quantifiable and because the number of spotted users is biased itself due to the limitations explained in chapter ‘3.3 Data Limitations’. The value of potential recreational usage itself is a combination of multiple values each of which have different meanings, scales and ranges, so there is no such way of combining these values into ‘the number of application users spotted in a an urban space’. Due to these reasons, all values have to be normalised into a single measure, which would be consistent for both actual and potential measures of recreational usage.

In order to quantify a qualitative measure of recreational usage, a fuzzy notion of likeliness has been used (Klir & Yuan, 1995). It describes how likely it is that a space is used for active recreational travels and is measured in the range of 0 to 1, where 0 means no usage and 1 means that space is definitely used for recreation. All values in between indicate how big is the chance for a space to be used for recreation. By using the fuzzy membership logics, it is possible to unify and consistently compare the actual and potential values and that way determine which variant of potential usage model best fits the actual situation.

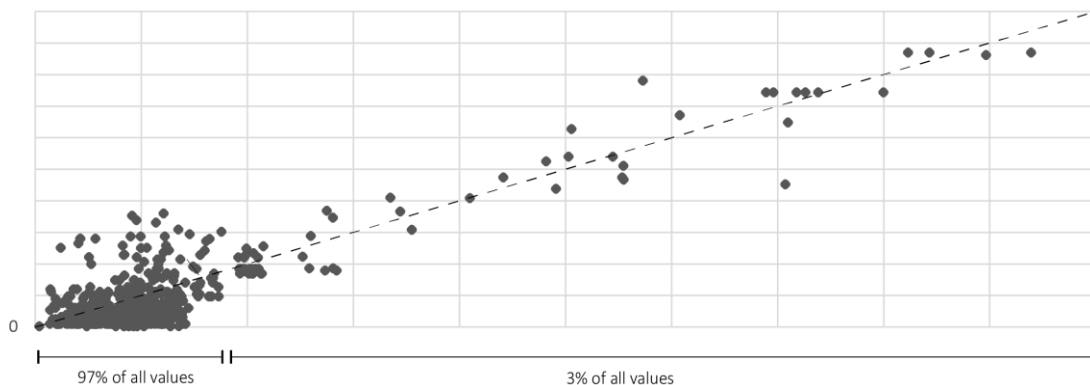
Actual recreational usage has been mapped into the coefficient of likeliness for a space to be used for recreation as can be seen in Figure 47. Values for all three cities have been normalised using the cumulative frequency of urban spaces in a network, which have recreational usage value above 0, thus are regarded as ‘used for recreation’, and the maximum number of recreational usage available in that city.



**Figure 47.** Graph of relation between the recreational usage of a space and total number of spaces which have the same or smaller value but not 0. The values of all 3 case study cities are normalised. The small graph represents the corresponding graph of normalised recreational usage and space likeliness to be used for recreation.

As it can be seen from the graph, the distribution of values differ per city, however all graphs converge and approximately even out at one point. The common point in this case denotes reading from which on the normalised cumulative frequency does not depend on the individual characteristics of the cities. Simply speaking, it means that in all cases only 3% of the all network edges have recreational usage value higher than 20% of the maximum. Therefore, this point has been used as a reference for the likeliness coefficient, which corresponds to different recreational usage values for different cities as shown in the small graph. For example, in case of Vilnius the maximum spotted recreational usage of an urban space is 592 unique users, which means that all spaces which have recreational usage 118 or more are regarded as 'used for recreation'. Accordingly, a space, which has recreational value of 59, is considered to have actual recreational usage value of 0.5, which can be understood as 50% chance of being used for recreation.

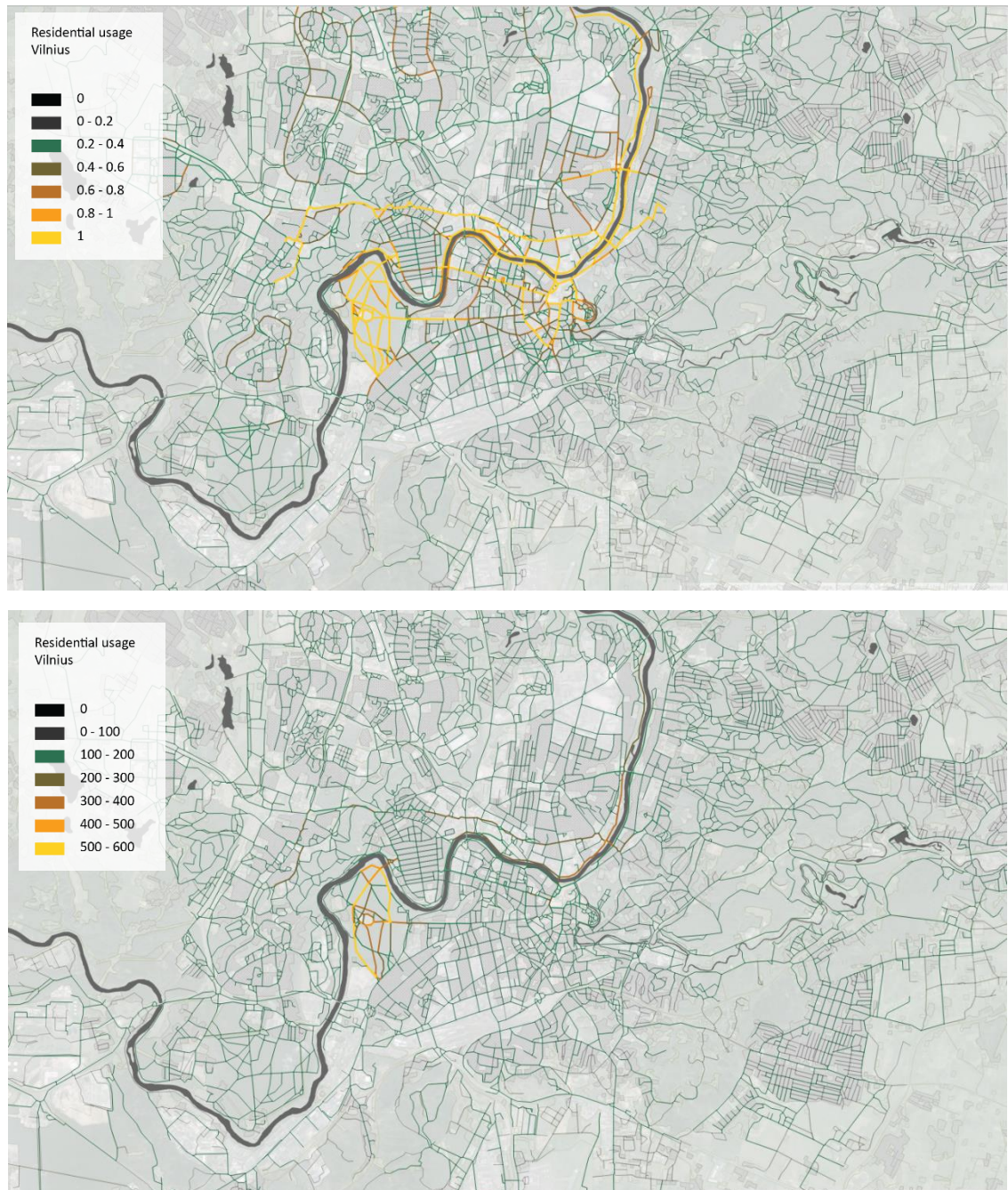
Reference points differ for each city both because of the different proportions of urban space network size and number of application users and because of different distributions of recreational activity which are dependant on individual characteristics of the built environment, i.e. higher number of attractive spaces shares out the total number of the users, while lower amount of attractive spaces concentrates the users within them. Finally, it must be noted that the choice of threshold for fuzzy normalisation does not have a real influence on correspondance between the actual and estimated values, since it merely shifts the scatter plot space in order to mitigate the spurious high correlation based on smaller amount of data while the bigger amount of data lies in an uncorrelated cloud. An example of such situation can be seen in Figure 48.



**Figure 48. Example of spurious high correlation caused by existent correlation of only small amount of data having higher values than the majority of data.**

Finally, fuzzy normalisation of actual recreational usage serves for visualisation purposes. Figure 49 shows the difference between normalised and non-normalised values. Looking at the map, it is obvious that normalised values enable more intuitive and comprehensible overview of network's recreational usage.





**Figure 49. Urban space network of Vilnius coloured according to its recreational usage value; top - colours correspond to normalised values, bottom - equal intervals of the number of distinct users per network edge.**

## 5.5 VALIDATION AND VERIFICATION

Validation and verification are procedures that are used for checking and ensuring that a product, service or system is able to fulfil its intended purpose (Maropoulos & Ceglarek, 2010). The difference between these procedures is that verification assures the developed method is able to fulfill its given task and validation is needed to check whether there is correspondance between measured / observed / estimated values and physical reality.

The particularity of this research is that it aims to provide a method which would allow to use data from sports tracking applications as a ground truth for any further relevant analysis, therefore it needs to be reassured that the data and its processing methods can serve this purpose. The matter in this case is that no other ground truth data exists, neither can be collected for such kind of validation. Verification in this case is easier, since every step of data processing is verified separately.

One way to ascertain that research results are not significantly biased by the choice of sports tracking application is through visually comparing the obtained recreational usage heatmap with the similar heatmaps suggested by other sports tracking applications. In this case Strava has been chosen as a reference application, since it has significantly more users than the chosen Endomondo application and provides clear and freely accessible images of heatmaps throughout the world. The difference between the heatmap provided by StravaLabs and developed during this research is that StravaLabs do not use an underlying space network and provide an image of aggregated GPS points in Euclidean and not in network space, while the developed heatmap is a value-added urban space network.

As it can be seen from Figure 50, even though some disparities exist between the compared heatmaps, the overall picture seems rather similar. A notable difference between the two heatmaps lies at the boundaries of a network, since in case of this research only those GPS tracks completely enclosed by the city boundary have been considered and therefore a part of data has been lost.

While it is not correct to state that similar patterns in different heatmaps can validate the process of assessing actual recreational usage of an urban space network, it is safe to assume that data acquisition, filtering and GPS trajectory snapping to the underlying network did not introduce any significant inadequacies to the result. Therefore, it can be stated that the values of actual recreational usage can be used as a ground truth data for determining, constructing and validating the model of potential recreational usage.



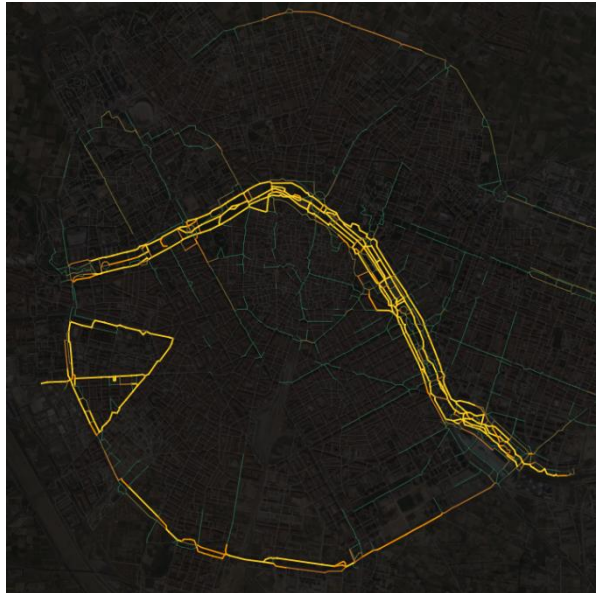
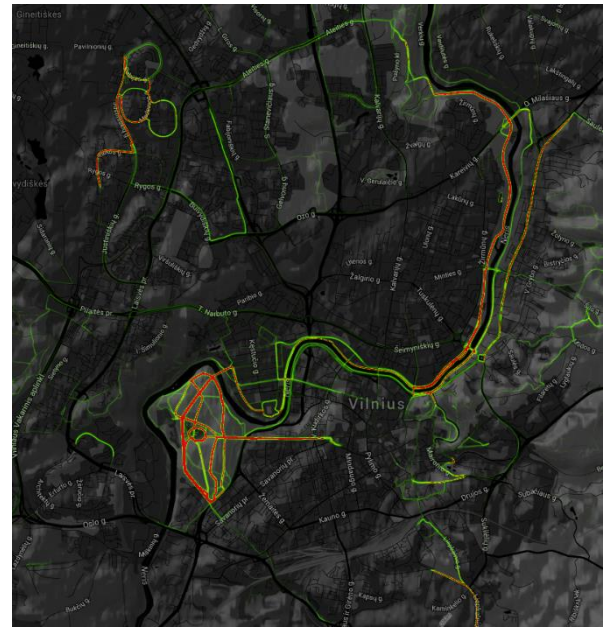
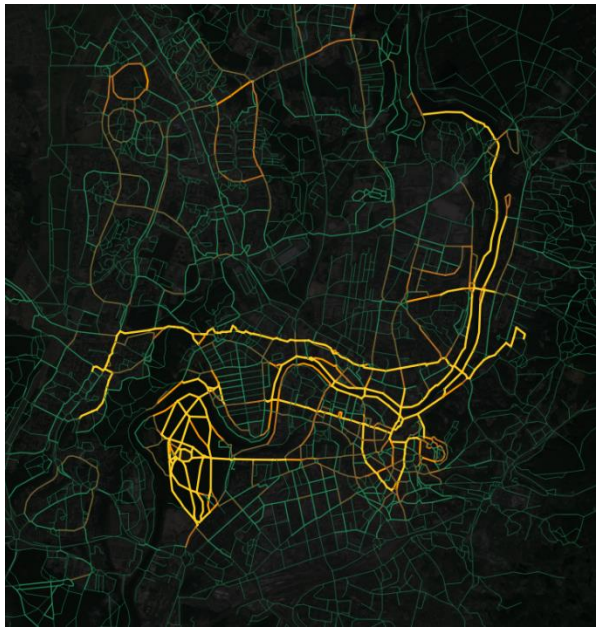


Figure 50. Comparison between the developed heatmap of recreational usage based on Endomondo sports tracking application (left) and StravaLabs heatmap (right). (<http://labs.strava.com/heatmap/> accessed 12 May 2015)



## 6 . RUNABILITY INDEX

---

While there is no consensus on what features of the built environment are definitely influential for the increase of physical activity, one of the most common notions used while trying to explain this phenomena is the Walkability Index (Troped et al., 2010). Generally, Walkability Index denotes how friendly a certain area is for walking. Cutumisu (2011) has defined walkability as an indication of the conduciveness to walking, running, biking, rollerblading, or other activities that involve non-motorized movement.

In contract to the definition Walkability by Cutumisu (2011), Choi (2013) has noticed that walking for pleasure or recreational reasons shows observably different behaviour from transportation based active travels. Recreational travels, be it walking or running, are generally conducted with less purposeful attitude, with more flexibility between moving and sojourning and not directed by the shortest distance route, as in the case of utilitarian trips.

Due to the existent differences in patterns, motivation and benefits of recreational and non-recreational movement, this research suggests introducing a separate index of space suitability for active recreational travels, which would be called the Runability Index. Generally, the Runability Index would aim to indicate the potential of space (street, footway, square, road, etc.) to be used for recreation. If defined correctly, the index should be correlated with the actual values derived from the sports tracking application data.

The most addressed indicators of Walkability Index mostly include three measures: land use mix, street network configuration and residential or population density (Giles-Corti et al., 2015; Sundquist et al., 2015; etc), which are often accompanied with the value of neighborhood greenness (James et al., 2015; etc.). The same indicators have been used as a pivot point for constructing the Runability Index. All the other measures have been adapted based on the distict characteristics of recreational travels compared with the transport-based ones. In case of this research the indicators are also influenced by the availability of data.

The characteristics and reasoning behind defining every indicator and aggregating them into a single measure of Runability Index, as well as compariosn between the estimated and actual values are explained in the following chapters.

## 6.1 VALUE OF GREENNESS

Probably the least controversial value, which has ever been associated with increase of physical activity for both recreational and transport purposes, is value of greenness, or more precisely, the amount of naturalness in a neighbourhood. Since naturalness as such is not measurable and the most common natural element in urban environment is different kinds of vegetation, the notion of greenery has superseded the vague notion of naturalness. Ever since, numerous researchers have found associations between proximity, access, size and quality of urban green spaces with the increase in physical activity, transport-related as well as recreational (Giles-Corti et al., 2005, Mowen et al., 2007 and Kaczynski et al., 2011). James et al. (2015) explain this association as vegetation's capability to reduce noise, air pollution, heat island effect, cast shadow and among others decrease stress and have direct restorative effect.

Normalized Difference Vegetation Index (NDVI) is a graphical indicator that can be used to assess whether the target under observation contains live green vegetation. NDVI calculations are based on the principle that healthy green plants absorb radiation in the visible region of the spectrum and reflect radiation in the near-infrared region (NASA.gov, 2015). It is computed using a visible red band and a near-infrared band acquired from vegetation growing seasons, i.e. in case of Europe, the overall suitable time for such data acquisition includes mostly late spring and early autumn months. (Northern Europe has a high possibility being covered with snow and loss of leaves during wintertime, while the Southern part of Europe experiences drought that affects vegetation during summertime). The following formula is applied for converting pixel values in satellite images of 30m resolution to derive the amount of healthy vegetation with the sampled area:

$$NDVI = \frac{(NIR - RED)}{(NIR + RED)}$$

where NIR – near infrared band; RED – red band of a satellite image.

Measurements can range from –1 (water) to 1 (dense, healthy green vegetation). Values close to 0 indicate such surfaces as bare soil or concrete (buildings, streets, parking lots, etc.). The boundary between different surface indicators is rather fuzzy and vegetation intensities may vary season to season.

A number of studies have used NDVI to evaluate greenness of urban areas in order to assess the greenness of urban environment, its impact on people's physical activity (Brownson et al, 2009; Troped et al., 2010; Lwin et al, 2011 and etc.), mental health (Sarkar et al., 2011) or BMI (Gordon-Larsen et al., 2006; Bell et al., 2008, etc.). All have found this variable to have a significant influence on the phenomena.

Thus, the greenness measure of a single urban space (edge of a network) has been computed as the average NDVI value of the pixels intersected by an edge under investigation using Landsat 8 satellite images as source data for computing NDVI indices (Figure 51). Since all the three case study cities fall within a single image, only one image per city has been used. All used images were taken in September 2014, with cloud coverage of less than 4% and clear vision above city boundaries.

The calculation procedure is performed within the database using PostGIS vector/raster functionalities and the average NDVI score per network edge is assigned as an attribute to that network edge. Results of the evaluation of network greenness can be seen in Figure 51.





**Figure 51. Urban space network coloured according to its greenness value, black edges correspond to heavily urbanised area with no vegetation, bright green edges – abundant vegetation; intermediate colours represent the fuzzy transition zone between two classes. Top: satellite image background. Bottom: Urban space network of Vilnius.**



## 6.2 VALUE OF LAND USE MIX

Land use mix has proven by numerous studies (Cervero & Kockelman, 1997; Ewing, 2001; 2009; Owen et al., 2004; etc.) to be one of the substantial indicators of human physical activity in a certain area. Later it has been used as an axiom for related researches (Gebel et al., 2007; Brown et al., 2009; van Dyck et al., 2010; Yamada et al., 2012; etc.). The review by Saelens and Handy (2008) of 13 prior reviews of relationships between walking and the built environment showed extensive correlation between walking and both land use mix and distances to walkable destinations, however, indicated that more recent studies support the relationship between mixed use and walking for transportation and not leisure.

In theory walkability is increased when it is more efficient than driving e.g. through congested areas where parking is often scarce and through well-connected streets that create fairly short and direct routes between destinations. Moreover, mixed use brings many diverse destinations in the same area (Brown et al., 2009). The same motivation, though, does not apply for recreational travels (i.e. running and walking) which usually do not have a destination and there is no treat between choosing means of transport. Even so, a review of 50 quantitative studies (Kaczynski and Henderson, 2007) found proximity to parks and recreational settings to be generally associated with greater physical activity.

According to Koohsari et al. (2015), proximity of recreational space can influence physical activity for leisure in at least three ways: first, recreational space can be a setting in where people engage themselves in physical activity. Second, it can be a destination to which people actively travel and change running, walking or cycling into a different activity. And third, recreational space can be used as part of a route to pass through (Figure 52). His further studies have found that proximity of a recreational space to the respondents place of residence is way less influential to one's recreational physical activity than a combination of measures, i.e. size, count, quality and accessibility.



**Figure 52. Different ways of how recreational spaces influence active physical travels.**

Another particularity of this research compared to the previously mentioned ones is that it is based on the urban space network instead of the neighbourhoods of the residential areas, i.e. it does not try to examine active recreational travel patterns in relation to where people live but in relation to the characteristics of an urban space and its neighbourhood. Due to the reasons described above, a special approach, different from a traditional Walkability Index calculation, is needed for measuring influence of land use mix on the presence of recreational activity in an urban space.

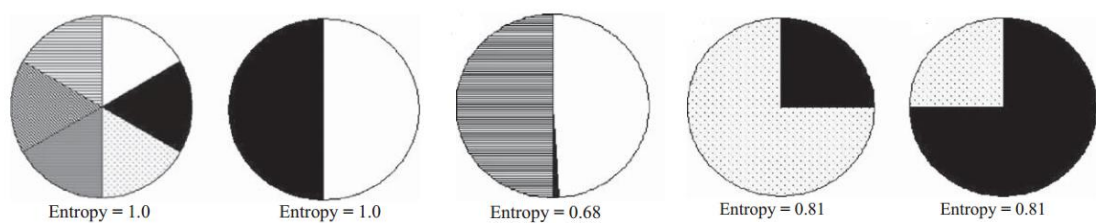
Previously described findings suggest that it is not the overall land use mix that encourages active recreational travels in a particular space but the balance between the residential (source) and recreational (target) spaces in its neighborhood. Therefore the value of land use mix for this research is regarded as the entropy score of residential and recreational land use areas within a chosen radius from the centroid of an urban space network edge. The size of radius is discussed further on.

Urban Atlas land use data has been used as a data source for the indicator of land use mix. Land use classes provided by the dataset have been classified into three groups as in Table 5.

<b>Land use group</b>	<b>Land use class</b>	<b>Notes</b>
<i>Residential</i>	Continuous Urban Fabric (S.L. > 80%)	All correspond to the residential or mostly residential urban areas
	Discontinuous Dense Urban Fabric (S.L. : 50% - 80%)	
	Discontinuous Medium Density Urban Fabric (S.L. : 30% - 50%)	
	Discontinuous Low Density Urban Fabric (S.L. : 10% - 30%)	
	Discontinuous Very Low Density Urban Fabric (S.L. < 10%)	
<i>Recreational</i>	Green urban areas	Urban recreational spaces, water bodies and natural, non-urbanised areas
	Sports and leisure facilities	
	Agricultural + Semi-natural areas + Wetlands	
	Forests	
	Wetlands	
	Water bodies	
<i>Other</i>	Isolated Structures	Land use classes which do not belong to the above categories
	Industrial, commercial, public, military and private units	
	Fast transit roads and associated land	
	Other roads and associated land	
	Railways and associated land	
	Port areas	
	Airports	
	Mineral extraction and dump sites	
	Construction sites	
	Land without current use	

**Table 5. Groups of land use classes corresponding to the calculation of the value of land use mix for the recreational active travels.**

Entropy scores appeared as variants of the Shannon index originally used to analyse accuracy of information transfer and later the measure was adapted to provide general indices of the evenness of spread across different categories (Krebs, 1999). Entropy score equals 1 when land use is maximally mixed (heterogeneous), i.e. the areas of all possible different classes are equal, and 0 when land use is maximally homogeneous (Figure 53).



**Figure 53. Examples of different land use configurations and entropy scores (Brown et al., 2009).**

Entropy's suitability for exploring land use mix values in relation to physical activity has been explicitly investigated by Brown et al. (2009) following the original work of Frank et al. (2005). They have calculated entropy score of the location of a residential building as a function of floor areas of different land use within 1km street network buffer.

In case of this research, their calculation is simplified to taking only 2 different groups of land use, disregarding the floor area and taking a crow flight buffer instead of a street network distance. Originally, only the sum of both group areas had to be taken into account while proceeding the calculations. However, in that case the measure would also favor zones of 98% industrial land use and only 1% of both residential and recreational use or other similar cases, therefore the formula was enhanced to take into account total area of a buffer as well (Figure 54).

The final formula used for entropy score calculation is as following:

$$LandUseMix = \frac{A}{\ln(N)}$$

where

$$A = -1 \sum_i \frac{b_i}{a} * \ln(\frac{b_i}{\sum_i b_i})$$

$b_i$  – area of a separate land use group within the buffer zone;

$a$  – total area of a buffer.

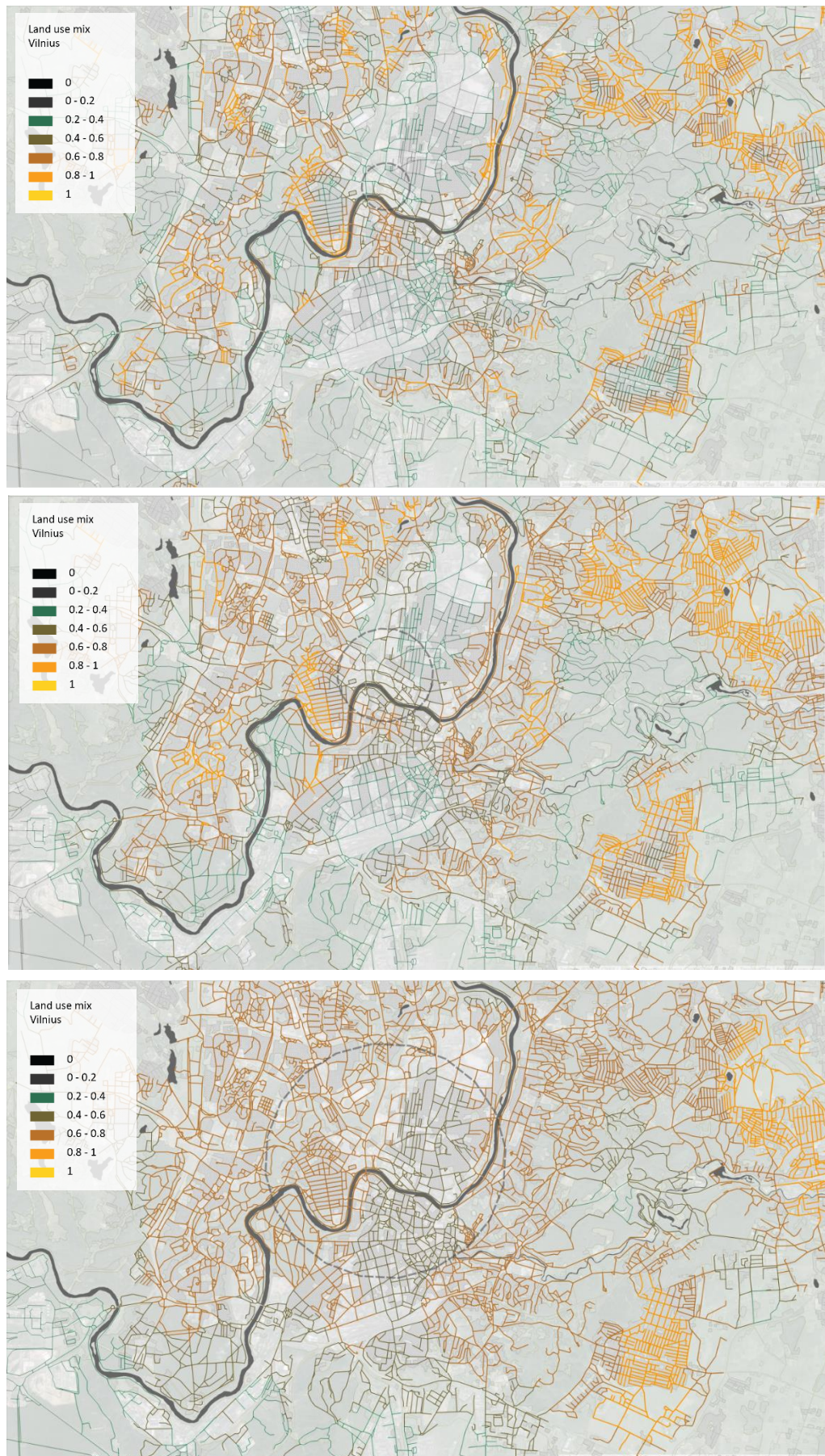


**Figure 54. Difference between the original calculation (left) of entropy score and the improved one (right). Grey areas represent residential use, green – recreational, white – other. Yellow network edges stand for high entropy, dark green – low. The circle represents 500m buffer around the centroid of network edge.**

As it can be seen from the formula and sample calculation results, the measure is highly dependant on the chosen buffer radius. The radius used for the neighborhood investigation in previously mentioned studies ranges from 500m to 3km where 500m captures an area that can be covered within a 10min walk and 3km correspond to the maximum acceptable walking distance. These values, however, may vary from country to country based on its climate, landscape and even culture.

In order to mitigate this dependency, three different buffers have been obtained for each of the case study cities to later decide which of them fits the potential recreational usage model best. The difference between results can be seen in Figure 55. 500m buffer corresponds to the whole buffer zone that can be covered within a 10min walk, 1km corresponds to the radius that can be walked from the center of a segment in a 10min walk; and 2.5km correspond to half of the average length of all GPS tracks obtained (5km), thus the average length of an active recreational travel or ‘runnable’ distance, which was the same for all three case study cities as explained later in chapter ‘6.3 Value of Network Centrality’





**Figure 55. Buffer radius influence on land use mix entropy score. From top down: 500m, 1km, 2.5 radius, urban space network of Vilnius**

### 6.3 VALUE OF NETWORK CENTRALITY

Fundamentally, centrality measures aim to quantify that in a network some nodes are more important (i.e. central) than others. In urban studies, centrality is related with such terms like accessibility, transport cost or effort. As already discussed in chapter '2.3 Urban Network Analysis', the theory of Space Syntax alongside with other measures, are also used to explain human walking behaviour.

Hillier & Iida (2005) have defined two aspects of human movement in an urban space: the 'to-movement' - selection of a destination from an origin; and the 'through-movement' - selection of the intervening spaces that must be passed through. Based on the two aspects there are two main types of analysis in Space Syntax:

- **Integration** (a variant closeness centrality) is related to 'to-movement'. It measures how close each segment is to all others, so it tells how accessible each segment is from all the others, and how much potential it has as a destination for movement: the destination potential for a segment.
- **Choice** (or betweenness) (Freeman, 1977) is related to 'through-movement'. It measures the degree to which each segment lies on routes between all other pairs of segments: the passing potential for a segment.

As later noticed by Choi (2013) the recreational or leisure-based active travels do not have a fixed destination point, but take place based on the quality-influencing aspect of the environment, thus the 'through-movement' in this case is the more relevant aspect. Therefore, the betweenness centrality has been chosen as one of the indicators, which takes account of street network configuration, for the proposed Runability Index.

The Space Syntax approach follows a dual representation of street networks where streets are turned into nodes and intersections into links. There are two ways of constructing such a graph – one is through representing streets as axial lines, i.e. the horizontal straight lines that one can take before having to make an angular turn to be able to progress. They also stand for distinct visual axial lines and indicate convex spaces. Turner's (2007) work has further argued that the conventional axial representation may be better replaced by another representation based on street centreline segments, which assumes that a street is defined by its centreline and is limited by the intersection points with the other street segments. The urban space network used for this project also corresponds to the latter way of representation.

There are three different cases, which need to be explored while talking about the shortest path between two nodes of such graph. The simplest one, called topological shortest path, stands for the smallest number of graph nodes that need to be passed until a certain node is reached. The metric shortest path means the lowest cost in a weighted graph where network edges are weighted by the distances from one midpoint to another. Similarly, the angular shortest path accounts for the lowest cost in a graph where weights are given based on the azimuth change at intersections.

Considering these aspects, it has been noted by Hillier & Iida (2005) that people tend to minimise the distance between start and destination points regarding all three: metric, topological and geometric angular costs. Later studies by Hillier et al. (2007), however, argue that locations, which appear accessible or remote, depend on human way finding skills and mental conceptualizations of the environment. Thus, the most accessible locations are not those closest to all others in terms of metric distance, but rather those closest in terms of perceived turns. This assumption also fits well with the behaviour of a recreational travel where one does not try to reach a destination in the fastest way possible but tends to follow a simple and conventional route (Choi, 2013). The theory behind it all states that the fewer the number of direction changes throughout a route taken, the more the street configuration will stimulate movement (Crucitti et al., 2006).



Formally, betweenness centrality measures how many times shortest paths pass through a certain node between all pairs of origins and destinations (nodes), where  $g_{jk}(p_i)$  is the number of geodesics between node  $p_j$  and  $p_k$ , which contain node  $p_i$ , and  $g_{jk}$  is the number of all geodesics between  $p_j$  and  $p_k$  as in the following equation (Freeman, 1977):

$$C_B(P_i) = \sum_j \sum_k g_{jk}(p_i) / g_{jk} (j < k)$$

Later work by Hillier et al. (2012) developed this equation further in order to normalise the measure across different radii, independent of the size of urban systems. They argued that the original measure suffered from a paradox of segregating systems generating higher values of betweenness. The newly introduced measure is termed NACH (Normalised Angular Choice). It is based on the concept of the following equation as explained in Hillier et al. (2012). The measure is made up of two components: angular choice (betweenness) and a total angular depth, which measures the sum of the angular costs between every origin to every destination:

$$NACH_r = \frac{\log(ACH_r + 1)}{\log(ATD_r + 3)}$$

where

$NACH_r$  - normalised angular choice at radius  $r$ ,

$ACH_r$  - angular choice at  $r$ ,

$ATD_r$  - total depth at  $r$ .

Another variable of centrality measure is the radius, which can also be classified into the same three categories: angular, topological and metric. The differences between them can be seen in Figure 56. Metric radius corresponds to the geodesic radius around a space centroid. Topological radius accounts for the maximum depth of traversing a graph. Angular radius takes into account the direction change when a 90-degree change is equivalent to one step (Hillier and Iida, 2005).



Figure 56. Difference between the angular (left), topological (middle) and metric (right) radius considered around a segment.

Even it is assumed that investigating different size of the catchment radius corresponds to the different phenomena (i.e. larger radius shows centrality measures for vehicular means of transport, while smaller radius corresponds to walkability), there is no well-established opinion about which radius has to be used in which case. The most common metric radius used for the investigation of neighbourhood walkability ranges from 300 to 1500m (Saelens & Handy, 2008) which corresponds to the supposed minimal walking distance as already discussed in chapter '6.2 Value of Land Use Mix'; and 3000m to  $n$  (all network size) for the vehicular means of transport. Meanwhile, suggested both topological and angular radius ranges from

2 to 7 steps depth (Kutumisu, 2011; Zhou, 2012). Dhanani et al. (2012) in their cross-comparison study have displayed that a common step depth cannot be defined globally suitable for all cases since it is highly dependent not only on the means of transport under investigation but also on the granularity of data, network model, and level of detail of geometric and segmental representation. Therefore, it needs to be calibrated according to every individual situation.

However, the walkable distance neither in metric nor in angular or topological terms is the same as the runnable distance. This is why these measures have been defined based on the actual data from the mobile sports tracking application. The average measures derived from the data can be seen in Table 6. The metric distance has been derived as the average Euclidian distance of the GPS trajectories. The topological distance stands for the average number of urban space network edges passed during one workout. The same counts for the average angular distance.

<b>Measure</b>	<b>Vilnius</b>	<b>Valencia</b>	<b>Gothenburg</b>
<i>Avg. metric distance</i>	5169m	5624m	4973m
<i>Avg. topological distance</i>	35	70	40
<i>Avg. angular distance</i>	4.93	6.42	5.02
<i>Avg. length of USN edge</i>	160,86m	71,06m	113,11m

**Table 6. The average measures of variables, which are used to define catchment radius for local centrality measures calculated for each of the case study cities separately.**

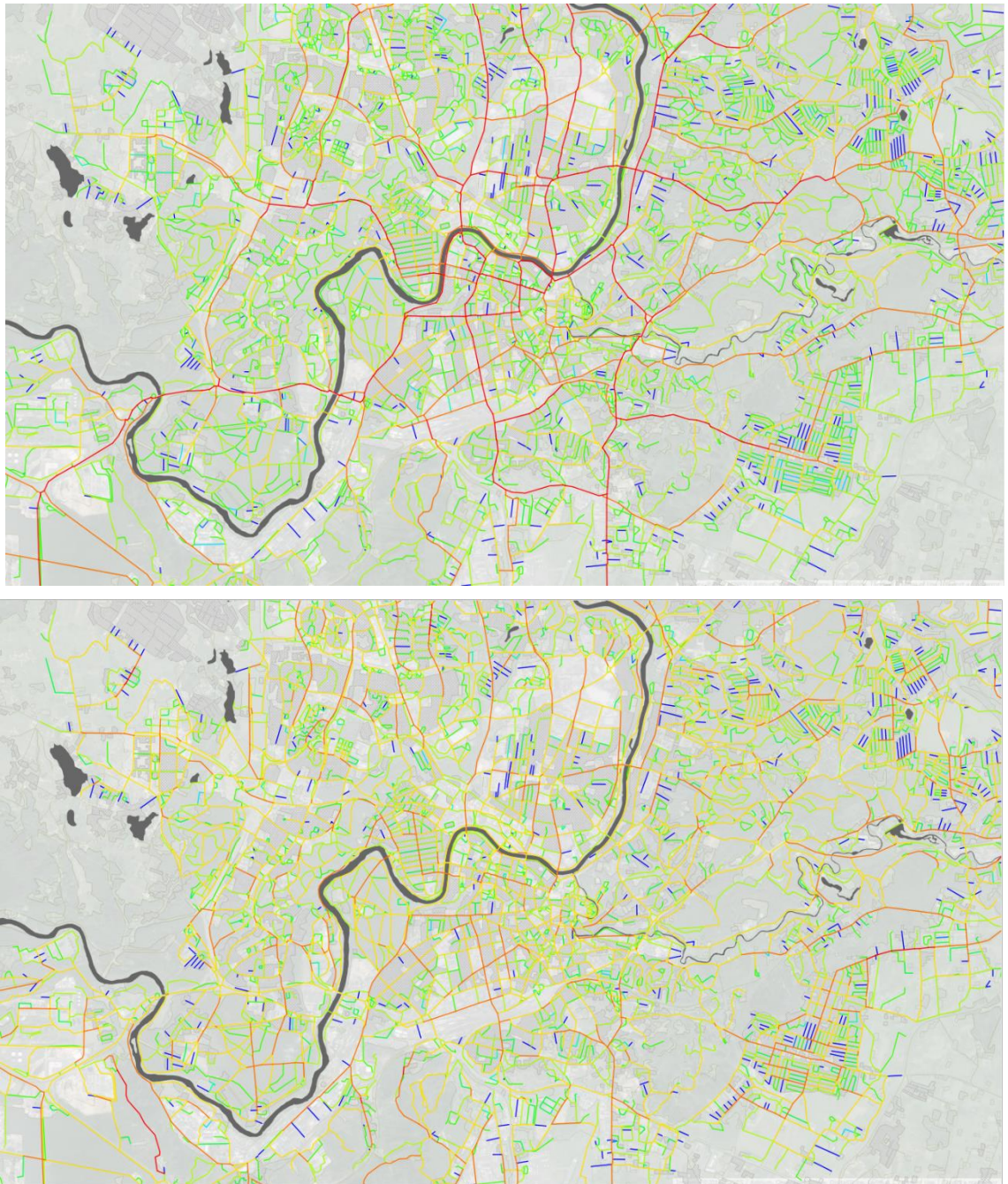
As it can be seen from the table, the average metric distance of a single active recreational travel is roughly the same for all three case study cities while the topological distance varies considerably. This variation is directly related with the average length of an urban space network edge – in case of Valencia where street network is very dense and block size relatively small, topological distance becomes significantly longer than in Vilnius and Gothenburg where street network is scarser and urban blocks are of a bigger size. The same accounts for the angular distance where it is obvious that higher street density yields more possibilities for direction change.

Taken all these measures into account the following numbers have been tested in order to explore which catchment radius best fits for estimating the potential recreational usage of a space (Table 7). Three cases were taken into account – local measure, when only the nearest neighbourhood is considered, walkable radius, which is based on assumption that 1000m is a usual walkable distance that can be covered in 5min and recreational, for which measures have been extracted from the GPS trajectories. For topological radius, 3 has been chosen as local catchment radius for all cities, while walkable and recreational have been calculated as a ratio between the average runnable distance and an average length of a network edge segment. Therefore, these measures differ for each case study city. Since average angular distance of a recreational travel hardly reaches up 3 steps, 2 is used as a walkable value and no local measure is computed.

<b>Catchment radius</b>	<b>Metric</b>	<b>Topological</b>	<b>Angular</b>
<i>Local</i>	500m	3	--
<i>Walkable</i>	1000m	16 for Vilnius 35 for Valencia 23 for Gothenburg	2
<i>Recreational</i>	2500m	21 for Vilnius 50 for Valencia 28 for Gothenburg	3
<i>Global</i>	n		

**Table 7. Tested catchment radius variables for the betweenness centrality measure.**

All network centrality measures have been calculated using Qgis “Space Syntax Toolkit” plug-in that provides a front-end for the UCL “depthmapX” software. Later calculated values are attributed to their urban space networks inside PostGIS. A few of the resulting images of the network centrality evaluation are shown in Figure 57.



**Figure 57. Urban space network of Vilnius coloured according to its NACH values. Top: global radius n, bottom: angular radius of 3 direction changes, blue meaning no betweenness, red - maximum betweenness values.**

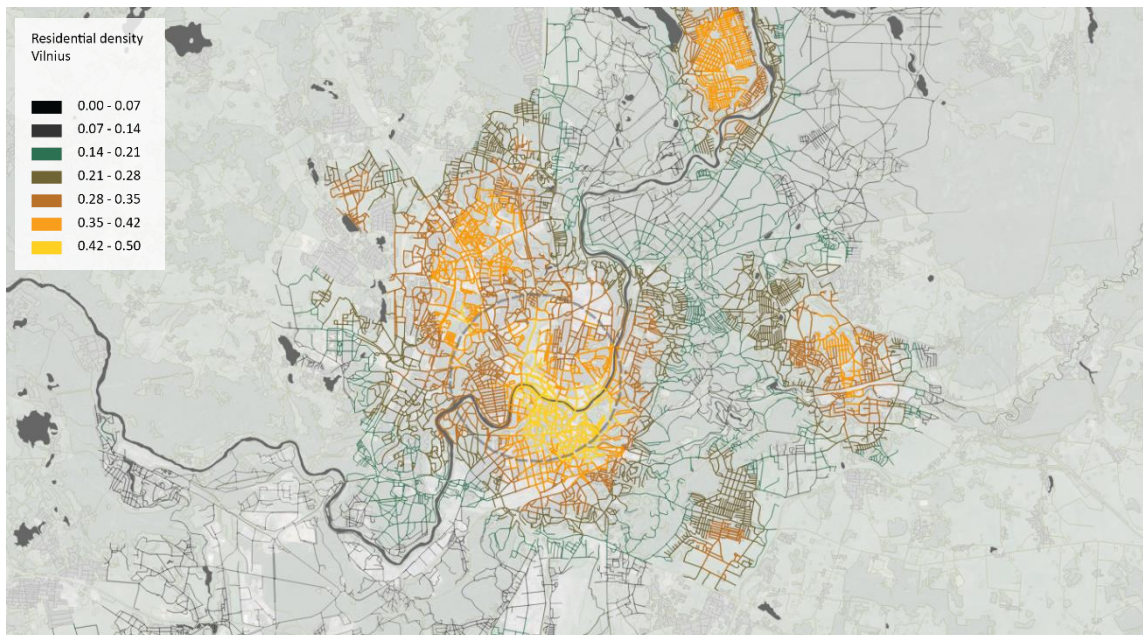


## 6.4 VALUE OF RESIDENTIAL DENSITY

Another measure, which is often incorporated into the calculation of Walkability Index, is the population or residential density in the area. The difference between the two is that population density defines the number of people living in a certain area and residential density defines number of housing units per area unit regardless of how these units are occupied. The deficiency of using such data is the lack of it, since not all European countries are able to provide up-to-date, standardized and comparable datasets (Valencia City Council, 2011, Vilnius.lt, 2015). The latest available pan-European dataset, provided by Eurostat is the GEOSTAT population grid of 1km grid size. The base data used for creating this dataset has been validated in 2003 (efgs.info, 2015). While the data is at least 12 years old and provided in a very rough grid, it can hardly be usable for the purpose of this research.

In order to substitute the population density data, the Urban Audit land use data has been used. The advantage of this data towards the GEOSTAT population grid is that it is provided per each building block, thus at much finer resolution and has a timestamp of 2007, which is out-dated as well, however newer than the population grid. The disadvantage of this data is that it does not provide data about neither number of inhabitants nor number of housing units in a block. The residential density in this case means the ratio of building footprint with the total ground area of the block, where buildings have primary use as residential. Although, generally, on a big scale population density with residential building density should show similar results, on a small scale in case of the same density of soil sealing, private residential houses will give the same results as a skyscraper (European Union, 2011).

The residential density as an attribute of an urban space network edge has been calculated as a sum of built-up areas with residential use buildings over the total area of the 2500m radius around a network edge. The radius of 2500m has been chosen based on the average runnable distance as explained in previous chapters. The relatively large radius has been chosen due to the roughness of data, so that the results would correspond more to the general trend and be less sensitive to the data limitations. The results for Vilnius case can be seen in Figure 58

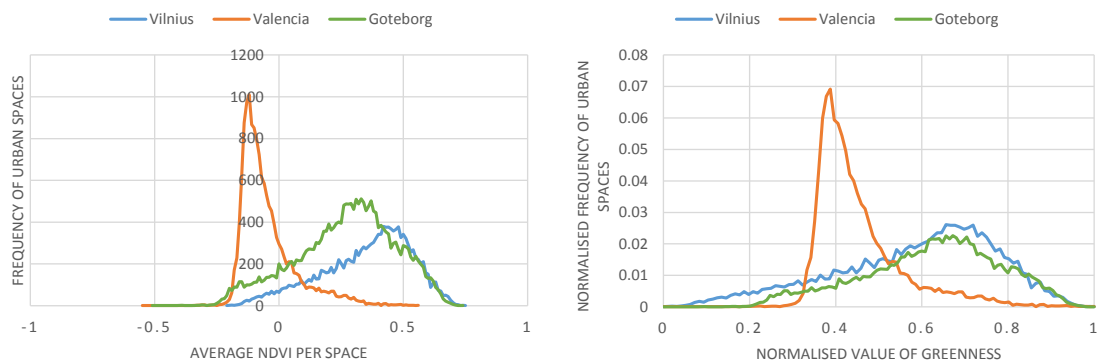


**Figure 58.** Urban space network of Vilnius coloured according to the residential density in 2500m radius around the network edge (grey circle showing the size of the radius).

## 6.5 UNIFICATION OF MEASURES

Similarly as with the actual recreational usage values, all three indicators of the potential recreational usage need to be translated from quantitative to qualitative measures. Differently than in case of actual recreational usage, the indicators of Runability Index do not have a clear threshold value which would mean the definite presence of recreational activities. Therefore, all indicators have been normalised per each city based on their maximum and minimum values accordingly. When normalised, the values do not represent their meaning directly but are mapped into the likeliness for the recreational activity to happen in a certain space. For example, the greenest space of the city is the most likely to attract recreational activities, as well as a street in the densest neighbourhood or in the most balanced area in terms of land use mix.

In particular, value of greenness needs to be normalized since the calculated average NDVI value is dependant on the vegetation intensity which varies throughout different seasons and different geographical locations, i.e. the greenness may yield higher values for humid climate areas and lower for droughty ones. However, in case of potential recreational usage, greenness is rather a perceptual measure which does not depend on the actual vegetation intensity but on the presence of vegetation as such. Furthermore, the boundary between green and non-green environment is fuzzy and depends on the particular situation.



**Figure 59. Histograms of greenness, left: actual frequency of average NDVI values; right: normalised frequency of normalised value of greenness.**

Figure 59 shows how greenness values are differently distributed within all three case study networks before and after normalisation. These differences appear both because of different characteristics of climate and vegetation, and different amount of greenery in cities. In order to mitigate the former cause, the values of greenness have been normalised based on network's own minimum and maximum values. The fact that low NDVI values imply presence of water features can be ignored in this case since urban space network edges do not run on water surfaces. While Vilnius and Gothenburg share similar histograms, Valencia clearly shows less greenery and less even distribution of it throughout the city.



## 6.6 COMPARISON

While the actual recreational usage is a rather straightforward measure obtained from the mobile sports tracking application data, the potential usage value can be calculated in many possible ways. First, both land use mix and NACH already have different variables to be explored and furthermore, the indicators can be combined into a Runability Index using various methods.

One method to combine the measures is by taking an average of them and assuming that for the space to be likely recreational, it needs to have high values of all greenness, land use mix and NACH. A similar approach is using geometric mean, which smoothens the cumulative effect, however, in case one of the values equals to 0, the final result is also 0. Even though physically all the four indicators are of a different nature and scale, and cannot be added into a single measurement, the averaging of values is merely a way to observe the general trend of values.

Another approach is related to hypothesis that a space is more likely to be recreational if it has a high value of either greenness, land use mix, NACH or residential density and contrary, it is less likely to be recreational if it has low value of the either. In this case, fuzzy operators of AND and OR are used which mathematically correspond to the minimum and maximum value of a set respectively. These possibilities result into 108 possible combination for each of the three cities, finally emerging into 324 scatter plots to be tested. Due to this reason, the decision needed to be automated and there the two values were computed: RMSE and R-squared.

RMSE or Root Mean Squared Error measures the difference between two datasets assuming that one of them is predicting the other. The differences between two datasets are also called residuals, and RMSE serves to aggregate them into a single measure of predictive power. The RMSE of a model prediction with respect to the estimated variable is defined as the square root of the mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (X_o - X_m)^2}{n}}$$

where  $X_o - X_m$  is the difference between two corresponding values.

RMSE has the same units as compared values and needs to be interpreted within the range of input values. In case of this research, different RMSE values can be compared between case study cities and variables because the input values have been normalised into the same range. Higher RMSE value means less correspondence between the actual and potential values, and the other way round respectively.

However, while RMSE accumulates the errors, it does not indicate how well outcomes can be predicted by the model. This is done using the R-squared, which is the proportion of variability in a data set that can be explained by the statistical model. This value ranges between 0 and 1, 0 meaning that the model has no ability to predict the phenomena and 1 mean the 100% ability of prediction. While these two measures are rather similar, they can still yield different results especially, when the compared prediction models have low correspondence with the actual data. R-squared is calculated as following:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where  $y_i$  – an observed value,

$f_i$  – predicted or modelled value,

$\bar{y}$  – the mean of the observed data.

In order to explore all the previously mentioned cases for defining the model of the potential recreational usage all calculations have been done following the schema as shown in Pseudocode 4.

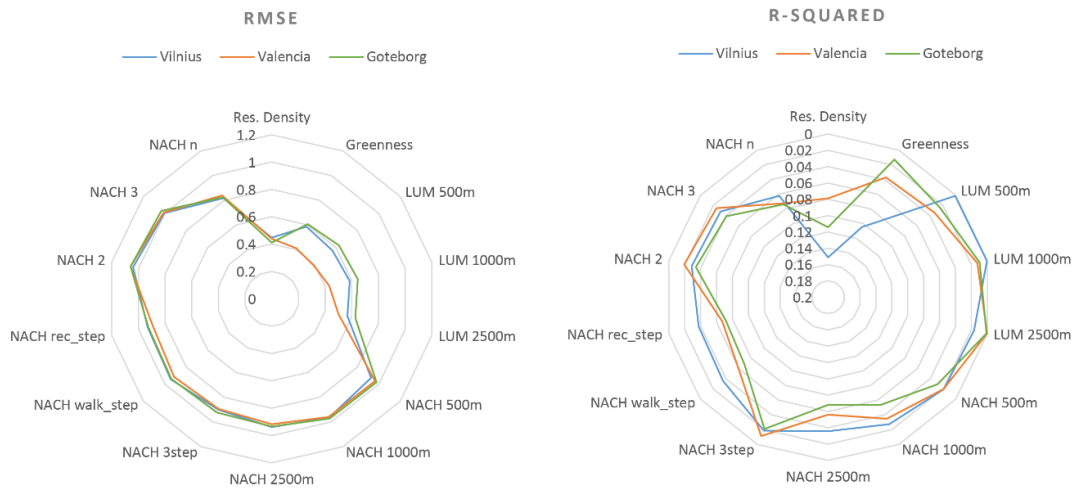
```
actual = normalised(recreational usage)
for city in [Vilnius, Valencia, Goteborg]:
    for operator in [average, geom_mean, AND, OR]:
        for landusemix in [LUM 500m, LUM 1000m, LUM 2500m]:
            for NACH in [NACHr500m, NACHr1000m, NACHr2500m, NACHr3step,
                          NACHrWalk_step, NACHrRec_step', NACHr2, NACHr3, NACH]:
                if operator = average:
                    potential = (greenness+ landusemix + NACH + r_density) / 4
                elif operator = geom_mean:
                    potential = (greenness * landusemix * NACH * r_density)^1/4
                elif operator = OR:
                    potential = max([greenness, landusemix, NACH, r_density])
                elif operator = AND:
                    potential = min([greenness, landusemix, NACH, r_density])
                calculate RMSE
                calculate R-squared
```

**Pseudocode 4. Determination of the most suitable combination of the potential recreational usage indicators in terms of different indicator variables and their aggregation method.**

All the results of comparison have been plotted in radar charts for easier evaluation as can be seen in Figure 61. All the numbers of the results can be seen in Appendix C.

Additionally, the same type of results have been plotted for every measure of greenness, land use mix, residential density and NACH separately without combining with the other values in order to check if any of them have good correlation with the actual recreational usage on its own (Figure 60).

As it can be seen from all the radar charts, none of the variants has exceptionally satisfying results and all values show low correlation between the actual and potential usage. Nevertheless, some trends can be noticed repeating throughout the results.



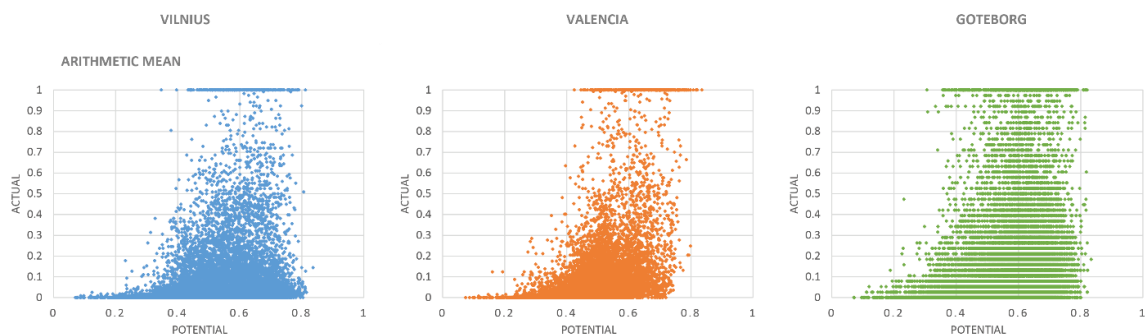
**Figure 60. Radar charts of the correlation measures between the single ingredient variables of the potential recreational usage and the actual recreational usage. Note that R-squared values are inverted in order to provide a simple interface for comparison.**



**Figure 61. Radar charts of comparison results between different data models of potential recreational usage and the actual recreational values. Note that R-square values are inverted in order to provide a simple interface for comparison.**

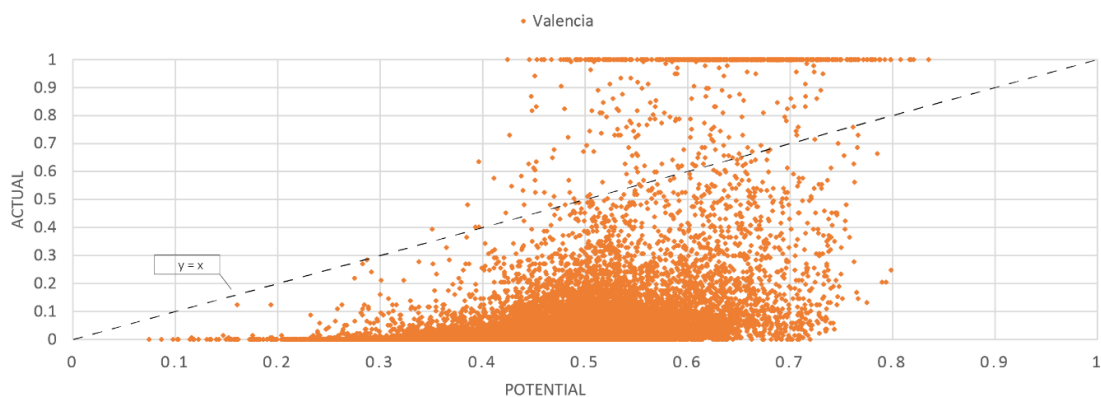
First of all, it is obvious that the city of Valencia has better results than the other case study cities and it is the only case where R-squared shows correlation (R-squared higher than 0.05). In addition, it can be noticed by looking deeper into the charts that out of all three of the land use mix variables, the 500m catchment radius performs slightly better than the larger ones. On the contrary, bigger catchment radii provide better results while looking at the NACH values with the global n radius at the peak. Due to these observations, the combination of greenness, land use mix at 500m radius and NACH at global radius have been chosen for the further investigation. Scatter plots for each of the cities have been explored for each of the value combination methods. All of the scatter plots can be seen in Appendix B. Scatter plots.

Looking at all the scatter plots it can be seen, that the most correlation can be observed when indicators of the potential recreational values are averaged to combine them into a single result (Figure 62).



**Figure 62.** Scatter plots for each of the case study cities, where actual recreational values are compared with the average result of greenness, LUM at 500m and NACH at global n radius.

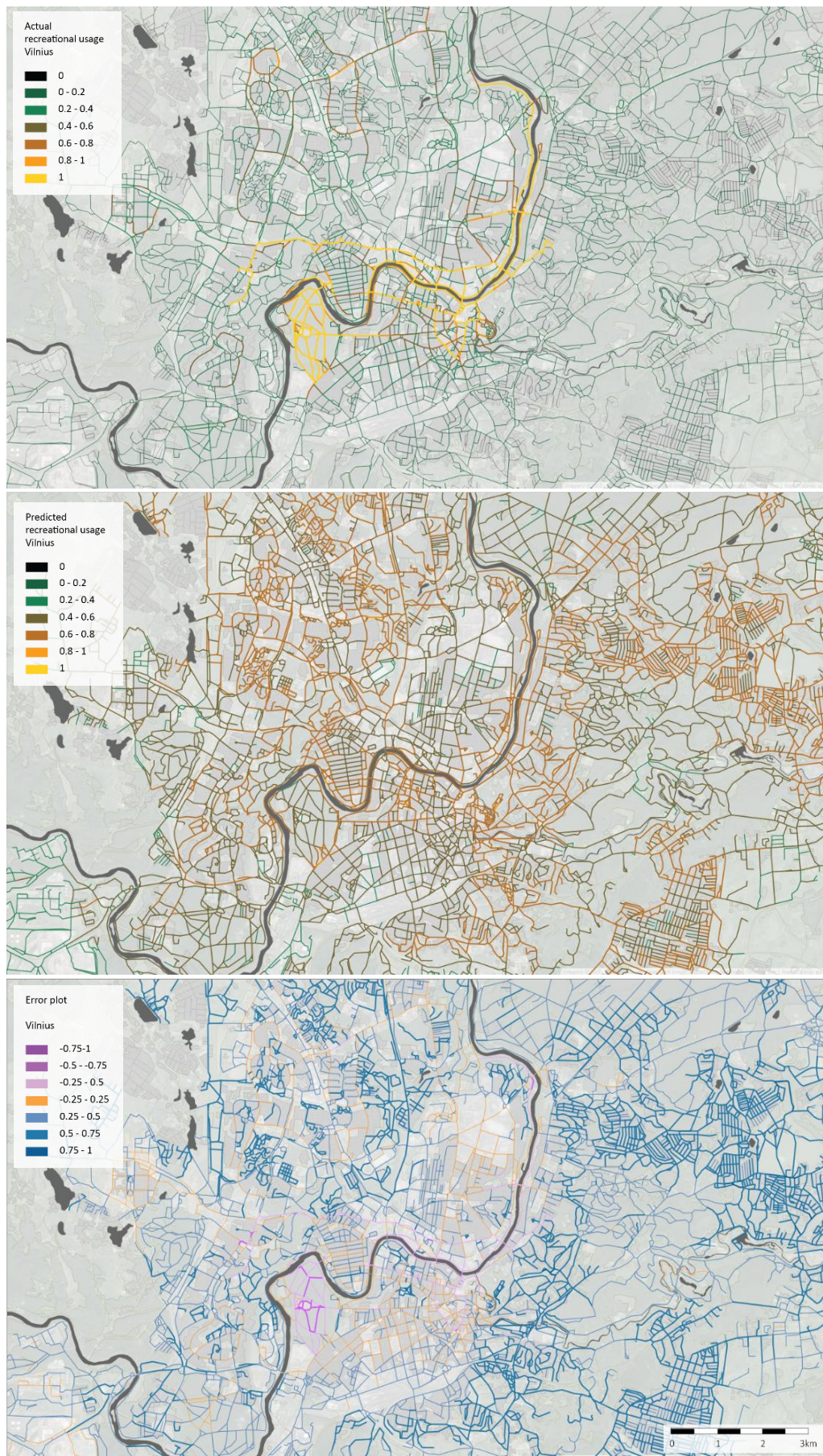
While at first glance, scatter plots seem to provide almost random distribution of data, it must be noticed that if regression line of  $y = x$  is fitted on each of scatter plots, the general tendency is that much more points remain below the line than above (Figure 63). This means that the model quite well explains where the recreational activity does *not* happen, however overestimates where it should happen, i.e. low estimated values do not have high actual values, however high estimated values have low actual ones.



**Figure 63.** Scatter plot of Valencia's case, actual vs. potential recreational values with the regression line of  $y = x$ .

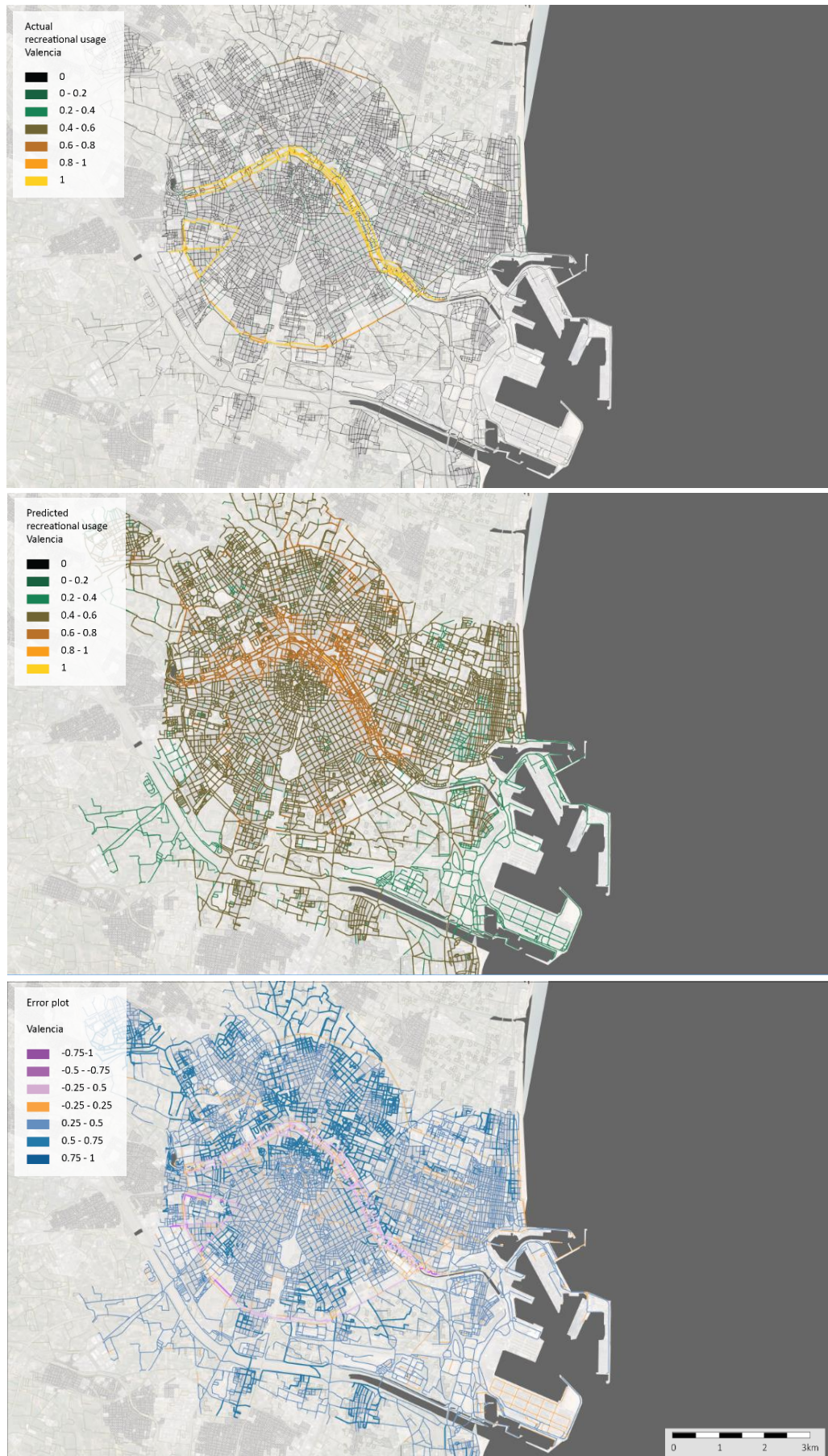
The benefit of having underlying urban space network as an intermediary between the datasets is that both actual and potential recreational values can be plotted on a map. Moreover, an error plot has been also made in a network space so that the deficiencies of estimated model can be easier identified according to the particularities of the built environment. Three versions of urban space network values have been plotted for each of the cities: actual values, estimated potential and an error plot, i.e. difference between the actual and estimated values. The same colour scheme and scale have been used for all case study cities for easier identification ().





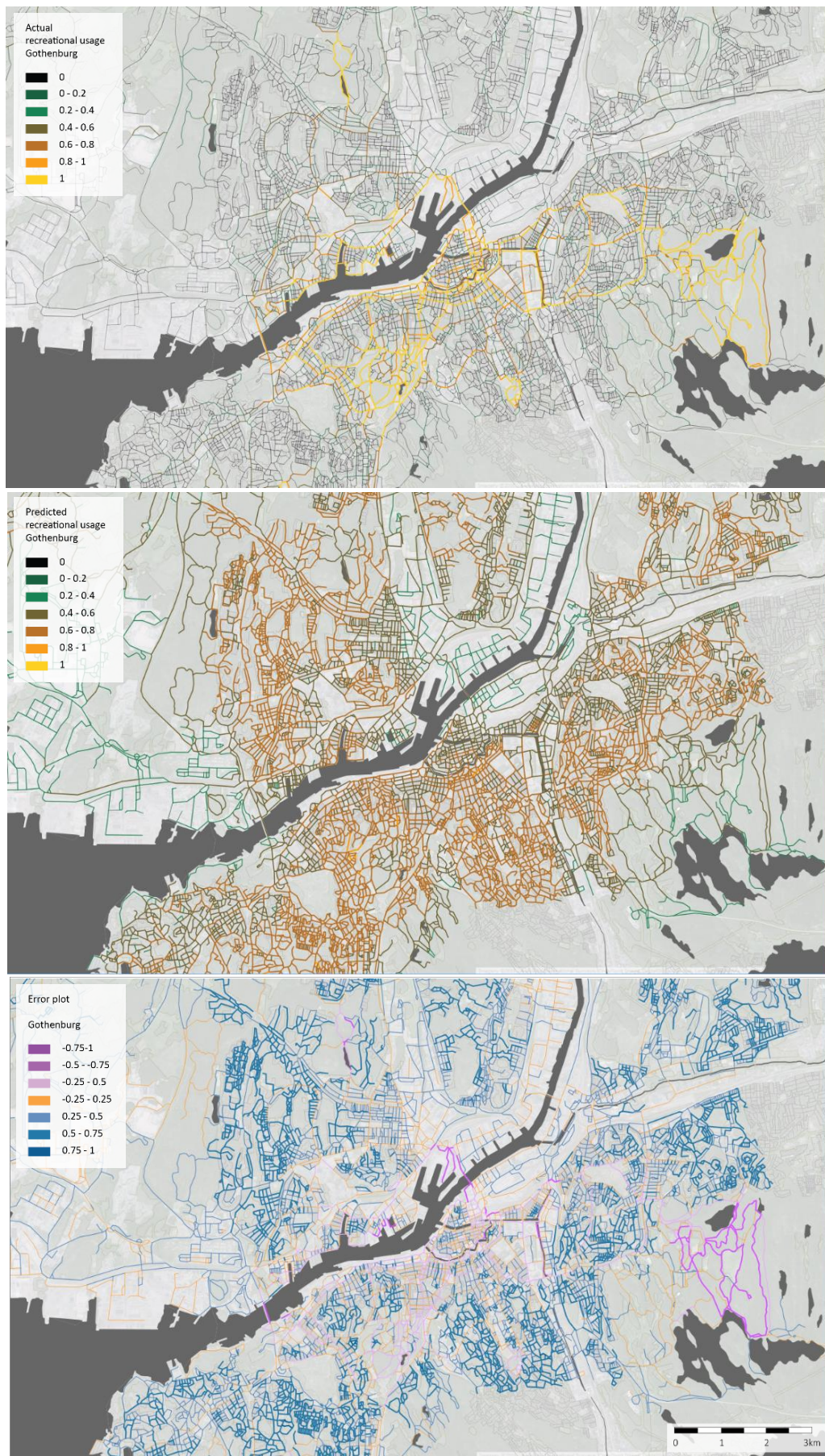
**Figure 64. Urban space network coloured according its actual recreational usage, potential recreational usage and error plot as a difference between the two, Vilnius case.**





**Figure 65.** Urban space network coloured according its actual recreational usage, potential recreational usage and error plot as a difference between the two, Valencia case.





**Figure 66. Urban space network coloured according its actual recreational usage, potential recreational usage and error plot as a difference between the two, Gothenburg case.**

## 6.7 LIMITATIONS

One of the first things that can be noticed while comparing the actual and estimated recreational usage is that actual recreational usage shows very sharp patterns for all three of the case study cities. However, this is not the case for the estimated values. The maps of estimated values tend to give patterns of gradual change within a network while going from one space to another, i.e. the estimated values are similar for neighbouring spaces. However, in reality, if a certain space is used for recreation, the recreational usage is not necessary the same for its neighbours.

Furthermore, it can be seen from the plotted networks, that the model tends to overestimate the outskirts of cities, where there is plenty of green areas, however, less urban fabric. These are the areas called territories-in-between. This is especially visible in case of Gothenburg and Vilnius, since the city of Valencia has much higher urban density than the other ones and the territories-in-between are left behind its boundary. Even though these territories have lower values of residential density, the Runability Index is heightened by high values of greenness and a good balance between residential and recreational areas.

While looking at the map error plots, it can be noticed that mostly coastal areas are underestimated. Even though they are very attractive for recreational activities, their score is mitigated by low values of greenery and the betweenness measure. Such attractive recreational spaces as riverbanks get lower values due to their position in a network. Another area that tends to be underestimated and especially relevant in case of Gothenburg, is industrial territories, such as harbours, which have low land use mix value and moderate greenness, however attract people due to straightforward and continuous routes alongside the banks.

Another critique related to the indicator of betweenness centrality as a measure of street network configuration is that larger catchments radius are mostly associated with such means of transport, which overcome long distances, thus walkability is associated with smaller radius. However, the recreational activities also tend to cover longer distances and the comparison between all different catchments radii has shown that the best correspondence exists when all network is considered. This, however, raises incongruity between suitability of measure for both active recreational travels and motorised means of transport, which must have converse movement patterns and preferences.

Finally, it is obvious from both the maps and the scatter plots that the frequencies of values are non-matching severely (Figure 67). While the model values are almost normally distributed, the actual values form an inverted pattern. This is most probably caused by the averaging of the ingredients of potential recreational usage model, which prevents the final measures from getting extremely high and extremely low final values and smoothen the distribution.

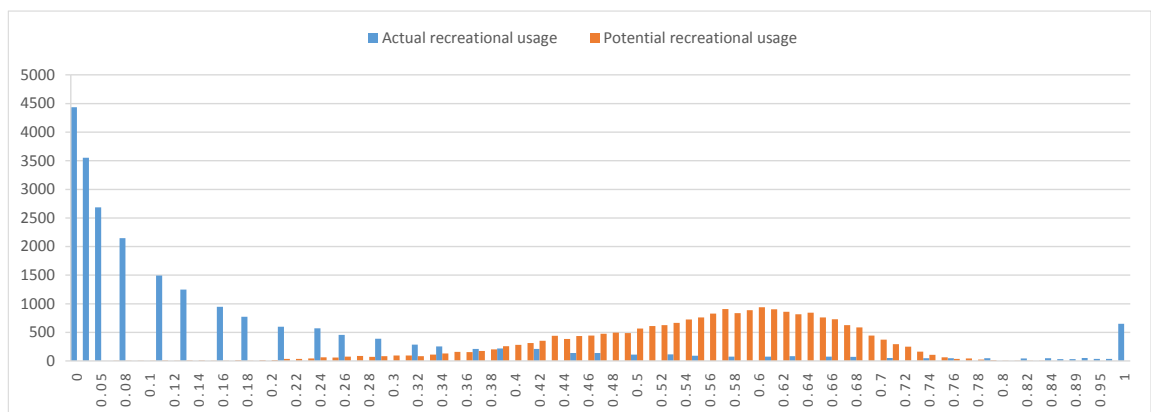


Figure 67. Frequency histogram for the values of actual and estimated potential recreational usage.



## 7. FUTURE WORK AND CONCLUSIONS

---

### 7.1 CONCLUSIONS

The conducted Master Thesis research has investigated how GPS data from mobile sports tracking applications can be used to assess, analyse and model the recreational usage of an urban space network. A reusable automated and non-labour intensive framework has been developed for mobile sports tracking application data acquisition, management, processing and analysis. The Runability Index has been introduced as an indication of space potential to be used for recreation, based on the well-known measure of walkability. Finally, the acquired and processed data from a mobile sports tracking application has been used as a ground truth for the calibration and validation of the Runability Index.

Endomondo sports tracking application has been chosen as a case study due to its popularity rate and relatively convenient open and free data access. A specific data acquisition method has been designed in order to collect GPS data from publicly accessible websites. The method is based on a Ruby script, which is able to access, and read HTML code of every single workout on Endomondo server and choose the required data, which is later filtered, transformed and uploaded in a PostGIS database on the TU Delft server. The developed data acquisition method is fully automated and only minor user interaction is required. Data acquisition process took approximately 1248h and resulted in more than 3.5 million valid GPS tracks of almost a million distinct users within the territory of Europe. Collected data covers a timespan of one year and includes evenly distributed data throughout a day, all seasons, weekdays and public holidays.

Midway through the data acquisition' process three case study cities have been chosen based on the ratio between city's population and a number of distinct users spotted within the territory of a city. All the chosen cities, i.e. Vilnius (Lithuania), Valencia (Spain) and Gothenburg (Sweden) are considered as "extra-large" by the Eurostat Urban Audit project, thus their population at the city core is between half and a million inhabitants. Nevertheless, all cities have distinct characteristics in means of climate, geographic location, urban development, amount and distribution of urban recreational spaces. The population and official territorial city boundaries are determined by the Eurostat Urban Audit project.

Before processing GPS data, a specific method has been developed for automated construction of an urban space network using Open Street Map data complemented with Urban Atlas Road Land Use data. The method relies on integration of datasets, generalisation and simplification through buffering linear features and combining all polygons. Later Segmented Voronoi (Delaunay) Graph is used to extract polygon centreline, which, after minor processing and additional simplification is used as a representation of an urban space network.

The constructed network is relevant for the desired type of analysis and differs from conventional street networks in that it includes paths for both motorised and non-motorised means of transport, which run through urban fabric as well as parks and urban forests. A particular characteristic of the urban space network is that it has low granularity, however, well-preserved space connectivity. Further research process has also proved that network construction method is able to provide topologically valid, clean and simple network, which is easy to handle and use for various algorithms. The chosen granularity (or level of detail) for generalisation and simplification is based on the resolution of later used satellite images and the positional error of GPS in urban environments without additional error correction methods. The urban space network proved an efficient intermediary between all later considered datasets.

Collected GPS tracks have been filtered from blundering fixes and snapped to an urban space network with 85% mapping accuracy using algorithms developed explicitly for this research. GPS tracks when



aggregated per single network edge form a measure, which is regarded as the actual recreational usage. The value is later normalised using fuzzy normalisation methods in order to be comparable with the developed Runability Index. The assessment method for identifying spaces of high and low recreational usage proved to be not biased by the choice of sports tracking application, neither by the timestamp of data, and was verified not to lose its quality during the further processing.

The model for estimating potential recreational usage has been constructed based on the aspiring measure of Walkability Index. The most commonly used indicators of the Walkability Index have been adapted in order to fit the characteristics of active recreational travels. The chosen indicators namely were greenness, land use mix, network centrality and residential density. The newly constructed value has been called a Runability Index. The index is supposed to be used as an indicator of space potential for active recreational travels.

The value of greenness for every single urban space (edge of a network) has been computed as the average NDVI value of the pixels intersected by an edge under investigation using Landsat 8 satellite images as source data and normalised for each of the case study cities. The value of land use mix has been calculated as an entropy score of residential and recreational land use areas within a chosen radius from the middle point of an urban space network edge. Three different radii have been considered. Finally, the value of network centrality has been calculated as a normalised angular choice (NACH) measure based on 9 different radii. 4 combination methods, namely arithmetic and geometric mean and fuzzy operators 'OR' and 'AND' have been considered for aggregating values into a single value of Runability Index.

108 different variants of Runability Index for each of the 3 case study cities have been tested in correspondence with the values of actual recreational usage obtained from the sports tracking application data. Based on the best values of correlation (RMSE and R-squared) the best-fit formula appeared to be an average of greenness, residential density, land use mix in the radius of 500m and a global NACH value.

Later investigations of the scatter plots, frequency histograms and error plots on a map have shown that the constructed Runability Index works well while indicating which spaces are not used for recreational travels however fails at predicting which of them are. The developed model tends to overestimate urban territories-in-between and underestimate such areas as central urban parks and especially coastlines. Furthermore, the frequencies of the estimated and the actual values severely differ, which shows that the Runability Index needs to be defined differently. Nevertheless, by looking at the correlation values between the different measures, it is obvious that such characteristics of the built environment as land use, greenness and street configuration have different impact on active recreational than on transport-based travels.

Finally, testing all processes and algorithms in parallel for three different case studies has ensured that the collected data as well as the developed methods would not be dependent on a specific urban structure and can be repeated for any of the European cities with sufficient application users. The process, however, ascertained that built environment differences between the cities have a high impact on the phenomena under investigation and therefore global explanations without individual calibrations are hardly possible.

To sum up, the Master Thesis project has developed a method of collecting, managing and processing publicly accessible GPS data from mobile sports tracking applications, which can be used for both investigating the usage of urban space network for recreational purposes and for using the data as a ground truth for validation and calibration of estimated potential recreational usage models. Finally, the developed Runability Index based on the well-known Walkability Index did not show satisfying correlation results while compared with the actual values, however demonstrated that leisure-based and transport-based active travels form different patterns throughout the city and cannot be predicted using the same indicators.

Based on the research results, a discussion and a number of recommendations have been provided for future research.

## 7.2 DISCUSSION AND RECOMMENDATIONS

To begin with, even though a robust and reliable data acquisition algorithm has been constructed in order to obtain sufficient amount of data only with minor human interaction, the whole data acquisition process took 1248h of continuous accessing, querying and copying a part of the publicly accessible Endomondo data. Furthermore, the data needed to be collected all over the world, even though only three cities were used as case studies. This tedious process of data acquisition could be avoided if Endomondo allowed accessing and querying their database; however, they refuse to do so due to privacy issues and that is a completely valid reason. Notwithstanding, a collaboration between a sports tracking application and a researcher would improve the efficiency of data acquisition.

Another important point, which needs to be explored while using mobile sports tracking application data as a ground truth, is the influence of bias caused by the certain characteristics of application users. Knowing such characteristics as user age group, occupation, education, social status, ethnicity, etc. might give a better overview of data validity and allow deeper investigation of recreational travel patterns. Currently user group analysis is not possible due to the privacy matters.

Another method limitation caused by external factors is the lack of data. Even though the integration of Urban Atlas and Open Street Map datasets improves the completeness of an urban space network, a number of paths and connections remain unknown. This problem, though, in a future research could be tackled by upgrading the GPS network-snapping algorithm. The missing paths could be added to the constructed urban space network based on the clusters of GPS tracks. This would definitely improve the mapping accuracy of the snapping algorithm itself, and may have positive influence on the analysis of network centrality. Finally, the data of residential density could be replaced with the population density data for more accurate results.

Furthermore, the construction of an urban space network could be enhanced by introducing varying buffer width for different kinds of paths. While primary or secondary streets tend to occupy much wider space, the standard 30m buffer does not always unite all representative lines into a single area, however, footpaths in a park may all merge into a single line because of narrow gaps in between. However, on the other hand, since the research is aiming to investigate non-motorised means of transport some heavy traffic roads should also be considered as barriers as well as rivers or ditches, so that only certain connections through them would be possible. Another problem caused by buffering is that it might connect spaces, which actually do not reach in reality due to topography, water features or any other kind of obstacles, including heavy traffic roads as well.

Currently, the running and walking activities have been considered equally without differentiating between them. On the one hand, they might also have different movement patterns, which could be explored separately, while on the other hand, various other types of recreational travels could be added among which recreational cycling, orienteering, roller skiing, skateboarding, etc.

Next point of discussion is the Runability Index itself, which is intended to indicate how much a certain space is suitable for the active recreational travels. As it has turned out from the comparison of theoretically defined and justified values and the actual recreational usage values obtained from the sports tracking application data, almost no correlation exists between the two. There might be a few reasons for this. One of the reasons is the lack of indicators included into the prediction model. Only four factors have been considered for defining Runability Index, however literature research has demonstrated that such factors as microclimatic characteristics, traffic, air pollution, presence of resting places, surface cover, width of a sidewalk, safety, etc. have influence on route choice for active recreational travels.

Exploring three diverse case study cities in parallel has demonstrated that the structure of city's urban network has influence on its recreational usage. Smaller amount of spaces suitable for recreation tend to concentrate people that way achieving higher recreational usage values, while bigger amount of qualitative spaces share citizens between them and the usage values become lower. This knowledge must be taken into account while constructing the Runability Index.

On the contrary, it must be acknowledged that human perception of the environment is not always compliant with the observed physical environment and might be influenced by factors, which are hard to be modelled. Even more, the revealed behaviour – what is actually happening – is not necessarily the same as preferred behaviour – what would be happening given a desired set of alternatives. That is, in case of an active recreational travel, it is rather a choice of 'if to go or not', than 'how to get there where I need to go' which is mostly the case for the transport-based travels. Thus if Walkability Index aims to find where in a network people are most likely to go, the Runability Index has an extra challenge of explaining 'whether or not' people are likely to use the certain space for recreation.

The same pattern can be noticed while looking at the provided heat maps, map error plots and frequency histograms. It is not only the attractiveness of a single space that enables presence of recreational activity but also its position in a broader network of spaces; not in a sense of being in an attractive area but in a sense of being connected to other attractive spaces. Therefore, these findings suggest that, in contrast to the Walkability Index, while defining the Runability Index it is the network-based analysis that must play bigger role than the neighbourhood-based analysis.

## REFERENCES

---

- Abraham S. and Lal P.S. (2010). Trajectory similarity of network constrained moving objects and applications to traffic security. In Chen H, Chau M, Li S-H, Urs S, Srinvasa S, and Wang G A (eds) *Intelligence and Security Informatics: Proceedings of the Pacific-Asia Workshop on Intelligence and Security Informatics 2010*. Berlin, Springer Lecture Notes in Computer Science Vol. 6122: 31–43
- Alm, E. L., Malbert, B., & Korhonen, P. (2002). The Göteborg Case Study.
- Anderson E. (2012). "Why is GPS data sometimes inaccurate?" Strava. Accessed 13 Jan 2015.  
<https://strava.zendesk.com/entries/21443922-Why-is-GPS-data-sometimes-inaccurate->
- Auld, J., Williams, C., & Mohammadian, K. (2008). Prompted recall travel surveying with GPS. In 2008 Transport Chicago Conference, (p. 16).
- Barsukov N. (2014). Generating running route maps. Accessed 01 Dec 2014.  
<http://barsukov.net/programming/2014/07/26/endomondo-code.html>
- Bell, J. F., Wilson, J. S., & Liu, G. C. (2008). Neighborhood greenness and 2-year changes in body mass index of children and youth. *American journal of preventive medicine*, 35(6), 547-553.
- Béra, R. & Claramunt, C. (2004). Can relative adjacency contribute to space syntax in the search for a structural logic of the city? *Proceedings of 3rd International Conference on Geo-graphical Information Science*, Springer Berlin / Heidelberg, LNCS, Vol. 3234/2004, 38-50.
- Bhat, C., & Lockwood, A. (2004). On distinguishing between physically active and physically passive episodes and between travel and activity episodes: an analysis of weekend recreational participation in the San Francisco Bay area. *Transportation Research Part A: Policy and Practice*, 38(8), 573-592.
- Biljecki, F. (2010). Automatic segmentation and classification of movement trajectories for transportation modes (Master Thesis, TU Delft, Delft University of Technology).
- Blanchard, P., & Volchenkov, D. (2008). *Mathematical analysis of urban spatial networks*. Springer Science & Business Media.
- Brandt, J. W., & Algazi, V. R. (1992). Continuous skeleton computation by Voronoi diagram. *CVGIP: Image understanding*, 55(3), 329-338.
- Braw, E. (2013). "Hamburg's answer to climate change" *The Guardian*. Accessed 08 Jan 2015.  
<http://www.theguardian.com/sustainable-business/hamburg-answer-to-climate-change>
- Bretagnolle, A., Daudé, E., & Pumain, D. (2006). From theory to modelling: urban systems as complex systems. *Cybergeo: European Journal of Geography*.
- Brown, B. B., Yamada, I., Smith, K. R., Zick, C. D., Kowaleski-Jones, L., & Fan, J. X. (2009). Mixed land use and walkability: Variations in land use measures and relationships with BMI, overweight, and obesity. *Health & Place*, 15(4), 1130-1141.
- Brownson, R. C., Hoehner, C. M., Day, K., Forsyth, A., & Sallis, J. F. (2009). Measuring the built environment for physical activity: state of the science. *American journal of preventive medicine*, 36(4), S99-S123.
- Cervero R., Kockelman K. (1997). "Travel demand and the 3ds: density, diversity, and design". *Transportation Research Part D*. 1997;2(3): p. 199–219.



Chin GKW, Van Niel KP, Giles-Corti B, Knuiman M (2008) Accessibility and connectivity in physical activity studies: The impact of missing pedestrian data. *Preventive Medicine* 46:41–45. doi: 0.1016/j.ypmed.2007.08.004

Choi, E. (2013). Understanding Walkability: Dealing with the complexity behind pedestrian behavior. In 9th International Space Syntax Symposium, Seoul, Sejong University 2013. Sejong University.

Cohen DA, McKenzie TL, et al. (2007). Contribution of public parks to physical activity. *American Journal of Public Health*; 97(3):509-514.

Cohen, D. A., Ashwood, J. S., Scott, M. M., Overton, A., Evenson, K. R., Staten, L. K., ... & Catellier, D. (2006). Public parks and physical activity among adolescent girls. *Pediatrics*, 118(5), e1381-e1389.

Crucitti, P., Latora, V., & Porta, S. (2006). Centrality in networks of urban streets. *Chaos: an interdisciplinary journal of nonlinear science*, 16(1), 015113.

Cullberg M., Montin S., Tahvilzadeh N. (2014). Urban Challenges, Policy and Action in Gothenburg --- GAPS project baseline study. Mistra Urban Futures.

Cutumisu, N. (2011). Movement-Attractors and Generic Neighbourhood Environment Traits (MAGNET).

Dhanani, A., Vaughan, L. S., Ellul, C., & Griffiths, S. (2012). From the axial line to the walked line: Evaluating the utility of commercial and user-generated street network datasets in space syntax analysis.

Douglas D. & Peucker T. (1973), "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature", *The Canadian Cartographer* 10(2), 112–122

efgs.info (2015). European Forum for Geography and Statistics. GEOSTAT Open Data. <http://www.efgs.info/data/geostat/open-data> Accessed 03 Jun 2015.

Endomondo.com (2015). Accessed 12 Jan 2015. [www.endomondo.com](http://www.endomondo.com)

English Nature, (2005). The English Nature Website. Accessed 01 Dec 2014. <http://www.english-nature.gov.uk>.

Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2), 189-200.

European Union (2011). Mapping Guide for a European Urban Atlas. Available at: <http://www.eea.europa.eu/data-and-maps/data/urban-atlas#tab-methodology>

European Urban Audit (2007). State of European Cities Report. Adding value to the European Urban Audit. Study contracted by the European Commission.

Ewing R. et al. (2009). "Traffic generated by mixed use developments: a six-region study using consistent built environmental measures". Paper presented at the annual Meeting of the transportation research board; January 12, 2009; Washington, dc.

Ewing R., Cervero R. (2001). Travel and the built environment. *Transportation Research Record*. 2001;1780: p. 87–114.

Ferrari, L., & Mamei, M. (2011). Discovering city dynamics through sports tracking applications. *Computer*, 44(12), 63-68.

- Ferrari, L., & Mamei, M. (2013). Identifying and understanding urban sport areas using Nokia sports tracker. *Pervasive and Mobile Computing*, 9(5), 616-628.
- Floyd, M. F., Spengler, J. O., Maddock, J. E., Gobster, P. H., & Suau, L. J. (2008). Park-based physical activity in diverse communities of two US cities: an observational study. *American Journal of Preventive Medicine*, 34(4), 299-305.
- Frank, L.D., Sallis, J.F., Conway, T.L., Chapman, J.E., Saelens, B.E., Bachman, W., (2006). Many pathways from land use to health: associations between neighbourhood walkability and active transportation, body mass index, and air quality. *Journal of the American Planning Association* 72 (1), 75–87.
- Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35-41.
- Gauvin, L., Richard, L., Craig, C. L., Spivock, M., Riva, M., Forster, M., ... & Potvin, L. (2005). From walkability to active living potential: an “ecometric” validation study. *American journal of preventive medicine*, 28(2), 126-133.
- Gebel K, Bauman A.E., Petticrew M. (2007). “The physical environment and physical activity: a critical appraisal of review articles”. *American Journal of Preventive Medicine*. 2007;32: p. 361–369.
- Gil, J. (2014). Analyzing the Configuration of Multimodal Urban Networks. *Geographical Analysis*, 46(4), 368-391.
- Giles-Corti, B., Broomhall, M. H., Knuiman, M., Collins, C., Douglas, K., Ng, K. & Donovan, R. J. (2005). Increasing walking: how important is distance to, attractiveness, and size of public open space?. *American journal of preventive medicine*, 28(2), 169-176.
- Giles-Corti, B., Macaulay, G., Middleton, N., Boruff, B., Bull, F., Butterworth, I., ... & Christian, H. (2015). Developing a research and practice tool to measure walkability: a demonstration project. *Health promotion journal of Australia*, 25(3), 160-166.
- Girres J-F, Touya G (2010) Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS* 14:435–459. doi: 10.1111/j.1467-9671.2010.01203.x
- Gómez, F., Jabaloyes, J., Montero, L., De Vicente, V., & Valcuende, M. (2010). Green areas, the most significant indicator of the sustainability of cities: Research on their utility for urban planning. *Journal of Urban Planning and Development*, 137(3), 311-328.
- Gordon-Larsen P, et al. (2006). “Inequality in the built environment underlies key health disparities in physical activity and obesity”. *Paediatrics*. 2006;117(2): p. 417–424.
- Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning B: Planning and design* 37 (4), 682-703.
- Handley, J., Pauleit, S., Slinn, P., Barber, A., Baker, M., Jones, C., Lindley, S., (2003). Accessible natural green space standards in towns and cities: a review and toolkit. English Nature research report number 526. English Nature, Peterborough.
- Harrison, C., Burgess, J., Millward, A., Dawe, G., (1995). Accessible natural green space in towns and cities: a review of appropriate size and distance criteria. English Nature research report number 153. English Nature, Peterborough.

- Hassan A. Karimi & Piyawan Kasemsuppakorn (2013) Pedestrian network map generation approaches and recommendation, *International Journal of Geographical Information Science*, 27:5, 947-962
- Henderson, K. A. (2005). Parks and physical activity. *Parks and Recreation*, 40(8), 20-26.
- Hillier B. and Hanson J. (1984), *The Social Logic of Space*, Cambridge: Cambridge University Press.
- Hillier B., & Stutz C. (2005) New methods in space syntax. *World architecture* (11), 54-55.
- Hillier, B., & Iida, S. (2005). Network and psychological effects in urban movement. In *Spatial information theory* (pp. 475-490). Springer Berlin Heidelberg.
- Hillier, B., Turner, A. Yang, T., Park, H. (2007). Metric and topo-geometric properties of urban street networks: some convergences, divergences and new results, *Proceedings 6th International Space Syntax Symposium*, ITU, Istanbul, Turkey, 12-15 June 2007.
- Hillier, W. R. G., Yang, T., & Turner, A. (2012). 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.
- Yamada, I., Brown, B. B., Smith, K. R., Zick, C. D., Kowaleski-Jones, L., & Fan, J. X. (2012). Mixed land use and obesity: an empirical comparison of alternative land use measures and geographic scales. *The Professional Geographer*, 64(2), 157-177.
- Yang, J. S., Kang, S. P., & Chon, K. S. (2005). The map matching algorithm of GPS data with relatively long polling time intervals. *Journal of the Eastern Asia Society for Transportation Studies*, 6, 2561-2573.
- James, P., Banay, R. F., Hart, J. E., & Laden, F. (2015). A Review of the Health Benefits of Greenness. *Current Epidemiology Reports*, 2(2), 131-142.
- Jang, B. K., & Chin, R. T. (1990). Analysis of thinning algorithms using mathematical morphology. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(6), 541-551.
- Jiang, B. (2009). Ranking spaces for predicting human movement in an urban environment. *International Journal of Geographical Information Science*, 23(7), 823-837.
- Jiang, B., & Claramunt, C. (2004). A structural approach to the model generalization of an urban street network\*. *Geoinformatica*, 8(2), 157-171.
- Kaczynski, A. T., Potwarka, L. R., & Saelens, B. E. (2008). Association of park size, distance, and features with physical activity in neighborhood parks. *American Journal of Public Health*, 98(8), 1451.
- Kaczynski, A.T., Henderson, K.A., (2007). Environmental correlates of physical activity: a review of evidence about parks and recreation. *Leisure Sciences* 29, (4), 315–354.
- Klaasen, I. T. (2003). Knowledge-based design: developing urban & regional design into a science (Doctoral dissertation, TU Delft, Delft University of Technology).
- Klir, G., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic* (Vol. 4). New Jersey: Prentice Hall.
- Kobayashi, T., & Miller, H. (2014). Exploratory visualization of collective mobile objects data using temporal granularity and spatial similarity. In *Data Mining for Geoinformatics* (pp. 127-154). Springer New York.

- Koohsari, M. J., Mavoa, S., Villanueva, K., Sugiyama, T., Badland, H., Kaczynski, A. T., ... & Giles-Corti, B. (2015). Public open space, physical activity, urban design and public health: Concepts, methods and research agenda. *Health & place*, 33, 75-82.
- Köppen climate classification Dfb. Kottek, M., J. Grieser, C. Beck, B. Rudolf, and F. Rubel (2006). "World Map of the Köppen-Geiger climate classification updated". *Meteorol. Z.*
- Krebs, C.J., 1999. *Ecological Methodology*, second ed Addison-Wesley, New York.
- Lin, H., Sun, G., & Li, R. (2015). The Influence of Built Environment on Walking Behavior: Measurement Issues, Theoretical Considerations, Modeling Methodologies and Chinese Empirical Studies. In *Space-Time Integration in Geography and GIScience* (pp. 53-75). Springer Netherlands.
- Lynch K. (1960). *The image of the city*. Cambridge, MIT Press.
- Lindsay, G. (2010). "Driving makes you fat, urban sprawl bankrupts you and other life-saving new urbanist epiphanies". *FastCompany*, Accessed 05 Dec 2014, <http://www.fastcompany.com/1650173/driving-makes-you-fat-urban-sprawl-bankrupts-you-other-life-saving-new-urbanist-epiphanies>
- Liu X., Jiang B. (2010), Defining and generating axial lines from street center lines for better understanding of urban morphologies, *International Journal of Geographical Information*, forthcoming.
- Lopez, R. (2004). Urban sprawl and risk for being overweight or obese. *American Journal of Public Health*, 94(9), 1574.
- Lwin, K. K., & Murayama, Y. (2011). Modelling of urban green space walkability: Eco-friendly walk score calculator. *Computers, Environment and Urban Systems*, 35(5), 408-420.
- Marchal F., Hackney J. and Axhausen K.W. (2004). Efficient map-matching of large GPS data sets - Tests on a speed monitoring experiment in Zurich. *Arbeitsbericht Verkehrs- und Raumplanung*, Institut für Verkehrsplanung und Transportsysteme, ETH Zurich, Zurich
- Maroko, A. R., Maantay, J. A., Sohler, N. L., Grady, K. L., & Arno, P. S. (2009). The complexities of measuring access to parks and physical activity sites in New York City: a quantitative and qualitative approach. *International journal of health geographics*, 8(1), 1-23.
- Maropoulos, P. G., & Ceglarek, D. (2010). Design verification and validation in product lifecycle. *CIRP Annals-Manufacturing Technology*, 59(2), 740-759.
- Miranda Carranza, P., & Koch, D. (2013). A Computational Method For Generating Convex Maps Using the Medial Axis Transform. In *9th International Space Syntax Symposium*, Seoul, October 31-November 3, 2013 (pp. 064-1). Sejong University Press.
- Modsching, M., Kramer, R., & ten Hagen, K. (2006, March). Field trial on GPS Accuracy in a medium size city: The influence of built-up. In *3rd Workshop on Positioning, Navigation and Communication* (pp. 209-218).
- Mooney, P. (2015). An Outlook for OpenStreetMap. In *OpenStreetMap in GIScience* (pp. 319-324). Springer International Publishing.
- Mora, A. M., & Squillero, G. (Eds.). (2015). *Applications of Evolutionary Computation: 18th European Conference, EvoApplications 2015, Copenhagen, Denmark, April 8-10, 2015, Proceedings* (Vol. 9028). Springer.

- Nagar, A., & Tawfik, H. (2007). A Multi-Criteria Based Approach to Prototyping Urban Road Networks. *Issues in Informing Science and Information Technology*, 4.
- NASA.gov (2015). Landsat 8 science data user's handbook. [landsathandbook.gsfc.nasa.gov/handbook.html](http://landsathandbook.gsfc.nasa.gov/handbook.html).
- NASA.gov (2015). Measuring Vegetation. <http://earthobservatory.nasa.gov/Features/MeasuringVegetation/> Accessed 14 Apr 2015.
- New London Landscape (2012). Green Infrastructure Ideas inspired by High Line for London competition. Accessed 01 Jan 2015, <http://www.newlondonlandscape.org/>
- NRPA (2014). Parks and Recreation in Underserved Areas: A Public Health Perspective. Available at: <http://www.nrpa.org/research-papers/>
- Ogniewicz, R. L. (1992). Discrete Voronoi Skeletons (Doctoral dissertation, Diss. Techn. Wiss. ETH Zürich, Nr. 9876, 1992. Ref.: O. Kübler; Korref.: T. Pun).
- Oksanen J, Suvanto S & D Eränen (2013). Project SUPRA: Looking for routes from massive workout data. Workshop on Analysis & Visualization of MOVement, Geoviz Hamburg, March 4-5, 2013.
- Owen, N., Humpel, N., Leslie, E., Bauman, A., Sallis, J.F., (2004). Understanding environmental influences on walking: review and research agenda. *American Journal of Preventive Medicine* 27 (1), 67–76.
- Piórkowski, M. (2009, June). Sampling urban mobility through on-line repositories of GPS tracks. In *Proceedings of the 1st ACM international workshop on hot topics of planet-scale mobility measurements* (p. 1). ACM.
- Portugali, J. (2011). *Complexity, cognition and the city*. Springer.
- Powell I.M., Slater S., Chaloupka F.J., Harper D. (2006). "Availability of physical activity–related facilities and neighbourhood demographic and socioeconomic characteristics: a national study". *American Journal of Public Health*. 96: p. 1676–1680.
- Pucher J., Buehler R. (2008). "Making cycling irresistible: lessons from the Netherlands, Denmark, and Germany". *Transport Reviews*. 28(4): p. 495–528.
- Qiuping L., Hongchao F., Xuechen L., Bisheng Y. & Lin L. (2014) Polygon-based approach for extracting multilane roads from OpenStreetMap urban roadnetworks, *International Journal of Geographical Information Science*, 28:11, 2200-2219,
- Quddus, M., Washington S. (2015). Shortest path and vehicle trajectory aided map-matching for low frequency GPS data. *Transport. Res. Part C*, <http://dx.doi.org/10.1016/j.trc.2015.02.017>
- Sadahiro, Y., Lay, R., & Kobayashi, T. (2013). Trajectories of moving objects on a network: detection of similarities, visualization of relations, and classification of trajectories. *Transactions in GIS*, 17(1), 18-40.
- Saelens, B.E., Handy, S.L., (2008). Built environment correlates of walking: a review. *Medicine and Science in Sports and Exercise* 40 (7), S550–S566.
- Sanders, I. (2008). *Complex systems thinking and new urbanism. New urbanism and beyond: designing cities for the future*. Rizzoli, New York, 275-279.
- Sandstrom, U. G. (2002). Green infrastructure planning in urban Sweden. *Planning Practice and Research*, 17(4), 373-385.



- Sarkar, C., Gallacher, J., & Webster, C. (2013). Urban built environment configuration and psychological distress in older men: Results from the Caerphilly study. *BMC public health*, 13(1), 695.
- Savino, S. (2011). A solution to the problem of the generalization of the Italian geographical databases from large to medium scale: approach definition, process design and operators implementation (Doctoral dissertation, Dissertation thesis, Università di Padova).
- Schüssler, N., & Axhausen, K. W. (2009). Processing Raw Data from Global Positioning Systems Without Additional Information. *Transportation Research Record: Journal of the Transportation Research Board*, 2105(4).
- Shaw S-L, Yu H., and Bombom L. S. (2008). A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS* 12: 425–41
- Shoval, N. (2008). Tracking technologies and urban analysis. *Cities* 2008,25, 21-28
- Sundquist, K., Eriksson, U., Mezuk, B., & Ohlsson, H. (2015). Neighborhood walkability, deprivation and incidence of type 2 diabetes: A population-based study on 512,061 Swedish adults. *Health & place*, 31, 24-30.
- Troped, P. J., Wilson, J. S., Matthews, C. E., Cromley, E. K., & Melly, S. J. (2010). The built environment and location-based physical activity. *American journal of preventive medicine*, 38(4), 429-438.
- Turner, A. (2007). From axial to road-centre lines: a new representation for space syntax and a new model of route choice for transport network analysis. *Environment and Planning B: Planning and Design*, 34(3), 539-555.
- Urbonaitė, I. (2013). Rekreacinių funkcijų raidos transformacijos posovietinio miesto viešosiose erdvėse. Vilniaus atvejis [Transformations in spatial expression of urban recreational functions in post-soviet spaces. Vilnius case], *Journal of Architecture and Urbanism* 37(3): 194–209. Vilnius: Technika. ISSN 2029-7955 (SCOPUS, ICONDA, Index Copernicus).
- Valencia City Council (2011), Statistics Summary of the City of Valencia. [http://www.valencia.es/ayuntamiento/webs/estadistica/Recull/Recull2012\\_Ingles.pdf](http://www.valencia.es/ayuntamiento/webs/estadistica/Recull/Recull2012_Ingles.pdf) Accessed 19 February 2015
- Valle D. C. (2013). Sport in the City Research on the relation between sport and urban design. Centro de Estudos de Arquitectura e Urbanismo, Faculdade de Arquitectura da Universidade do Porto.
- Van der Spek, S. C., Van Langelaar, C. M., & Kickert, C. C. (2013). Evidence-based design: satellite positioning studies of city centre user groups. *Proceedings of the ICE-Urban Design and Planning*, 166(4), 206-216.
- Van der Spek, S., McArdle, G., Demšar, U., & McLoone, S. (2014). Classifying pedestrian movement behaviour from GPS trajectories using visualization and clustering. *Annals of GIS*, 20(2), 85-98.
- Van der Spek, S., Van Schaick, J., De Bois, P., & De Haan, R. (2009). Sensing human activity: GPS tracking. *Sensors*, 9(4), 3033-3055.
- Van Dyck, D., Cerin, E., Cardon, G., Deforche, B., Sallis, J. F., Owen, N., & De Bourdeaudhuij, I. (2010). Physical activity as a mediator of the associations between neighborhood walkability and adiposity in Belgian adults. *Health & place*, 16(5), 952-960.
- Vilnius.lt (2015). Vilnius city municipality administration website. Accessed 08 Feb 2015. [www.vilnius.lt](http://www.vilnius.lt)

Vinnitskaya, I. (2012). "What Can Architecture Do for Your Health?". ArchDaily. Accessed 08 Jan 2015. <http://www.archdaily.com/?p=244063>

Vinnitskaya, I. (2013). "Where Does Zoning Fit Into Our Future City Planning?". ArchDaily. Accessed 08 Jan 2015. <http://www.archdaily.com/?p=337042>

Wang Y, et al. (2008). "Will all Americans become overweight or obese? Estimating the progression and cost of the US obesity epidemic". *Obesity*. 16(10): p. 2323–2330.

Wray, S., Hay, J., Walker, H., Staff, R., (2005). Audit of the Towns, Cities and Development Workstream of the England Biodiversity Strategy. English Nature research report number 652. English Nature, Peterborough.

Zhou, J. (2012). Urban Vitality in Dutch and Chinese New Towns: A Comparative Study Between Almere and Tongzhou. TU Delft.

## APPENDIX A. COMPARISON VALUES BETWEEN DIFFERENT VARIATIONS OF POTENTIAL RECREATIONAL USAGE AND THE ACTUAL RECREATIONAL USAGE

Operator					Variables		Vilnius		Valencia		Gothenburg	
							RMSE	R <sup>2</sup>	RMSE	R <sup>2</sup>	RMSE	R <sup>2</sup>
	Greenness						0.5855	0.1046	0.4106	0.0370	0.6065	0.0126
	Residential density						0.4478	0.1511	0.4391	0.0788	0.4101	0.1142
	LUM 500m						0.5664	0.0010	0.3913	0.0335	0.6265	0.0246
	LUM 1000m						0.5833	0.0005	0.4301	0.0125	0.6448	0.0097
	LUM 2500m						0.5675	0.0166	0.4999	0.0002	0.6256	0.0009
	NACH 500m						0.9281	0.0191	0.9644	0.0194	0.9789	0.0287
	NACH 1000m						0.9645	0.0272	0.9581	0.0345	0.9724	0.0532
	NACH 2500m						0.9378	0.0358	0.9168	0.0561	0.9330	0.0680
	NACH 3step						0.8995	0.0184	0.8944	0.0113	0.9229	0.0207
	NACH walk_step						0.9449	0.0355	0.9141	0.0597	0.9391	0.0689
	NACH rec_step						0.9313	0.0377	0.8914	0.0671	0.9268	0.0716
	NACH 2						1.0402	0.0286	1.0587	0.0193	1.0578	0.0339
	NACH 3						1.0040	0.0315	1.0116	0.0251	1.0317	0.0407
	NACH n						0.8163	0.0622	0.8386	0.0723	0.8236	0.0737
Average	Greenness	Res. density	LUM 500m	NACH 500m			0.6505	0.0017	0.56	0.0735	0.7143	0.0021
				NACH 1000m			0.6627	0.0011	0.5573	0.0911	0.7116	0.0002
				NACH 2500m			0.6532	0.0004	0.5431	0.113	0.698	0.0001
				NACH 3step			0.6438	0.0051	0.5398	0.0656	0.6976	0.008
				NACH walk_step			0.6559	0.0005	0.5423	0.1167	0.7003	0
				NACH rec_step			0.6512	0.0003	0.5346	0.1242	0.6961	0.0001
				NACH 2			0.6856	0.0004	0.59	0.0733	0.7394	0.0012
				NACH 3			0.6746	0.0005	0.5748	0.0811	0.7314	0.0009
				NACH n			0.612	0.0003	0.5158	0.1322	0.6597	0.0005
	Greenness	Res. density	LUM 1000m	NACH 500m			0.6635	0.0001	0.5808	0.0542	0.7247	0
				NACH 1000m			0.6762	0	0.578	0.0727	0.7221	0.0015
				NACH 2500m			0.6669	0.0002	0.5637	0.0952	0.7086	0.0036
				NACH 3step			0.6572	0.0016	0.5603	0.0477	0.708	0.0017
				NACH walk_step			0.6697	0.0002	0.563	0.0977	0.7108	0.0034
				NACH rec_step			0.6651	0.0003	0.5554	0.1049	0.7066	0.0038
				NACH 2			0.6996	0.0001	0.6108	0.0545	0.75	0.0002
				NACH 3			0.6886	0.0001	0.5956	0.062	0.742	0.0004
				NACH n			0.6256	0.0031	0.5366	0.1114	0.6701	0.0058
	Greenness	Res. density	LUM 2500m	NACH 500m			0.6627	0.0009	0.608	0.0364	0.7217	0.0024
				NACH 1000m			0.6758	0.002	0.6052	0.0532	0.7191	0.0085
				NACH 2500m			0.6666	0.0039	0.5909	0.0749	0.7056	0.0133
				NACH 3step			0.6569	0	0.5873	0.0296	0.705	0.0001
				NACH walk_step			0.6694	0.0037	0.5903	0.076	0.7079	0.013
				NACH rec_step			0.6648	0.0042	0.5828	0.082	0.7037	0.0139
				NACH 2			0.6995	0.0032	0.6381	0.0371	0.7474	0.0042
				NACH 3			0.6884	0.0033	0.6228	0.0439	0.7392	0.0051
				NACH n			0.6255	0.0116	0.5641	0.0873	0.6671	0.0179
Geometric mean	Greenness	Res. density	LUM 500m	NACH 500m			0.3675	0.0093	0.2174	0.0869	0.4191	0.014
				NACH 1000m			0.3746	0.0084	0.2148	0.0977	0.4141	0.0084
				NACH 2500m			0.3623	0.0065	0.2082	0.1137	0.3964	0.0053
				NACH 3step			0.3526	0.0113	0.2118	0.079	0.3939	0.0187
				NACH walk_step			0.3647	0.0068	0.2074	0.116	0.399	0.0056
				NACH rec_step			0.3597	0.0063	0.2042	0.1225	0.3934	0.0051
				NACH 2			0.396	0.0075	0.2276	0.0883	0.4478	0.0127
				NACH 3			0.384	0.0076	0.2215	0.0922	0.4382	0.0119
				NACH n			0.3173	0.0025	0.1975	0.1362	0.3449	0.0029
	Greenness	Res. density	LUM 1000m	NACH 500m			0.373	0.0054	0.2281	0.0656	0.4261	0.0026
				NACH 1000m			0.3806	0.0045	0.2248	0.0799	0.4214	0.0005
				NACH 2500m			0.3691	0.003	0.2171	0.0988	0.4042	0
				NACH 3step			0.3581	0.007	0.2206	0.0593	0.4004	0.0053
				NACH walk_step			0.3722	0.0032	0.2169	0.0993	0.407	0
				NACH rec_step			0.3671	0.0028	0.2135	0.1057	0.4014	0
				NACH 2			0.4051	0.0038	0.2398	0.0679	0.4568	0.0021
				NACH 3			0.3924	0.0038	0.2328	0.0726	0.4466	0.0017

	Greenness	Res. density	LUM 2500m	NACH n	0.3222	0.0003	0.2062	0.1175	0.3509	0.0003
				NACH 500m	0.3597	0.0013	0.2516	0.0407	0.4081	0.0001
				NACH 1000m	0.3667	0.0007	0.2479	0.0547	0.4035	0.002
				NACH 2500m	0.356	0.0002	0.2383	0.0741	0.3864	0.0044
				NACH 3step	0.3459	0.002	0.2411	0.0343	0.3837	0.0002
				NACH walk_step	0.3591	0.0003	0.2384	0.0736	0.3895	0.004
				NACH rec_step	0.3545	0.0002	0.2343	0.0793	0.3841	0.0046
				NACH 2	0.3917	0.0004	0.2669	0.0429	0.4391	0.0003
				NACH 3	0.3792	0.0004	0.2579	0.0476	0.4285	0.0006
				NACH n	0.3129	0.0005	0.2252	0.0888	0.3356	0.0074
OR	Greenness	Res. density	LUM 500m	NACH 500m	0.6735	0.0455	0.4663	0.0373	0.6958	0.0204
				NACH 1000m	0.6735	0.0455	0.4663	0.0373	0.6958	0.0204
				NACH 2500m	0.6735	0.0455	0.4663	0.0373	0.6958	0.0204
				NACH 3step	0.6735	0.0455	0.4663	0.0373	0.6958	0.0204
				NACH walk_step	0.6735	0.0455	0.4663	0.0373	0.6958	0.0204
				NACH rec_step	0.6735	0.0455	0.4663	0.0373	0.6958	0.0204
				NACH 2	0.6735	0.0455	0.4663	0.0373	0.6958	0.0204
				NACH 3	0.6735	0.0455	0.4663	0.0373	0.6958	0.0204
				NACH n	0.6735	0.0455	0.4663	0.0373	0.6958	0.0204
	Greenness	Res. density	LUM 1000m	NACH 500m	0.6763	0.0273	0.4798	0.0167	0.7028	0.0156
				NACH 1000m	0.6763	0.0273	0.4798	0.0167	0.7028	0.0156
				NACH 2500m	0.6763	0.0273	0.4798	0.0167	0.7028	0.0156
				NACH 3step	0.6763	0.0273	0.4798	0.0167	0.7028	0.0156
				NACH walk_step	0.6763	0.0273	0.4798	0.0167	0.7028	0.0156
				NACH rec_step	0.6763	0.0273	0.4798	0.0167	0.7028	0.0156
				NACH 2	0.6763	0.0273	0.4798	0.0167	0.7028	0.0156
				NACH 3	0.6763	0.0273	0.4798	0.0167	0.7028	0.0156
				NACH n	0.6763	0.0273	0.4798	0.0167	0.7028	0.0156
	Greenness	Res. density	LUM 2500m	NACH 500m	0.6625	0.0109	0.5257	0.0009	0.6877	0.0096
				NACH 1000m	0.6625	0.0109	0.5257	0.0009	0.6877	0.0096
				NACH 2500m	0.6625	0.0109	0.5257	0.0009	0.6877	0.0096
				NACH 3step	0.6625	0.0109	0.5257	0.0009	0.6877	0.0096
				NACH walk_step	0.6625	0.0109	0.5257	0.0009	0.6877	0.0096
				NACH rec_step	0.6625	0.0109	0.5257	0.0009	0.6877	0.0096
				NACH 2	0.6625	0.0109	0.5257	0.0009	0.6877	0.0096
				NACH 3	0.6625	0.0109	0.5257	0.0009	0.6877	0.0096
				NACH n	0.6625	0.0109	0.5257	0.0009	0.6877	0.0096
	Greenness	Res. density	LUM 500m	NACH 500m	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH 1000m	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH 2500m	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH 3step	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH walk_step	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH rec_step	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH 2	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH 3	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH n	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
	Greenness	Res. density	LUM 1000m	NACH 500m	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH 1000m	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH 2500m	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH 3step	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH walk_step	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH rec_step	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH 2	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH 3	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH n	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
	Greenness	Res. density	LUM 2500m	NACH 500m	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH 1000m	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH 2500m	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH 3step	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH walk_step	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH rec_step	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH 2	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH 3	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH n	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
AND	Greenness	Res. density	LUM 500m	NACH 500m	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH 1000m	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH 2500m	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH 3step	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH walk_step	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH rec_step	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH 2	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH 3	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
				NACH n	0.4584	0.021	0.3227	0.0464	0.5256	0.0281
	Greenness	Res. density	LUM 1000m	NACH 500m	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH 1000m	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH 2500m	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH 3step	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH walk_step	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH rec_step	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH 2	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH 3	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
				NACH n	0.4751	0.0229	0.3509	0.0395	0.5383	0.0136
	Greenness	Res. density	LUM 2500m	NACH 500m	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH 1000m	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH 2500m	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH 3step	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH walk_step	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH rec_step	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH 2	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH 3	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056
				NACH n	0.4754	0.0179	0.3767	0.0353	0.5351	0.0056

**Table 8. Comparison values of the correlation between different variants of the data model for the potential recreational usage and the actual recreational usage**

## APPENDIX B. SCATTER PLOTS



Figure 68. Scatter plots of actual recreational usage values with corresponding greenness, LUM 500m and NACH n radius in four different combination methods for all case study cities.





**Figure 69. Scatter plots of actual recreational usage values with corresponding greenness, LUM 500m and NACH n radius in four different combination methods for all case study cities without taking into account residential density values.**

## REFLECTION

---

Studies of the built environment's impact on citizens' physical activity have become an aspiring topic in the recent years. This commotion is created due to increased awareness that encouraging physical activity and creating qualitative recreational spaces within cities has positive impact on such various factors as value of real estate, sustainability, urban sprawl, health care expenses, productivity and workforce availability. It has been learned to think of space not as the background to human activity, but as an intrinsic aspect of everything, human beings do. The importance of scientific knowledge in these matters is that when trying to improve a situation in urban planning, there should be evident knowledge of both present situation and consequences of any action.

Although there is a consensus on the benefits of such shifts, there is little scientific knowledge on how cities are being used for recreational purposes and even harder it is to monitor any changes in people behaviour after changes are made. Yet the data suitable for this kind of research is voluntary produced by people using sports tracking applications.

Thus the aim of Master Thesis research was to develop a method to acquire, manage and process the data from sports tracking applications in such a way that it would serve as a ground truth not only for examining urban recreational travel patterns but also for modelling the phenomena. In other words, the goal was being able to define where recreational activities happen, where they do not and finally, use this knowledge to give an indication to every space of how likely it is that the space *is* or *will be* used for recreation. The Runability Index, introduced by this research, intends to be used for creating knowledge-based models of cities recreational systems, developing existing and creating new strategies for urban recreation systems and their regeneration and assessment of structural elements, investigating the relationship between various characteristics of the built environment and the recreational physical activity.

Generally, the Master Thesis research is a fusion between the fields of Urbanism and Geomatics for the Built Environment. Urban studies are mostly empirical and try to explain "what will probably be the case", however the practical approach of Geomatics aims to explain, "what is or can be the case" based on the collected evidence, which in this case was data taken from a mobile sports tracking application. Broadly speaking, the application of the research is a societal one, but the result is a technically developed scientific instrument, which can eventually contribute to the advance of empirical scientific knowledge.

The research, as well as the science of Geomatics is concerned with the acquisition, analysis, management and visualisation of geographic data with the aim of gaining knowledge and a better understanding of the built environments. The GPS tracking techniques have been used in order to collect appropriate data, which was later processed and analysed using GIS tools for big data management, storage, manipulation and visualisation. The collected data has been used in tandem with OpenStreetMap and Eurostat Urban Atlas datasets in order to calibrate and validate the introduced urban space index, based on the well-known measure of walkability, used in urban sciences.

The Master Thesis research makes use of most of the techniques taught by the Master Geomatics, including data collection and analysis, spatial information modelling and the visualisation of data for solving real-world problems in an innovative way.

*Rusnė Šilerytė  
MSc Geomatics*

*r.sileryte@student.tudelft.nl  
Delft, 2015*