University of Thessaly
School of Engineering
Department of Planning and Regional Development

The use of crowd-sourced geographic data in spatial planning

by

Spyridon Spyratos

A dissertation submitted

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Thesis consulting committee:

Associate professor Dimitrios Stathakis (Supervisor)

Professor Dimitrios Economou

Professor Vassilios Pappas

Volos Greece, February 2017

Certified by the members of the Dissertation Committee:

1st member (Supervisor)	Associate Professor Dimitrios Stathakis (Supervisor) Associate Professor in the Department of Planning and Regional Development, University of Thessaly
2nd member	Professor Dimitrios Economou Professor in the Department of Planning and Regional Development University of Thessaly
3rd member	Professor Vassilios Pappas Professor in the Department of Architecture University of Patras
4th member	Professor Marie Noelle Duquenne Professor in the Department of Planning and Regional Development University of Thessaly
5th member	Professor Dimitrios Kotzinos Professor in the Department of Computer Science University of Cergy Pontoise
6th member	Assistant Professor Vassilios Tselios Assistant Professor in the Department of Planning and Regional Development, University of Thessaly
7th member	Assistant Professor Stamatis Kalogirou Assistant Professor in the Department of Geography Harokopio University of Athens

ABSTRACT

Spatial and urban planning is a technical and political process which requires various types of data. For example, data about existing land use, environmental pollution data and data about citizens' satisfaction with regard to urban facilities and services. However, for many cities, the required data is not available or is not up-to-date. Its collection using traditional methods, such as in-situ surveys, requires a significant volume of human and financial resources. In the present PhD research, we examined the use of crowd-sourced data for enriching or replacing data collected using traditional methods. To this end, two case studies have been performed. The first, is about the use of crowd-sourced data for land/building use mapping. In this case study, we developed a method for estimating building block use using place data from Foursquare social media application. This method can be trusted for the estimation of the land use categories "*Hotels, restaurants & cafes*" and "*Retail*". The second case study, is about the use of crowd-sourced data for estimating citizens' satisfaction with regard to urban facilities and services. To this end, we proposed two indicators which are estimated using Foursquare place data. The proposed indicators can provide robust estimates about citizens' satisfaction with regard to "*Sport facilities*", "*Cultural facilities*" and "*Streets & buildings*". Both case studies proved that by using appropriate methods such as those developed in this PhD research, crowd-sourced data can be used for complementing official statistics collected through traditional methods. The added value of its use, is that it can provide low cost, up-to-date and globally harmonized estimates of land use patterns and of citizen's satisfaction levels for specific types of services and facilities. These estimates can be used for comparing cities, for identifying changes in land use patterns and in citizens' satisfaction levels, and lastly for designing effectively in-situ surveys.

Keywords: Crowdsourcing, Volunteered Geographical Information, Land use, Urban Planning, Spatial Planning, Social media, Foursquare.

ΠΕΡΙΛΗΨΗ

Ο χωροταξικός και πολεοδομικός σχεδιασμός είναι μια τεχνική και πολιτική διαδικασία που απαιτεί την ύπαρξη δεδομένων διαφορών τύπων, όπως για παράδειγμα, δεδομένα για τις υπάρχουσες χρήσεις γης, δεδομένα για την μόλυνση του περιβάλλοντος και δεδομένα για την ικανοποίηση των πολιτών σχετικά με τις αστικές υποδομές και υπηρεσίες. Η συλλογή των δεδομένων αυτών μέσω των παραδοσιακών μεθόδων, όπως οι επιτόπιες έρευνες, απαιτούν ένα σημαντικό αριθμό ανθρώπινων και οικονομικών πόρων. Στην παρούσα διδακτορική έρευνα εξετάζουμε τη χρήση των συλλογικών (crowdsourced) δεδομένων για τον εμπλουτισμό ή την αντικατάσταση των δεδομένων που συλλέγονται με τις παραδοσιακές μεθόδους. Για το σκοπό αυτό, διεξήχθησαν δύο μελέτες περιπτώσεων (case studies). Η πρώτη, αφορά την χρήση συλλογικών δεδομένων για τη χαρτογράφηση των χρήσεων γης/κτηρίων. Σε αυτή τη μελέτη περίπτωσης αναπτύξαμε μια μέθοδο για την εκτίμηση της χρήσης κτιριακών συγκροτημάτων χρησιμοποιώντας δεδομένα τόπων από το μέσο κοινωνικής δικτύωσης Foursquare. Η μέθοδος αυτή αποδείχθηκε αξιόπιστη για την εκτίμηση των κατηγοριών χρήσης γης "*Ξενοδοχεία, εστιατόρια και καφέ*" και "*Λιανικό εμπόριο*". Η δεύτερη μελέτη περίπτωσης, αφορά την χρήση συλλογικών δεδομένων για την εκτίμηση της ικανοποίησης των πολιτών σε σχέση με τις αστικές υποδομές και υπηρεσίες. Για το σκοπό αυτό, προτάθηκαν δύο δείκτες οι οποίοι υπολογίζονται βάση δεδομένων τόπων από το Foursquare. Οι προτεινόμενοι δείκτες μπορούν να παρέχουν αξιόπιστες εκτιμήσεις σχετικά με την ικανοποίηση των πολιτών όσον αφορά τις "*Αθλητικές εγκαταστάσεις*", τις "*Πολιτιστικές εγκαταστάσεις*" και τους "*Δρόμους & κτίρια*". Και οι δύο μελέτες περιπτώσεων απέδειξαν ότι χρησιμοποιώντας τις κατάλληλες μεθόδους, όπως αυτές που αναπτύχθηκαν στην παρούσα διδακτορική έρευνα, τα συλλογικά δεδομένα μπορούν να χρησιμοποιηθούν συμπληρωματικά ως προς τα επίσημα στατιστικά δεδομένα που συλλέγονται μέσω παραδοσιακών μεθόδων. Η προστιθέμενη αξία της χρήσης των δεδομένων αυτών, είναι ότι παρέχουν αξιόπιστες, χαμηλού κόστους, ενήμερες, και εναρμονισμένες σε υπερεθνικό επίπεδο εκτιμήσεις χρήσεων γης και εκτιμήσεις επιπέδων ικανοποίησης των πολιτών για συγκεκριμένους τύπους υπηρεσιών και υποδομών. Οι εκτιμήσεις αυτές δύναται να χρησιμοποιηθούν για τη σύγκριση πόλεων, τον εντοπισμό αλλαγών στις χρήσης γης και στα επίπεδα ικανοποίησης των πολιτών, και τέλος, για τον καλύτερο σχεδιασμό επιτόπιων ερευνών.

Λέξεις κλειδιά: Συλλογικά Δεδομένα, Πληθοπορισμός, Εθελοντική Γεωγραφική Πληροφορία, Χρήσεις γης, Πολεοδομία, Χωροταξία, Μέσα κοινωνικής δικτύωσης, Foursquare.

TABLE OF CONTENTS

1. INTRODUCTION....1
1.1. RESEARCH AIM....1
1.2. BACKGROUND....2
1.2.1. THE RISE OF THE UGGC AND OF THE VGI....4
1.2.2. CROWD-SOURCING VS. COMMONS-BASED PEER PRODUCTION5
1.3. TYPOLOGY OF CITIZEN-CONTRIBUTED GEOGRAPHIC DATA....7
1.4. CHARACTERISTICS OF CCGD....10
1.4.2. QUALITY OF INITIAL DATA SUBMISSIONS....11
1.4.3. QUALITY OF DATASETS....13
1.4.4. COST OF DATA COLLECTION....15
1.5. OUTLINE OF THE THESIS....16
2. CASE STUDIES....17
2.1. CASE STUDY A –ESTIMATING BUILDING BLOCK USE....17
2.1.1. RELATED WORK....19
2.2. CASE STUDY B - EVALUATING SERVICES AND FACILITIES....20
3. DATA....22
3.1. FOURSQUARE PLACE DATA....22
3.2. CASE STUDY A - ESTIMATING BUILDING BLOCK USE....24
3.2.1. FOURSQUARE DATA....24
3.2.2. AMSTERDAM STUDY AREA....26
3.2.3. VARESE STUDY AREA....30
3.3. CASE STUDY B - EVALUATING SERVICES AND FACILITIES....31
3.3.1. FOURSQUARE DATA....32
3.3.2. EUROBAROMETER DATA....34

4. METHODOLOGY 35
4.1. CASE STUDY A - ESTIMATING BUILDING BLOCK USE 35
4.1.1. FOURSQUARE PLACE DATA PREPARATION 36
4.1.2. BUILDING BLOCK USE (BBU) ESTIMATION 39
4.1.3. BBU ACCURACY ASSESSMENT 45
4.1.4. REPRODUCIBILITY OF THE PROPOSED METHODOLOGY 48
4.2. CASE STUDY B- EVALUATING SERVICES AND FACILITIES 49
4.2.1. FOURQUARE PLACE DATA PREPARATION 49
4.2.2. ESTIMATION OF THE PROPOSED INDICATORS 52
4.2.3. ACCURACY ASSESSMENT 55
5. RESULTS 56
5.1. CASE STUDY A - ESTIMATING BUILDING BLOCK USE 56
5.1.1. BEST ACCURACY RESULTS 56
5.1.2. IMPACT OF d AND c PARAMETER VALUES 58
5.1.3. DENSITY OF "RETAIL" AND "HOTELS, RESTAURANTS AND CAFES" LU 62
5.1.4. SPATIAL VARIATION OF THE ACCURACY RESULTS 63
5.1.5. ASSESSMENT OF THE METHODOLOGY IN THE CITY OF VARESE, ITALY 65
5.1.6. SELECTION OF APPROPRIATE PARAMETER VALUES 66
5.2. CASE STUDY B - EVALUATING SERVICES AND FACILITIES 68
5.2.1. ACCURACY OF INDICATORS 68
5.2.2. IMPACT OF THE WEIGHT VALUES ON THE OF THE ACCURACY 72
6. DISCUSSION & COCLUSIONS 74
6.1. CASE STUDY A - ESTIMATING BUILDING BLOCK USE 74
6.1.1. DISCUSSION 74

6.1.2. CONCLUSIONS....76
6.2. CASE STUDY B - EVALUATING SERVICES AND FACILITIES....78
6.2.1. DISCUSSION....78
6.2.2. CONCLUSIONS....79
6.3. GENERIC FRAMEWORK FOR USING CROWD-SOURCED DATA....81
6.4. OVERALL DISCUSSION & CONCLUSIONS....82
REFERENCES....84

TABLE OF TABLES

Table 1: The number of Foursquare places in cities under investigation, and the population of age 15 years old and over. 31

Table 2: Number and percentage of Amsterdam LU subcategories for each of the 8 main Amsterdam LU categories, and number and percentage of Foursquare place categories that have been aligned to them. 37

Table 3**:** Confusion matrix for the "*Retail*" LU category. TP represents the number of building blocks that have "*Retail*" LU in both the estimated and the reference BBU datasets. TN represents the number of building blocks that do not have "*Retail*" LU in both the estimated and the reference BBU datasets. 46

Table 4: Alignment between the Eurobarometer categories and the Foursquare place categories 50

Table 5**:** Best accuracy results for each LU category of the Amsterdam study area. In the last two columns we present the d and c parameter values for which the highest Cohen's Kappa coefficient values were achieved. 57

Table 6: Best accuracy results for the two subareas of the Amsterdam study area. 64

Table 7: Accuracy assessment results of the Varese study area. 65

Table 8: Linear regression analysis results between the percentages of "Very satisfied" and "Totally satisfied" citizens as recorded in the Eurobarometer survey and the proposed indicators, where s and b are the proposed indicators, and w is the weight used for the estimation of the indicators. Significance levels are noted as follows: $*** p \leq 0.001$, $** p \leq 0.01$, $* p \leq 0.05, ns\ p > 0.05$ 69

TABLE OF FIGURES

Figure 1: Typology of Geographic Data 9

Figure 2: Many Facebook and Foursquare places are erroneously located in Amstel canal, Amsterdam, The Netherlands 12

Figure 3: Few Facebook and Foursquare places are located in the industrial area, at the right side of the figure, while many places that represent offices and retail facilities are located at the right side of the figure. 15

Figure 4: Density maps based: (i) on the total number of places that are located in each pixel; and (ii) on the total number of Foursquare users that have checked in places that are located in each pixel. The study area is divided in pixels of 300m by 300m size. ...25

Figure 5: Main Study area: The area bounded by the A10 motorway. 26

Figure 6: Building blocks with "*Retail*" LU (i) and "*Offices*" LU (ii). A building block can have multiple LU categories as long as it contains at least one surface of these LU categories. 29

Figure 7: Varese study area. The buildings in the Municipality of Varese are highlighted in light grey and the boundary of Varese Municipality in black. In dark grey are the 150 randomly selected building blocks for which a LU ground survey was performed. 30

Figure 8: Density maps based on the total number of places that are located in each pixel. The cities of Brussels and Athens are divided into pixels of 300m by 300m size.....33

Figure 9: Linear correlation between the Foursquare places and the population aged 15 years and older. 33

Figure 10: Methodology followed for the preparation of Foursquare place data, for the estimation of BBU, for assessing the accuracy of these estimations and for testing the reproducibility of the proposed methodology. 35

Figure 11: Number of Foursquare places and number of surfaces of the non-residential use dataset for each Amsterdam LU category. 38

Figure 12: Foursquare places pj and surfaces of the non-residential use dataset si located at the intersection of Rozengracht and Eerste Bloemdwarsstraat streets in Amsterdam. Both datasets are classified according to the Amsterdam LU categories. 39

Figure 13: Method used for assessing the accuracy of the geometric allocation of Foursquare places to their closest building block based on the application of different d parameter values.....40

Figure 14: Percentage of Foursquare places for which both their address and their geographic location refer to the same building block (positional accuracy) vs number of Foursquare places, using the d parameter values in the range [0m … 50m]. For example, for the parameter value d =6m there are 21,000 places from which, based on the 9,845 place sample, 70% is positionally accurate. 41

Figure 15: Percentage of Foursquare places for which both their address and their geographic location refer to the same building block (positional accuracy) vs number of places, using c parameter values in the range [0…50].....43

Figure 16: BBU estimation using Foursquare place data by taking into consideration the parameters c and d. In the building block $b2$, using parameter values $cp \geq c = 0$ and $dp \leq d = 2m$, the $a5$ and $a6$ LU categories are assigned. 44

Figure 17: Building blocks estimated to have retail use (i) versus building blocks that have retail use in the reference dataset (ii) in the "*Bos en Lommer*" neighbourhood, Amsterdam.....47

Figure 18: Methodology followed for the preparation of Foursquare place data, for the estimation of the proposed indicators and for assessing their accuracy. 49

Figure 19: Number of Foursquare places of the 17 cities, for each Eurobarometer category. 51

Figure 20: Cohen's Kappa coefficient of the estimated BBU dataset calculated using the parameter values $d = [0m \dots 50m]$ and $c = 0$.....58

Figure 21: Cohen's Kappa coefficient of the estimated BBU dataset calculated using the parameter values $c = [0 \dots 50]$ and $d = 20m$. 59

Figure 22: Precision per LU category of the estimated BBU dataset calculated using the parameter values $d = [0m \dots 50m]$ and $c = 0$.....60

Figure 23: Sensitivity per LU category of the estimated BBU dataset calculated using the parameter values $d = [0m \dots 50m]$ and $c = 0$.....60

Figure 24: Precision per LU category of the estimated BBU dataset calculated using the parameter values $c = [0 \dots 50]$ and $d = 20m$. 61

Figure 25: Sensitivity per LU category of the estimated BBU dataset calculated using the parameter values $c = [0 \dots 50]$ and $d = 20m$. 61

Figure 26: Linear regression plots between the number of places of the best estimated BBU dataset and the number of surfaces in the reference BBU dataset of each building block for the LU categories "*Hotels, restaurants & cafes*" and "*Retail*". The number of building blocks for each combination of estimated places and reference surfaces is presented in logarithmic scale. 62

Figure 27: The two subareas of the Amsterdam study area: the Amsterdam-Centrum and the A10-periphery, which is the area that remains if we exclude the Amsterdam-Centrum from the area that is enclosed by the A10 motorway. 63

Figure 28: The Kappa coefficients of the BBU datasets that were estimated using parameter values d =25 and c as defined by the calibrations on 20 randomly selected samples of the Amsterdam study area. 67

Figure 29: Correlation between the proposed indicators of the "*public spaces*" category and the percentage of "Very satisfied" and "Totally satisfied" citizens with regard to "*public spaces*" as recorded in the Eurobarometer survey. 70

Figure 30: Correlation between the proposed indicators of the "*Sports facilities*" category and the percentage of "Very satisfied" and "Totally satisfied" citizens with regard to "*Sports facilities*" as recorded in the Eurobarometer survey. 71

Figure 31: The impact of the weight value (w) on the coefficient of determination (R^2) between the s indicator and the percentage of "Very satisfied" citizens 73

Figure 32: The impact of the weight value (w) on the coefficient of determination (R^2) between the b indicator and the percentage of "Totally satisfied" citizens.....73

Figure 33: A generic framework for using crowd-sourced data for multi-thematic applications 81

LIST OF ABBREVIATIONS

Abbreviation	Meaning
API	Application Programming Interface
BBU	Building Block Use
CCGD	Citizen-Contributed Geographic Data
CORINE	CoORdination of INformation on the Environment
EU	European Union
FN	False Negatives
FP	False Positives
GD	Geographic Data
GNSS	Global Navigation Satellite System
LC	Land Cover
LU	Land Use
LUCAS	Land Use/Cover Area frame Survey
PGI	Professional Geographic Information
PhD	Doctor of Philosophy
Private GD	Private Geographic Data
QA	Quality Assurance
QC	Quality Control
SGD	Social Geographic Data
TN	True Negatives
TP	True Positives
UGC	User-Generated Content
UGGC	User Generated Geographic Content
UK	United Kingdom
VGI	Volunteered Geographic Information

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Professor Demetris Stathakis for the continuous support during the last four years. His guidance helped me to set targets, to overcome research limitations and importantly to remain concentrated and focused on my research.

Besides my main advisor, I would like to thank my advisors at the Joint Research Center Michael Lutz, Chrisa Tsinaraki, and Massimo Craglia.

I would like to thank Jacopo Grazzini, Elena Roglia, Francesco Pantisano, Sven Schade, Chris Jacobs-Crisioni and several anonymous reviewers for their valuable comments and suggestions during the course of my PhD research.

My sincere thanks also goes to Alessandro Annoni and Ioannis Kanellopoulos who provided me the opportunity to join the Digital Earth and Reference Data Unit as a PhD researcher.

This PhD research has been supported through a research grant (No IES-2011-201108) of the Joint Research Centre of the European Commission.

1. INTRODUCTION

In the Introduction Section the research aim and the characteristics of the crowd-sourced/ citizen-contributed geographic data are presented. In Section 1.1., we present the research aim and we highlight the two research questions. In Section 1.2., we provide the necessary background and we present the terms that are used to describe citizen-contributed data. In Section 1.3., we propose a typology of citizen-contributed geographic data so as to decrease the terminological ambiguity. In Section 1.4., we present the characteristics of citizen-contributed data and finally, in Section 1.5., we outline the structure of this thesis.

1.1. RESEARCH AIM

Spatial planning is a technical and political process which requires various types of up-to–date data. For example, such data includes Land Use (LU) data, environmental pollution data, traffic data and data about citizens' satisfaction with regard to urban facilities and services. The analyses of such data can highlight major environmental, social and economic issues in the area under investigation, can contribute towards a more holistic understanding of the environment, and finally, can facilitate evidence-based decision making. As a result, the availability of such data is essential to urban planners and managers.

However, for many cities, the required data is not available or is not up-to-date. The on-demand collection of this data using in-situ surveys requires a significant volume of human and financial resources that increase the cost of spatial and urban planning studies. In addition, in many cases, the required data, such as the land use data, is not collected at frequent time intervals. As a result, urban planners cannot easily monitor changes in the urban environment.

Nowadays, various internet-based applications request their users to collect geographical, environmental or geo-referenced data (Spyratos et al., 2014). In most cases, these applications offer free or low cost access to the data that has been contributed by theirs users. In the literature, various terms are often being used interchangeably to describe various types of geo-referenced data contributed intentionally by citizens to Internet applications. The most widely used terms are the Volunteered Geographic Information (VGI) (Goodchild, 2007) or crowd sourced geographic information. Examples of well-

known Internet applications that collect citizen-contributed geographic data are the OpenStreetMap (Haklay and Weber, 2008) and the Foursquare social recommendation application (Foursquare, 2016a). The users of these applications are contributing geographical data about roads, buildings and places.

The aim of this PhD research is to assess whether data contributed by citizens to Internet-based applications can be used in the context of spatial and urban planning studies, as either a low cost alternative or an auxiliary input to data collected through traditional in-situ surveys.

The feasibility of this proposal can be tested by answering the following two research questions:

1. Can we use crowd-sourced data for estimating land use?
2. Can we use crowd-sourced data for estimating citizens' satisfaction with regard to the urban facilities and services?

These two research questions are answered by the two case studies which are presented in Section 2.

1.2. BACKGROUND

Social and technological developments, such as the high number of highly educated people and the widespread use of smartphone devices, facilitate the ability of citizens to collect and publish data on the Internet. Citizens are contributing data to Internet-based applications for various purposes, such as socially-oriented and scientific. The collected data covers various environmental domains, such as place data (Foursquare, 2016a), air quality (ISPEX, 2016) and land cover observations (Geo-Wiki, 2016). Citizen-Contributed Geographic Data (CCGD) that are collected through these platforms differs from geographic data collected by professionals such as urban planners during land use surveys or structured interviews for the following reasons (Spyratos et al., 2014):

- First, the CCGD data collectors might have significantly diverse level of scientific and technical background (Budhathoki et al., 2010). For example, while the licenced land surveyors have formal qualifications that prove their capacity on surveying, the citizens that contribute data to Internet platforms, such as to the OpenStreetMap, might not have any qualifications.

- Second, the equipment and the methods used for the collection of CCGD are often unknown or of low quality. For example, citizens are using the Global Navigation Satellite System (GNSS)-receivers of their mobile device for mapping the location of bus stops. These GNSS receivers have lower accuracy compared to professional GNSS-receivers (Zandbergen and Barbeau, 2011).
- Third, in most of the cases, the quality of CCGD is not controlled by established quality assurance procedures (Haklay et al., 2010). The resolution of errors on the collected data rely on the willingness of other users.
- CCGD is collected by users who, in most cases, decide independently what types of data they will collect, at what time, and from which area. In contrast to data collected in the context of professional routines and practices, the CCGD data collection is not coordinated or designed a priori by an organization.

CCGD is gratuitously contributed by the citizens to Internet-based applications, and as result these applications offer timely data at very limited cost (Goodchild and Li, 2012). Due to its availability, CCGD has been used as auxiliary input in environmental monitoring (NoiseWatch, 2013; USGS, 2013) and research studies (Fritz et al., 2013).

It is still unclear whether, how and what types of CCGD can contribute towards a better and more holistic understanding of the environment (Spyratos et al., 2014). Goodchild and Li (2012) support that VGI is often not reliable data source for scientific research, since "*its quality is highly variable and undocumented, it fails to follow scientific principles of sampling design, and its coverage is incomplete*". Lee (1994) mentions that using volunteers to monitor the environment is not a new idea, and that for more than 100 years, the National Weather Service of United States of America has trained volunteers to report daily rainfall and air temperature measurements. The statement of Goodchild and Li (2012) and the statement of Lee (1994) are both valid, since they refer to different types of CCGD. CCGD is not a homogenous category of data, and it includes data that differs in terms of the purpose of data collection, the data quality and the characteristics of contributors (Spyratos et al., 2014).

In the next sections we describe some of the most widely used terms for describing CCGD such as the VGI (Goodchild, 2007), User Generated Geographic Content (UGGC), and crowd sourced geographic information. Since these terms are often being used

interchangeably to describe diverse types of geo-referenced data, in the Section 1.3 we present a typology of CCGD data.

1.2.1. THE RISE OF THE UGGC AND OF THE VGI

Technological developments such as the asynchronous JavaScript and XML group of technologies (Cormode and Krishnamurthy, 2008), facilitate the Internet users' ability to distribute their content and to some extent to edit, rate or comment the content produced by others. The content that is produced and shared by the users of Internet platforms, is often termed as "User-Generated Content" (UGC) or "User-Contributed Content". Vickery and Wunsch-Vincent (2007: 9) defined UGC as a *"(a) content made publicly available over the Internet; (b) which reflects a certain amount of creative effort; and (c) which is created outside of professional routines and practices.*"

UGC Internet-platforms act as mediators between the users and the producers of digital content. They allow users to participate in the content production, but this participation does not necessarily result in any ownership rights over the UGC data. The services provided by UGC platforms are mainly based on the content contributed by their users. Consequently, the users of these platforms represent a digital workforce that creates added value for a community, for individual citizens and/or for the owner of the internet platform.

UGC is often associated with indirect geographic references, e.g., through place names, or with explicit geographic references, such as geographic coordinates. An important technological factor that facilitates the collection of explicit geographic UGC on the web is the wide availability of GNSS-enabled mobile devices. UGC that has an implicit or an explicit geographic reference is termed user-generated geographic content (UGGC) (Goodchild, 2008). UGGC is an emerging web-based phenomenon, whose diffusion has been mostly driven by technology.

Another term used to describe CCGD is the VGI. VGI was originally described by Goodchild as web phenomenon of *"…the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information, a function that for centuries has been reserved to official agencies. They are largely untrained and their actions are almost always voluntary, and the results may or may not be accurate"* (Goodchild, 2007: 212).

The term VGI is widely accepted among researchers in the academic field of geographical information science. However, one should note that according to its Goodchild's definition, this term does not necessarily describe geographical data that was intentional collected by citizens during a voluntary activity. The term VGI has been criticized by several researchers because the characterization of this information as volunteered implies an intentionality or altruism on behalf of data collectors that may not be apparent in the so called VGI projects (Elwood, 2008). Due to this fact, in a recent study Elwood, Goodchild and Sui (2012: 575) specified the original definition by suggesting a more strict one as follows :*"VGI as geographic information acquired and made available to others through the voluntary activity of individuals or groups, with the intent of providing information about the geographic world".*

In principle, VGI, is not a recent phenomenon. There is a long volunteer tradition in the collection of environmental and geographic observations. For instance, since 1890, the United States National Weather Service has been engaging citizens in meteorological monitoring activities in the context of the Cooperative Weather Observer Program (NOAA, 2014). VGI describes a subset of UGGC. The term VGI embodies the notion of volunteering for data collection. It describes a science-oriented phenomenon that is supported by technology. Consequently, we suggest that the term VGI should not be used interchangeably with the term UGGC. Taken into account all these critics, in Section 1.3, we propose a typology of CCGD where we re-define the VGI term.

1.2.2. CROWD-SOURCING VS. COMMONS-BASED PEER PRODUCTION

The term Crowd-sourcing was coined by Jeff Howe and Mark Robinson in 2006 (Howe, 2006b). This term describes a web-based model of production that follows the paradigm of outsourcing. It is a practice that some enterprises or organizations use for assigning tasks to willing individuals or groups of individuals, which may or may not be compensated financially. Various and varying definitions of crowdsourcing exist in the literature (for a detailed presentations see Estellés-arolas & González-ladrón-de-guevara 2012). Howe gives the following "original" definition of crowdsourcing:

> *"Crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often*

> *undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers.* (Howe, 2006a)

Crowd-sourcing has raised ethical and legal concerns because of the way that is being implemented by some profit-oriented companies (Felstiner, 2011; Fort et al., 2011). Brabham, focusing on the positive aspects of crowdsourcing, describes it *"as a model capable of aggregating talent, leveraging ingenuity while reducing the costs and time formerly needed to solve problems"* (Brabham, 2008: 87). Haklay, takes a critical approach towards crowd-sourcing and highlights its negative potentials, as crowd-sourcing can be in specific situations *"a highly exploitative activity, in which participants are encouraged to contribute to an alleged greater good when, in reality, the whole activity is contributing to an exclusive enterprise that profits from it*" (Haklay, 2010: 683).

As Howe (2006a) mentioned crowdsourcing is used interchangeably with the commons-based peer production concept even if they are different. The former follows the profit-oriented outsourcing mode of production while the latter the non-profit community based mode of production. The term "commons-based peer production", was firstly coined by Benkler (2002). Benkler, in a later work of his, described commons-based peer production as (Benkler, 2006: 60):

> *"a new modality of organizing production: radically decentralized, collaborative, and nonproprietary; based on sharing resources and outputs among widely distributed, loosely connected individuals who cooperate with each other without relying on either market signals or managerial commands".*

Crowd-sourcing and commons-base peer production differ on two aspects. First on the entity that controls the production process, and second on the entity that controls the product. Applied to geographical information field, crowd-sourced mapping activities are activities centrally coordinated and initiated by a body. The data that is collected by the contributors is usually controlled and owned by the coordinator of the activity, who may or not may compensate financially the contributors. An example of a crowd-sourced map product is Google Map Maker (Google, 2016). On the contrary, mapping activities that follow the paradigm of commons-based peer production as is the OpenStreetMap project (OpenStreetMap, 2016), are initiated by a community and coordinated collaboratively by volunteers who take actively part in the design of the data collection process. Data collected by the volunteers in the context of commons-based peer production projects, is owned by the public and is openly available for use and reuse.

1.3.TYPOLOGY OF CITIZEN-CONTRIBUTED GEOGRAPHIC DATA

Since many terms are often being used interchangeably to describe geographic data contributed by citizens to Internet applications, in this section we address this terminological ambiguity by proposing a typology of CCGD. This typology, defines CCGD categories based on the characteristics of the data contributors and the purpose of the data collection activity. This is because the purpose of data collection activity has a major effect on the characteristics of the collected data, such as its quality and its cost per observation.

In the literature, two relevant typologies (Antoniou et al., 2010; Craglia et al., 2012) have been proposed with an aim to decrease that terminological ambiguity. Antoniou et al. (2010) have proposed a categorization between spatially implicit and explicit applications, based on their declared objectives. For example, as they mention "*Flickr and Picasa Web are more socially oriented, as they aimed at allowing people to share their photo albums, and thus are regarded as spatial implicit web applications*". In contrast, as Antoniou et al. (2010: 100) support "*spatially explicit applications, like Geograph and Panoramio, urge their contributors to interact directly with the spatial features (i.e. to capture spatial entities in their photos) while at the same time encourage that photos, and thus the content, be spatially distributed*". A second typology of CCGD have been proposed by Craglia et al. (2012). This typology describes four data types, which are determined based on the implicit or explicit nature of two dimensions. These dimensions are "*first, the way the information was made available, and second, the way geographic information forms part of it*" (Craglia et al., 2012: 402).

In this research, we propose a typology of CCGD which is based on the purpose of the data collection activity. This typology has partially based on the existing typologies, presented in the previous paragraph. The unique feature of the proposed typology (see Figure 1) is that there is a clear differentiation between those application collecting data for scientific purpose (VGI) and those collecting data for socially-oriented purposes (Social Geographic Data). These two data types can be produced using crowd-sourcing or commons-base peer production methods and they are defined as follows (Spyratos et al., 2014):

- Volunteered Geographic Information (VGI). The VGI category, as defined by the proposed typology, describes geographic or geo-referenced data collected during

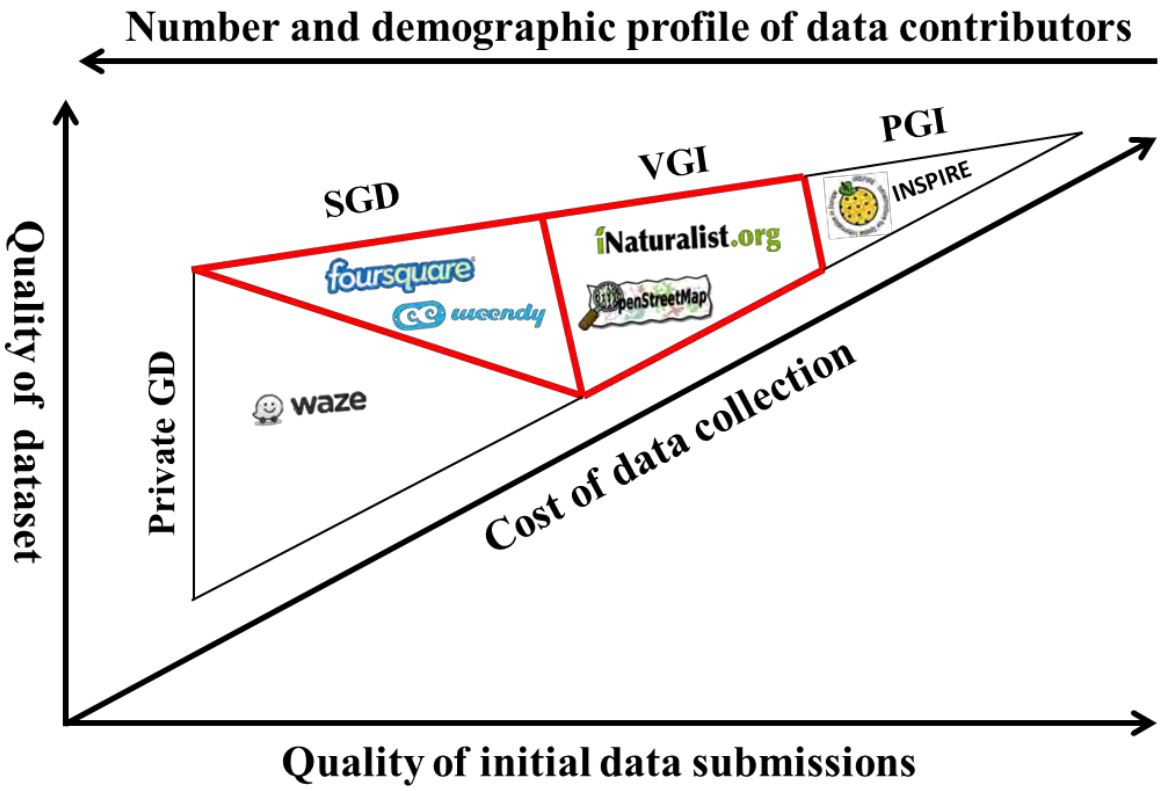
science-oriented voluntary activities. For instance, the VGI data is collected by volunteers in context of crowdsourcing projects such as the Google map maker (Google, 2016) or common-based peer production projects such as the OpenStreetMap (OpenStreetMap, 2016) .

- Social Geographic Data (SGD). This category describes geographic or geo-referenced data that, first, it has been intentionally generated by citizens or by their devices for socially-oriented purposes, and second, it has intentionally been publicly shared over the Internet. For example, this category includes, among other data, Foursquare place data (Foursquare, 2016b) and geo-located public tweets (Twitter, 2016)

Independently of the above two CCGD types, in this typology we additionally describe two other data categories. These categories are out of the scope of this PhD research, which is specifically focuses on CCGD. The reason for that, is that we want to make a clear differentiation between CCGD data and data collected exclusively by professionals or data that has data not been publicly share by their authors. To this end, we propose the following two data categories (Spyratos et al., 2014):

- Professional Geographic Information (PGI) (Parker et al., 2012). PGI category describes geographic data that has been collected exclusively by qualified professionals, such as surveyors and urban planners. PGI data can be distributed under various terms and it might be available either for a charge or openly. Openly available PGI data could be contributed to crowdsourcing or common-based peer production projects.
- Private Geographic Data (Private GD). Private GD is data collected by citizens or by devices, which can either be associated with the characteristics of an individual or of private premises or data intended for a particular person, group or service. For example, this category includes private geo-located Facebook messages, and GNSS positional data contributed to navigation services such as the Waze application (Waze, 2015).

Figure 1: Typology of Geographic Data



Source: Spyratos et al., 2014

1.4.CHARACTERISTICS OF CCGD

The aim of this PhD research is to assess whether and how CCGD can be used as auxiliary or primary input to urban planning studies. To this end, we need to analyse the characteristics of this data category, which are (Spyratos et al., 2014):

1. the number and demographic profile of potential data contributors,
2. the quality of initial data submission,
3. the overall quality of the datasets, and
4. the cost of data collection.

These characteristics are affected by a series of factors such as the purpose of the data collection activity and the utilized data collection tools. In the following sections we examine these four data characteristics in detail.

1.4.1. NUMBER AND DEMOGRAPHIC PROFILE OF POTENTIAL DATA CONTRIBUTORS

The quality of the collected data, both in terms of data accuracy and data representativity is affected by the number and importantly by the demographic profile of the data contributors (see Figure 1). The demographic profile and the number of data collectors depend, to a large extent, on the following two factors (Spyratos et al., 2014):

a) The level of technical and scientific knowledge required for data collection.
b) The time, technical equipment and other resources needed for data collection (Haklay, 2010).

These two factors create important barriers to the less technically knowledgeable citizens or to those that do not have access to the required resources, for participating in data collection activities. As a result, if the participation of marginalized members of the society in data collection activities is not supported by organizations or community-based initiatives, the collected datasets are socially-biased. This is because the areas that are on the interest of the less privileged citizens will be underrepresented in the collected datasets. For instance a study by Stephens (2013) has revealed that the percentage of male contributors to OpenStreetMap is double that the percentages of females. Another study by Budhathoki et al. (2010) revealed that only the 50% of the OpenStreetMap contributors had no experience on geographical information systems. These statistics proves, as one

would expect, that the demographic profile of VGI data collectors is not representative of the society.

The entry barriers for VGI and SGD data collectors differs. VGI is collected in the context of a scientific inquire, and a result the required scientific knowledge is much higher compared to SGD. This fact, has a result, that the number of citizens that are potentially able to collect SGD, is higher than the number of those who are able to collect VGI.

1.4.2. QUALITY OF INITIAL DATA SUBMISSIONS

The quality of initial data submissions (see bottom axes of Figure 1) refers to the quality of the data that a citizen has submitted to an application, before any automatic or manual Quality Assurance (QA) or Quality Control (QC) mechanism might applied. According to Oort (2006), the elements that determines the quality of geographic data and datasets are the following: 1) lineage; 2) positional accuracy; 3) attribute accuracy; 4) logical consistency; 5) completeness; 6) semantic accuracy; 7) usage, purpose, constrains; 8) temporal quality; 9) variation in quality; 10) meta-quality; and 11) resolution. For the description of each of the above elements we refer the interested reader to Oort (2006). The quality of initial data submissions depends on factors such as (Spyratos et al., 2014):

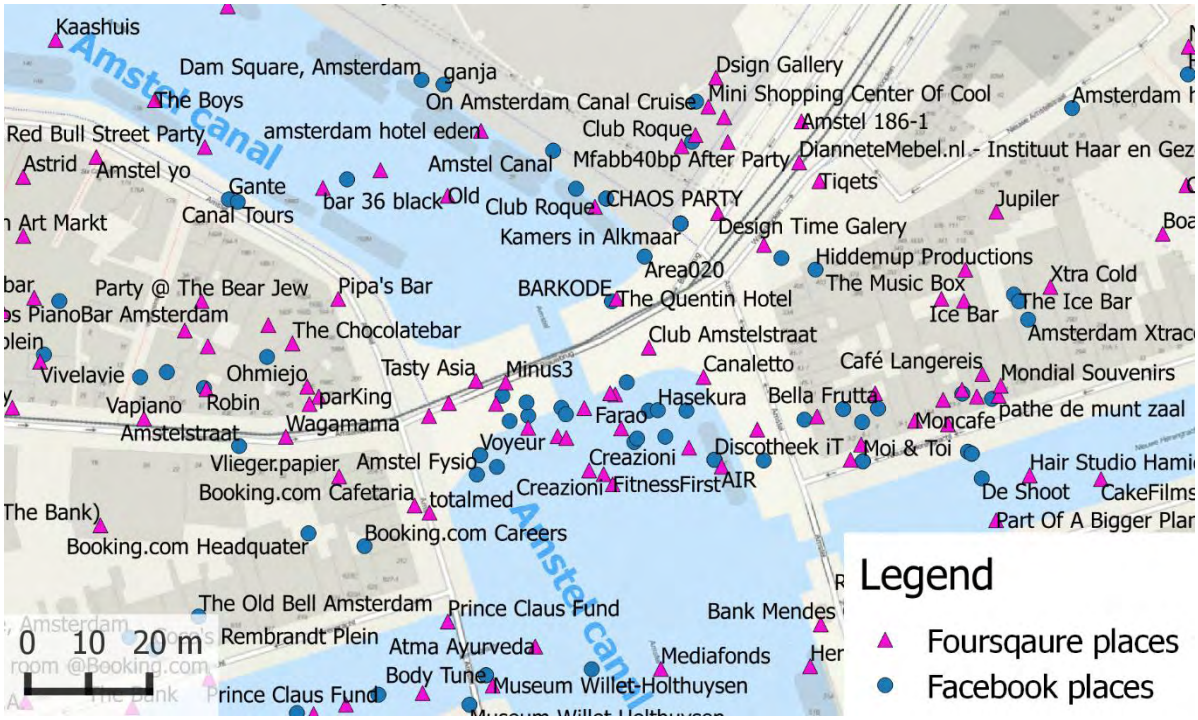
c) The desired data accuracy
d) The scientific and technical knowledge of data collectors (Delaney et al., 2008; See et al., 2013).
e) The precision and accuracy of the utilized equipment, sensors, and auxiliary data e.g. satellite images.

The factor (c) depends to a great extent, on the purpose of the data collection activity. Citizens, are typically desire higher accuracy when they collect data for scientific purposes (VGI) than when they collect data for socially-oriented purpose (SGD). The reason is that a volunteer aims at describing a geographic phenomenon or feature as accurately as possible. A user of a social media application requires a level of accuracy that will allow him/her to efficiently share a geo-referenced content (e.g. "I am in Volos and it is hot, 44C°"), regardless whether this observation depicts the reality or not. The factor (d) have clearly an important impact on the data quality, since data collected by experts is of higher quality than data collected by untrained amateurs. Finally, the factor (e) refers to the quality of the sensors and of the secondary data that are utilized by the data contributors. For example the GNSS receivers of mobile phone devices that are

typically used for determining the location of SGD, have lower accuracy compared to professional GNSS receivers (Zandbergen and Barbeau, 2011).

For example, as Figure 2 shows many Facebook and Foursquare places, which belong to the SGD category, are mistakenly pinned in the Amstel canal, Amsterdam, the Netherlands. The place data positional accuracy, though, is suitable for the intended purposes of the Facebook and Foursquare users, but it is not suitable for mapping purposes. This PhD research, through the case studies A and B, examines methods for filtering and improving the quality of this data, and thus making it suitable for urban planning purposes.

Figure 2: Many Facebook and Foursquare places are erroneously located in Amstel canal, Amsterdam, The Netherlands



Sources: Place data, Facebook Graph API and Foursquare Venues API; Basemap, MapQuest-OpenStreetMap contributors.

1.4.3. QUALITY OF DATASETS

CCGD datasets are highly heterogeneous, since they composed by data which has been contributed by citizens with different scientific and technical background, and who might use different data collection methods and equipment. As a result, the quality of these datasets might vary significant across time and space. For example, the quality of the CCGD datasets in wealthy areas, is expected to be higher than in deprived areas. This is because citizens who live in deprived areas have lower level of technical knowledge, less access to equipment and less available leisure time for data collection activities compared to members of the middle class (Haklay, 2010). The quality of datasets that consist of geographic or geo-referenced data in a given area and time it depends to a great extent on the on the following factors (Spyratos et al., 2014):

f) The quality of the initial data submissions.
g) The number and the demographic profile of contributors and the number of contributions.
h) The degree of coordination for the data collection activity.
i) The existence and the application of QA/QC mechanisms.

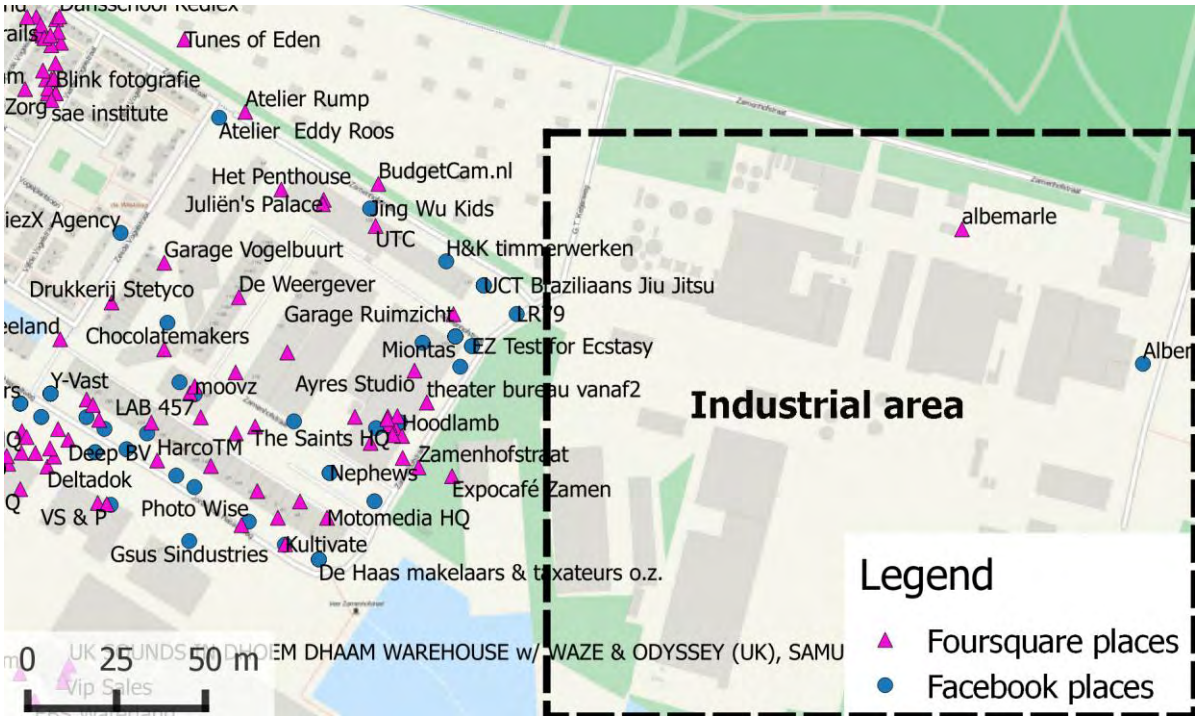
Clearly the quality of the initial data submissions (factor f), described in the previous section, determines to a great extent the quality of the datasets that they compose. The number and the demographic profile of contributors and the number of contributions (factor g), affect the thematic and spatial completeness of a dataset (Haklay, 2010; Stephens, 2013). The more the data contributors are, and the more representative of the society their demographic profile is, the higher is the possibility that all the feature and phenomena of citizens' interest will be equally represented in a dataset. The third factor that affects the quality of the citizen-contributed datasets is the horizontal or hierarchical coordination of a data collection activity. When data collection activities are coordinated, using a "top-down" or "bottom-bottom" approach, the spatial and temporal completeness of a dataset is expected to be higher. This is because, any potential unmapped feature or phenomena at given areas will be mapped by citizens who will volunteer to do so.

Finally, the last factor that affects the quality of a dataset is the application of QA/QC mechanisms. QA/QC mechanisms can be solely manual, where humans assess the truthfulness of each data submission, or they can be supported by automated procedures, where the truthfulness of a data submission is automatically assessed based on predefined

rules (for example see NBN, 2013). Manual QA/QC mechanisms can vary from being horizontally structured where users have distributed and equal authorities on editing observations to more hierarchical, where employees, community representatives, or elite users have increased authorities compared to average users. The existence of QA/QC mechanisms increase the quality of the data. For example, a dataset that its quality is controlled by horizontally structured QA/QC mechanism, due to the application of the so called "Linus' law", is expected to be of higher quality compared to a dataset that does not (Haklay et al., 2010). Linus' law" (Raymond, 2001), states that the higher the number of users or contributors of a product is, for instance the users of a map, the higher is the probability that an error will be identified and fixed by someone. Additionally, when citizens are able to assess the truthfulness of a data submission and the credibility of a data contributor, each CCGD can be associated with an indication of truthfulness, a rating. Finally, attention need to be given to the differences between VGI and SGD, as on SGD the rating refers to the attractiveness of a place, photo or measurement and not to its accuracy.

Many studies, proved that the quality of VGI datasets, and in particular of OpenStreetMap datasets is lower compared to PGI (Girres and Touya, 2010; Haklay, 2010; Koukoletsos et al., 2012). However, it worth mentioning, that many VGI datasets are also composed of PGI, since users are uploading open public data collected by professional mapping agencies to VGI applications. Regarding the quality of SGD datasets, the Antoniou et al. (2010) study demonstrated that the spatial distribution of photo on Flickr, which is a SGD application for photo sharing, has several peaks in populated and popular areas, while the respective spatial distribution of Geograph photos, which is a VGI platform for the collection of geographically representative photographs over the United Kingdom (UK), is more flat over the regions of interest. The reason of such difference is, first the aim of Geograph platform, and second that the spatial distribution of SGD observations is more likely to be limited to the attractive areas of the users' activity space. For example, Figure 3 shows that in an industrial area of Amsterdam there are only two places, while in an area with office and retail facilities there are many. This is because, users of social media they do not find attractive to declare their presence in industrial facilities.

Figure 3: Few Facebook and Foursquare places are located in the industrial area, at the right side of the figure, while many places that represent offices and retail facilities are located at the right side of the figure.



Sources: Place data, Facebook Graph API and Foursquare Venues API; Basemap, MapQuest-OpenStreetMap contributors.

1.4.4. COST OF DATA COLLECTION

The cost of data collection per observation varies based on the type of the collected data. For instance for the collection of PGI, qualified staff is employed, while for the collection of VGI and SGD, volunteers and users of Internet applications are offering for free their services. The collection of VGI has a cost that usually includes the cost for the coordination of the voluntary activities and for the utilized equipment. In theory, equipment of higher quality have also higher cost. For example, professional GNSS receivers are more accurate and expensive than those built-in mobile phones. Regarding the SGD, there are no coordination costs, since SGD data collection activities are not coordinated by any organization. In addition, since SGD data collectors are not desire high precision, the cost of the utilized equipment is lower compared to the one used for VGI. As results, the collection of SGD incurs less costs that the collection of VGI.

1.5.OUTLINE OF THE THESIS

The remainder of the PhD thesis is structured as follows. Chapter 2 describes the two case studies, which are used for answering the two research questions of this PhD research. In Chapter 3 we present the data used for these case studies. Chapter 4 describes the methodology and the methods used for each case study. In Chapter 5 we present the results of both case studies, and we test the reproducibility of the case study A in the city of Varese in Italy. In Chapter 6, we discuss the methodologies and their limitations, we present our conclusions and future work directions, as well as we propose a generic framework for using crowd-sourced data.

2. CASE STUDIES

In this Section we present the two cases studies, selected for answering to the two research questions of this PhD. The first case study, presented in Section 2.1, is about the reuse of data contributed by citizens to social media applications (SGD) for mapping building uses. The second case study, presented in Section 2.2, is about the reuse of data contributed by citizens to social media applications (SGD) for estimating citizens' satisfaction with regard to urban facilities and services.

2.1. CASE STUDY A –ESTIMATING BUILDING BLOCK USE

Data about existing LU is essential for urban planning, transportation and real estate studies. LU patterns have a major effect on the urban environment since they affect the land prices (Cheshire and Sheppard, 1995), the crime rates and crime types (Stucky and Ottensmann, 2009), and the factors that affect public health (Frank et al., 2006). Additionally, LU patterns affect the usage of urban infrastructure, for example, the demand for transport (Litman and Steele, 2014) and energy (Pérez-Lombard et al., 2008).

Despite its usefulness, for many urban areas, LU data is not available, is outdated, or is not enough detailed in terms of LU categories. For example at European scale, three projects collect and publish LU and/or Land Cover (LC) data. However, none of these projects provides data that can be used for identifying the LU of built up areas at a very detailed level. These projects are:

1. The European Urban Atlas project that provides inter-comparable, high-resolution LU and LC data derived by the analysis of satellite imagery, topographic maps and navigation data. Urban Atlas LC/LU data is available for 305 urban major European urban agglomerations for the reference year 2006 (European Union, 2011).
2. The CoORdination of INformation on the Environment (CORINE) project (Büttner and Barbara, 2007), which provides LC data derived from the interpretation of satellite images. CORINE LC data cover are available for many European countries for the reference years 1990, 2000, 2006 and 2012 (EPA, 2014).

3. The Land Use/Cover Area frame Survey (LUCAS), is in contrast to the CORINE and the Urban Atlas projects, a ground survey. Land use and land cover data is collected by surveyors through in-situ observations. The latest LUCAS survey was carried out at European Union (EU) 27 member states level in 2012 and in total 270 000 point locations were visited (Eurostat, 2015).

The main reason for the absence of LU data, is that its collection through traditional methods, i.e. in-situ LU surveys, requires a significant volume of human and financial resources. Other methods for the collection of LU data, which do not require an extensive in-situ LU survey, are the analysis of remotely sensed images and of business registries. Using remotely sensed images is possible to differentiate the areas with different LC such as built-up areas, the agricultural land, the green spaces and the water bodies, but is not possible to identify the use of the built-up areas. Using business registries is feasible to identify commercial facilities within urban fabric. However, is not possible to identify other types of facilities which are often not included in business registries such as schools and sport facilities.

In this case study, considering the high cost of in-situ LU surveys and the lack of detailed LU data, we propose the production of low cost LU data based on data contributed by citizens to social media applications, namely SGD.

To this end, we purpose the use of place data from the Foursquare application for the estimation of the LU of built-up areas. Foursquare users are contributing information about places, for example, the name of a place, its type (hotel, cafe, etc.) and its location. The places that are contained within a building, define its use. Thus we assigned LU categories to buildings by taking into consideration the type of the places that area contained in them.

As main study area we selected the city of Amsterdam, the Netherlands. The reason for that choice is for that city there is available detailed and relatively updated in-situ collected LU data. This data is essential for assessing the accuracy of the estimated, using Foursquare place data, LU data. In Amsterdam, and especially in its historic centre most of the buildings are very narrow. That has as a consequence many Foursquare places to be falsely described to be located in the nearby buildings. To overcome these problems the allocation of LU was performed at the spatial scale of urban building block level and not at building level.

Finally, we tested the reproducibility of the methodology by repeating it in the city of Varese in Italy. For this city, due to the absence of existing LU data, we have performed an in-situ LU survey on a sample of 150 randomly selected building blocks.

2.1.1. RELATED WORK

Previous studies have explored and developed methodologies for estimating LU using geographic data generated by citizens or by their devices. Several studies have used crowd sourced feature data from OpenStreetMap for estimating land and building use. Huang et al. (2013) and Fan et al. (2014) developed methodologies for identifying the primary use of buildings, by using the geometric and topological characteristics of the building footprints. Jokar Arsanjani et al. (2013) estimated LU and LC patterns using OpenStreetMap feature data and they evaluated the accuracy of these estimations by comparing the results with Urban Atlas (European Union, 2011) data.

Other studies used data from social media application such as Twitter and Foursquare for estimating LU in large-scale urban areas. Frias-Martinez et al. (2012) used the intensity and the time of geo-located tweets for identifying large scale urban clusters with commercial, residential, industry or recreational LU. Noulas et al. (2013) used the category and the check-in number of Foursquare places for estimating the type of the dominant activity in neighborhoods of Madrid and Barcelona. These two research studies are relevant to this case study but they differ in two important methodological approaches. First they do not deal with mixed LU, and second they do not estimate LU in a detail scale e.g. building or building block. Finally, several studies (Pei et al., 2014; Soto and Frias-Martinez, 2011; Toole et al., 2012) have used mobile phone activity data for estimating dominant land use types of almost neighborhood-level areas.

To the best of our knowledge, there is no study that has used place data for estimating LU at building or building block level.

2.2.CASE STUDY B - EVALUATING SERVICES AND FACILITIES

Approximately 54% of the world's population currently lives in urban areas (United Nations, 2015). The quality and the availability of urban facilities and services such as public spaces (Cattell et al., 2008; Giles-Corti et al., 2005), sport facilities (Powell et al., 2006; Prins et al., 2010; Van Lenthe et al., 2005) and cultural facilities (BOP Consulting, 2013) play an important role in wellbeing of urban dwellers. Statistics about citizens' satisfaction with regard to urban facilities and services are required by urban planners, policy makers and most importantly by society. The availability of such statistics can facilitate evidence-based decision making processes, and it can be proved to be a step towards citizens' participation in the governance of urban areas. Policies that take into account and respond to societal concerns can contribute to improving the quality of life of urban dwellers.

Often statistics about citizens' satisfaction with regard to urban facilities and services are not available, are outdated, or are not relevant to the public concerns. The estimation of citizens' satisfaction using questionnaires, demands a considerable amount of human and financial resources. This cost can be very high when the target of the survey is to compare citizens' satisfaction levels across many urban cities. Since this cost cannot be drastically reduced, there is an increasing need for maximizing the societal benefits of these surveys. For that purpose, low cost methods that estimate citizens satisfaction and identify which questions, when and where are relevant to the public can be used to assist researchers on better designing public opinion surveys.

The rational of this case study is to investigate whether we can use crowd-sourced place data from social media applications (SGD) for estimating citizens' satisfaction with regard to urban facilities and services.

To this end, we propose two indicators for estimating citizens' satisfaction using data from the Foursquare social media application. Both indicators are based on our hypothesis, which is that the higher the number of places that belong to a facility or service type on social media is, the higher the citizens' satisfaction rate with regard to this facility or service type would be. To test that hypothesis, data from the Flash Eurobarometer survey No 336 about "Quality of Life in European Cities" (European Commission, 2013), which we henceforth refer to simply as Eurobarometer were used as reference for evaluating the accuracy of the proposed indicators.

Several studies have used social media data to study the environment. A study by Venerandi et al. (2015) and its preliminary study by Quercia and Saez (2014) used Foursquare place data for estimating socioeconomic deprivation on three UK urban areas. Preotiuc-Pietro et al. (2013) used Foursquare place data to compare the urban landscape of seventeen cities across the United States. Salesses et al. (2013) measured peoples' perception of safety, class and uniqueness across four cities by requesting volunteers to rate crowd-sourced geo-tagged images. Finnaly, a study by Floris and Zoppi (2015) estimated tourists' satisfaction about touristic destinations and services in Sardinia, Italy, using crowd-sourced data collected by travel and booking Internet applications.

To the best of our knowledge, there is no study that estimates citizens' satisfaction with regard to urban facilities and services using place data from social media applications.

3. DATA

In this Section we present the data used in this PhD research. First, in Section 3.1, we present the Foursquare place data which is used in both case studies. In Section 3.2. we present the data used in the case study A, and in Section 3.3 the data used in the case study B.

3.1. FOURSQUARE PLACE DATA

In this PhD, we reused data contributed by citizens to the Foursquare social media application for the production of low cost LU maps and for the estimation of citizens' satisfaction with regard to urban facilities and services. According to the typology proposed in Section 1.3, this type of data is referred to as Social Geographic Data (SGD). In Section 1 we have presented the characteristics of that citizen-contributed data and how they differ from data collected exclusively by professionals.

Foursquare (Foursquare, 2016a) is a social media application that allows users to discover and evaluate places around the world. As of January 2016, Foursquare had 50 million monthly active users (Isaac, 2016). Foursquare users are contributing information about places that either do not exist in the Foursquare database or their description is not accurate or updated. Foursquare place data is up-to-date, is available at global scale and at limited cost to third parties. This data includes the name, the geographic location, and importantly for both case studies the category of a place such as park or football court. The Foursquare place categories are classified based on a detailed and well-structured place type classification. The description of Foursquare places also includes the number of Foursquare users that have expressed their subjective positive evaluation, widely known as likes, the number of Foursquare users that have declared a visit in a Foursquare place, widely known as check-ins, and the number of total check-ins. In July 2014, the check-in feature of Foursquare has been moved to a separate application which is interconnected to the Foursquare and is named the "Swarm" (Foursquare, 2014). Foursquare users determine the location of a place by adding a marker of point geometry on top of a web map. However, places are surfaces and not abstract points of the Euclidian geometry (Burrough and Frank, 1996). As a result the determination of the point location of the Foursquare places is based on the Foursquare users' subjective judgment.

Two major limitations are introduced in both case studies of this thesis by the use of Foursquare application as a data source. The first limitation is introduced by the fact that the demographic profile of the users of location-based social media applications, such as the Foursquare, is not representative of the society. This is because the age group 18 to 29 year old is overrepresented compared to other age groups (Zickuhr, 2013). Thus, it is expected that the place data will represent mostly the activity space of young smartphone owners that use Foursquare the most. As a result, this demographic bias which is equally distributed across space and time, it causes a systematic error in the Foursquare data. The second limitation is introduced by the fact that Foursquare users decide independently which places, of what type and from which area, will be added in the Foursquare dataset. Therefore, places of the Foursquare users' main interest, are better represented in the Foursquare place dataset compared to other types of places. As proved by this PhD research and shown in Figure 11 Foursquare users are biased in favour of contributing content about commercial or recreational places.

3.2.CASE STUDY A - ESTIMATING BUILDING BLOCK USE

In this Section we present the data which was used for the LU estimations and for the evaluation of their accuracy. In Section 3.2.1, we present the citizen contributed data used for the estimation of LU, and we explain why we selected to use Foursquare as a data source. In Section 3.2.2, we present the official data used for the Amsterdam study area and in Section 3.2.3, the official data used for the second study area, the city of Varese.

3.2.1. FOURSQUARE DATA

Place data is a data type that can be reused for the production of low cost LU maps. This is because each place is associated with the description of its type, for example restaurant or retail shop. Thus, the type of a place can be used for defining the use of the surface that this place covers. In this study, we define places as surfaces where socio-economic activities occur. Places, from the humanistic perspective, are defined as the "*enclosed and humanized space*" (Tuan, 2001). In line with the above definition, places can be described as spaces enriched with human experiences and meaning (Couclelis, 1992). As a result, we expect, that place data will reflect how space is experienced and subjectively perceived by the users of social media applications.

Two well-known social media applications were taken into consideration as potential sources of citizen-contributed place data. These are the Facebook (Facebook, 2016) and the Foursquare (see Section 3.1) applications. The Facebook place data is accessible via the Facebook Graph Application Programming Interface (API) (Facebook, 2015) and the Foursquare place data via the Foursquare venues API (Foursquare, 2015b). As of February 2015, there were available 24,486 Facebook places and 37,482 Foursquare places within the Amsterdam study area. Due to the higher volume of Foursquare place data, the Foursquare application was selected to be used in this research.

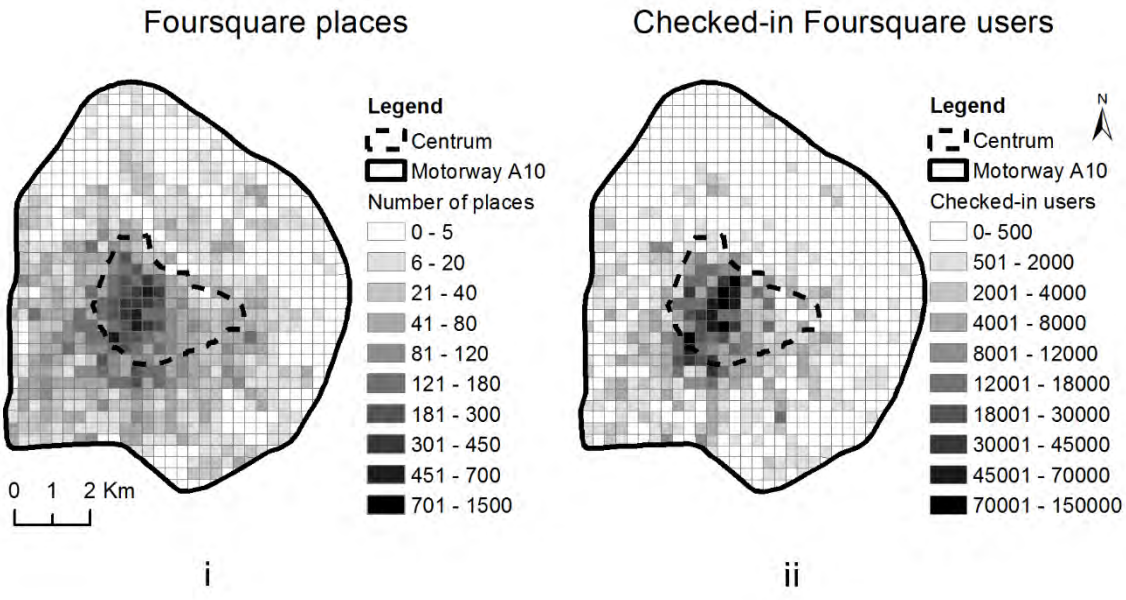
The combined use of both Facebook and Foursquare place data was not implemented, since it will introduce methodological and semantic problems that will affect the consistency of the methodology. This is due to the following two reasons (Spyratos et al., 2016):

- Facebook uses a different place type classification from Foursquare. Consequently, the combination of the two sources will introduce semantic problems.

- Facebook users determine the location of places using the GNSS receivers of their smartphones. As a result, these places are mostly located outdoors where GNSS signal is available. On other hand, Foursquare users determine the place locations on top of web maps and as a result they are free to locate a place wherever they prefer. This difference will introduce problems in the determination of the value of the d parameter, which is described in Section 4.1.2.

In Figure 4, are shown the spatial distributions of the Foursquare places (i) and of Foursquare checked in users (ii), in the city of Amsterdam. Not surprisingly, the majority of both places and checked in users are located in the city centre of Amsterdam, where the most popular and populated areas of the city are located.

Figure 4: Density maps based: (i) on the total number of places that are located in each pixel; and (ii) on the total number of Foursquare users that have checked in places that are located in each pixel. The study area is divided in pixels of 300m by 300m size.

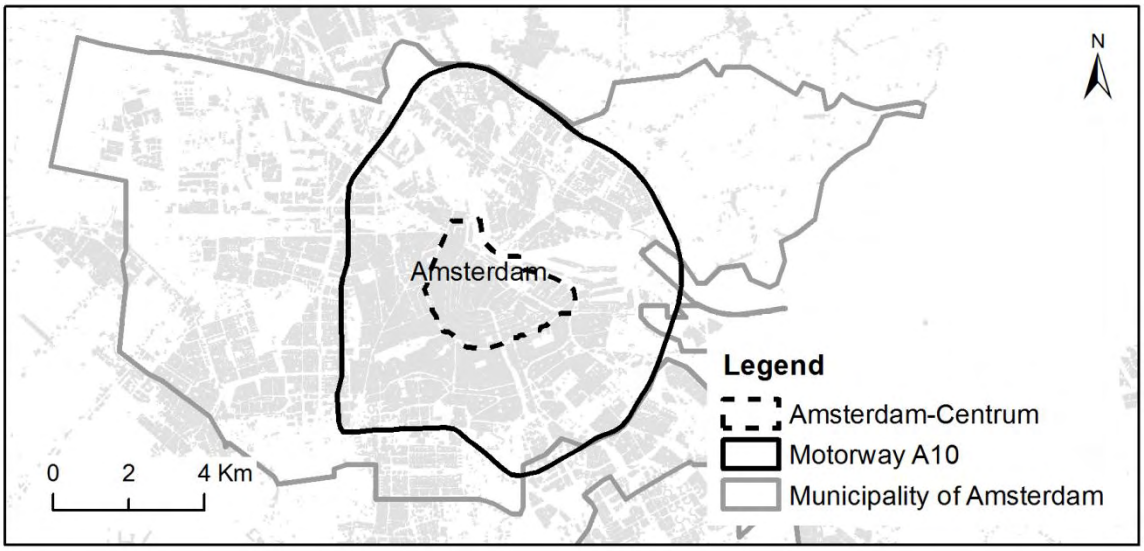


Source: Spyratos et al., 2016

3.2.2. AMSTERDAM STUDY AREA

The selection of the main study area was based on the availability of high quality, detailed and updated reference LU data. This is because, first reference LU data will be used for assessing the quality of the LU data which is estimated using citizen-contributed place data, and second citizen contributed data is available for almost all the cities globally. To this end, we have reviewed the availability of LU data and its quality for many major European cities, including the cities of Athens and Thessaloniki, and we have decided to use as a main study area for the case study A, the city of Amsterdam in the Netherlands. This is because, the city of Amsterdam, offers relatively updated and detailed data on built up area LU. The exact study area, is the area of the Amsterdam city which is bounded by the A10 motorway (see Figure 5) and it has a size of 72.12 km^2. The reason for that selection is that the A10 motorway creates a physical boundary that encloses both sparsely and densely built-up areas, which are well representing the diversity of urban environment.

Figure 5: Main Study area: The area bounded by the A10 motorway.



Source: Spyratos et al., 2016

In Amsterdam study, a reference Building Block Use (BBU) dataset was used for the evaluation of the accuracy of the estimated BBU dataset. This reference BBU dataset was produced using two official datasets (see Section 3.2.2.1), the official buildings dataset and the non-residential use dataset of Amsterdam. These datasets belong to the Professional Geographic Information (PGI) category, defined in Section 1.3. This is

because this data has been collected by professionals and in addition the mapping activities have been centrally coordinated by a mapping agency. Both the building and the non-residential use datasets are presented in more detail in the following paragraphs.

The buildings dataset is the "*Basisregistraties Adressen en Gebouwen*" and it was retrieved from the geo-data portal of the VU University Amsterdam (VU geoplaza, 2015). In this study, we used the data version which was available on the geo-data portal on February 2015 and which had been last updated with new buildings in August 2013. The original building dataset includes 143,804 building units within the study area. In this study, we assign uses to building blocks and not to buildings. A building block is defined as a unit of attached buildings which are separated by others by a street or other open space. To derive the building block dataset we merged the attached buildings. Moreover, we removed building blocks with less than 100m^2 footprint area, since they mostly represent garages and sheds of residential premises. The cleaned building block dataset formally represented by the set $\boldsymbol{b}$ in (1), has 9,827 members (Spyratos et al., 2016).

$$\boldsymbol{b} = (b_1, b_2, \ldots., b_{m-1}, b_m), m = 9{,}827\ , \qquad (1)$$

where b is a building block, and m is the total number of building blocks within the Amsterdam study area.

The non-residential use dataset is the "*viet-woonfuncties Functiekaart*" and it was retrieved from the geo-data portal of the city of Amsterdam (City of Amsterdam, 2015). The non-residential use dataset includes surfaces of buildings with non-residential use. A surface is defined as the area of the footprint of one or more buildings that a facility, e.g. a supermarket, covers. In this study we used the data version which was available on the Internet on February 2015, and which included building surfaces with non-residential use that had been surveyed during the time period March 2010 - November 2014. Within the study area, the non-residential use dataset includes 23,313 surfaces and is represented by the set $\boldsymbol{s}$, which is defined in (2) (Spyratos et al., 2016).

$$\boldsymbol{s} = \{s_1, s_2, \ldots., s_{n-1}, s_n\}, n = 23{,}313 \qquad (2)$$

where s is a surface of the non-residential use dataset, and n is the total number of these surfaces within the Amsterdam study area.

Every surface of the non-residential use dataset $s_i(b^s, a^s), i = 1 \dots n$, that is a member of the $\boldsymbol{s}$ set, has the attributes b^s and a^s, where (Spyratos et al., 2016):

- b^s is the building block that geometrically contains s_i. The b^s attribute is a member of the $\boldsymbol{b}$ set which is defined in (1).
- a^s is the official Amsterdam LU category of the s_i surface. The a^s attribute is a member of the $\boldsymbol{a}$ set which is defined in (3).

The official Amsterdam non-residential LU classification is represented by the $\boldsymbol{a}$ set, has 8 members and it is described in (3).

$$\boldsymbol{a} = \{a_1, a_2, \dots, a_7, a_8\} \quad (3)$$

Every $a_z,\ z = 1 \dots 8$, represents one of the following LU categories (Spyratos et al., 2016). Details about the types of facilities that these 8 LU categories and their subcategories describe, can be found in City of Amsterdam (2011).

- $a_1 =$ *"Bedrijven"* / *"Industries"*,
- $a_2 =$ *"Kantoren"* / *"Offices"*,
- $a_3 =$ *"Detailhandel"* / *"Retail"*,
- $a_4 =$ *"Horeca"* / *"Horeca, Hotels, restaurants & cafes"*,
- $a_5 =$ *"Maatschappelijk"* / *"Societal"*,
- $a_6 =$ *"Vrije tijd"* / *"Leisure"*,
- $a_7 =$ *"Parkeren & Openbaar vervoer"* / *"Parking & public transport"*,
- $a_8 =$ *"Opslag & onduidelijk"* / *"Storage & unclear"*.

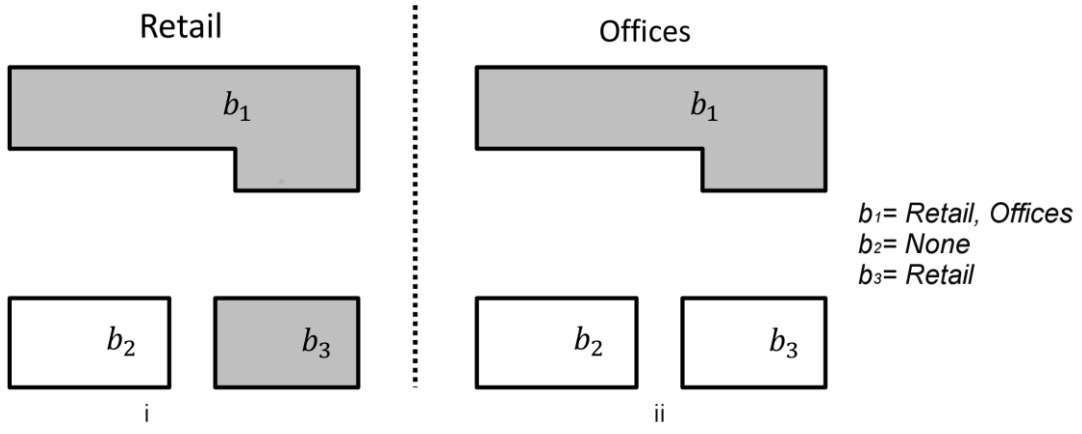
3.2.2.1. PRODUCTION OF THE REFERENCE BBU DATASET

For the accuracy assessment of the estimated BBU dataset, whose production is described in Section 4.1.2, we used a reference BBU dataset. This reference dataset needs to have the same properties with the estimated BBU dataset so as to be comparable. The reference BBU dataset was produced using the building block dataset and the non-residential use dataset, which were both described in Section 3.2.2. Buildings and building blocks typically host various types of facilities. For example, a building in the centre of a city, may host retail facilities on the ground floor and offices on the upper floors.

An urban building block $b_k, k = 1 \dots m$, can contain many non-residential surfaces s_i of the same LU category or of different LU categories. We therefore calculated the reference BBU dataset $\boldsymbol{R}_{z,k}$ using formula (4), which computes for each building block b_k whether there exist or not s_i surfaces for each of the 8 a_z LU categories (Spyratos et al., 2016). For example, as shown in Figure 6 the building block b_1 has both the "*Retail*" and the "*Offices*" use, while the building b_3 have only the "*Retail*" use.

$$\boldsymbol{R}_{z,k} \begin{cases} 1 \; if \; \exists \; s_i(b^s, a^s) \in \boldsymbol{s}, & where \; a^s = a_z, \quad b^s = b_k \\ 0 \; if \; \nexists \; s_i(b^s, a^s) \in \boldsymbol{s}, & where \; a^s = a_z, \quad b^s = b_k \end{cases} \qquad (4)$$

Figure 6: Building blocks with "*Retail*" LU (i) and "*Offices*" LU (ii). A building block can have multiple LU categories as long as it contains at least one surface of these LU categories.

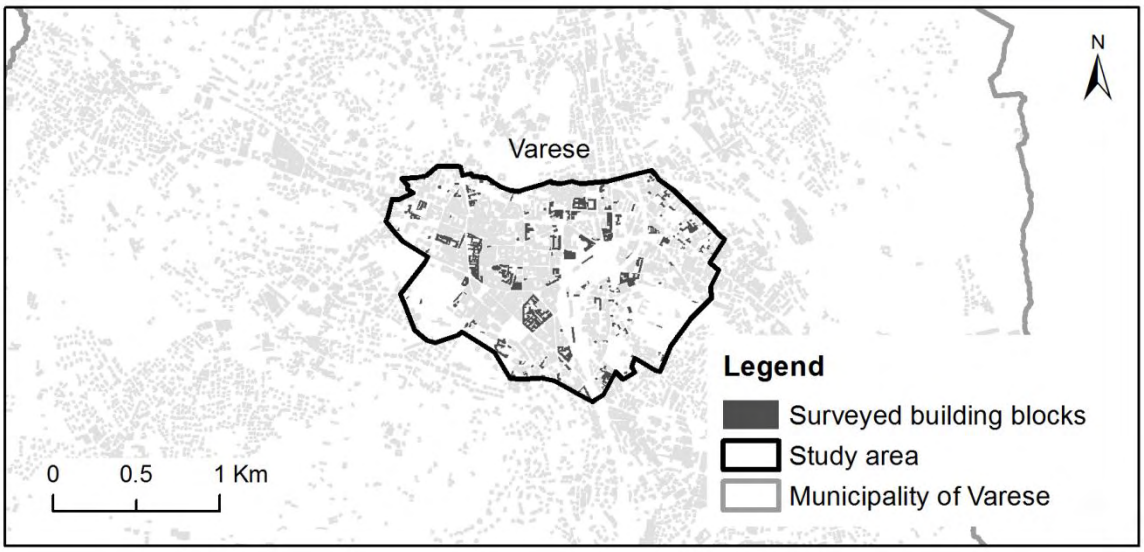


Source: Spyratos et al., 2016

3.2.3. VARESE STUDY AREA

The city of Varese, in Italy, was selected as a second study area (see Figure 7). The reason for this selection is to test the reproducibility of the proposed methodology in an area with different urban morphology and characteristics. In comparison to the city of Amsterdam, the city of Varese has almost 10 times less population and no canals. The Varese study area covers 1.93 km^2, size which is almost 36 times smaller than the one of Amsterdam.

Figure 7: Varese study area. The buildings in the Municipality of Varese are highlighted in light grey and the boundary of Varese Municipality in black. In dark grey are the 150 randomly selected building blocks for which a LU ground survey was performed.



Source: Spyratos et al., 2016

The municipality of Varese has provided us the city's building dataset. In total there were 5,411 building units within the study area. As in Amsterdam, in Varese study we merged the attached buildings to derive the building block dataset and we further removed building blocks with less than 100m^2 footprint area. In total 641 building blocks remained within the study area. For the city of Varese we have decided to perform a ground survey of non-residential use. This is, first, because there was no LU data available, and second, because we wanted to use updated data about existing LU. The LU categories that were used are the 8 non-residential LU categories used in the Amsterdam study area (see City of Amsterdam, 2011). The LU survey was performed in July 2015 on a random sample of 150 building blocks, which counts for the 23.4% of the total. During the ground survey, we recorded the non-residential use that were visible at the exterior of the buildings or they were described in the mailboxes and doorbells of the buildings.

3.3.CASE STUDY B - EVALUATING SERVICES AND FACILITIES

In this case study, we assess the feasibility of estimating indicators that reflect citizens' satisfaction with regard to urban facilities and services using data from social media applications (SGD). For estimating these indicators we selected Foursquare social media application as a data source instead of others for two reasons. First, the Foursquare, since it is exclusively a location-based social media application, it includes higher number of place data compared to alternative sources such as the Facebook (see Section 3.2.1). Second, Foursquare places are classified based on a detailed and well-structured place type classification. This fact facilitates the identification and classification of places that belong to each of the Eurobarometer categories under investigation.

Moreover, the estimated indicators were compared with official statistics from the Eurobarometer survey (European Commission, 2013). For this comparison, we selected 17 cities which are listed in Table 1. The criterion for selecting these 17 cities was to include a representative sample of large and very large cities of Europe. The administrative boundaries of these urban cities and their population of 15 years old or more are available in the appendix of the Eurobarometer survey (European Commission, 2013). In the following two Sections we describe the data that we used in this study in more detail.

Table 1: The number of Foursquare places in cities under investigation, and the population of age 15 years old and over.

	Urban City	Population of age 15 years and over	Number of Foursquare places
1	Lisboa	477,239	26,366
2	Praha	1,077,005	63,397
3	Helsinki	514,611	42,517
4	Brussels	916,829	58,600
5	Amsterdam	661,407	47,320
6	Barcelona	1,418,437	67,076
7	Athens	659,664	24,519
8	Kobenhavn	464,858	31,919
9	Sofia	1,055,205	38,404
10	Napoli	807,815	9,502
11	Warszawa	1,502,571	46,900
12	Paris	1,844,243	92,675
13	London	5,807,285	170,884
14	Berlin	3,035,226	79,082

15	Hamburg	1,557,324	42,922
16	Budapest	1,550,299	72,673
17	Bucuresti	1,718,888	45,020

3.3.1. FOURSQUARE DATA

Foursquare place data, described in Section 3.1, was collected for the 17 European cities. To this end, first we downloaded the administrative boundaries of each urban city from various local and national geo-data portals. Then using the Foursquare Venues API (Foursquare, 2015b), we harvested data about places that are located within the administrative boundaries of the selected 17 urban cities. For example in Figure 8, we present density maps of Foursquare places that are located within Athens' and Brussels' administrative boundaries. Due to API rate limits, the data was collected gradually city by city, between the 26th of January 2015 and the 4th of May 2015. In total there were 959,776 places within the 17 selected urban cities. As shown in Figure 9, the urban population of fifteen years and older is strongly correlated with the number of total Foursquare places in each city. The value of the coefficient of determination is $R^2 = 0.813$ ($p \leq 0.01$), which means that the population of age fifteen years and older explains 81.3% of the total variation in the number of places.

Figure 8: Density maps based on the total number of places that are located in each pixel. The cities of Brussels and Athens are divided into pixels of 300m by 300m size.

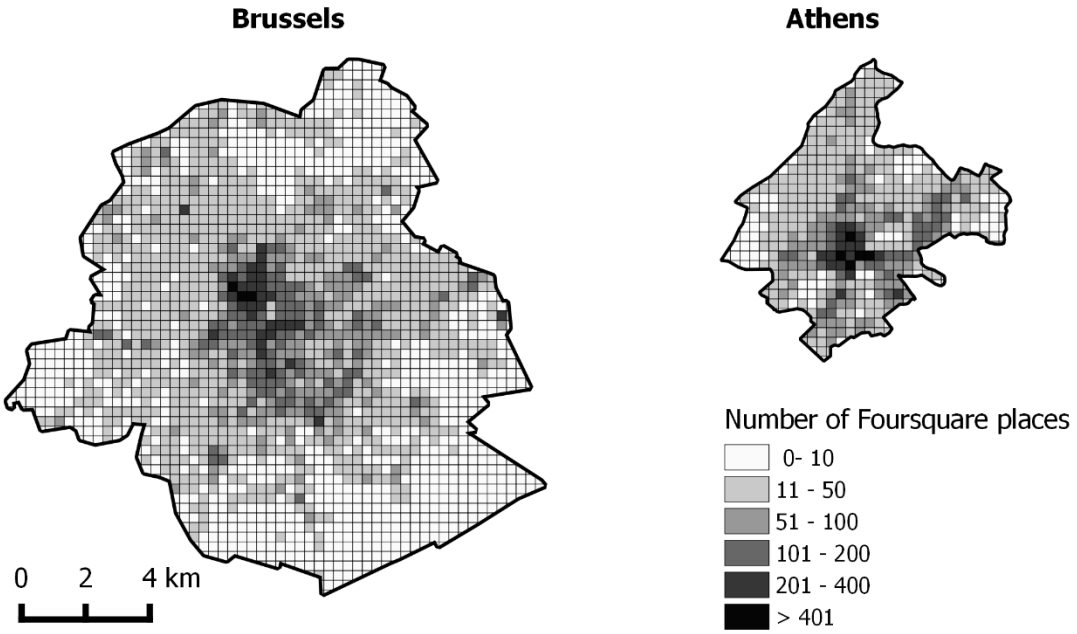
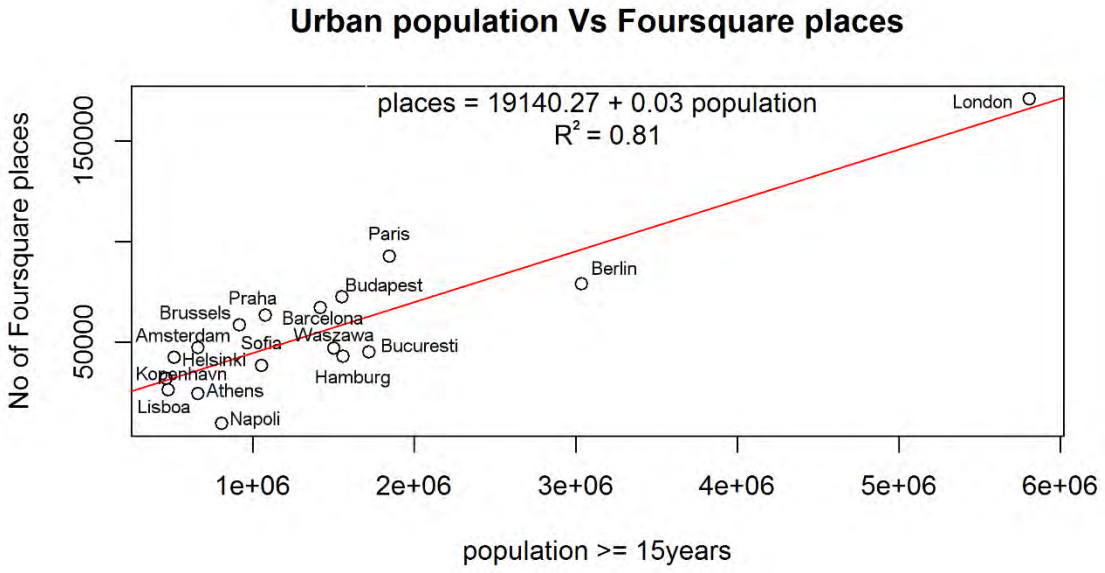


Figure 9: Linear correlation between the Foursquare places and the population aged 15 years and older.



3.3.2. EUROBAROMETER DATA

Statistics available in the Eurobarometer survey were used as a reference for evaluating the accuracy of the estimated indicators. The Eurobarometer survey was performed between the 15th of November and the 7th of December 2012. All interviews were carried out via telephone. The Eurobarometer covers the population of 83 cities, aged 15 years and older. For more details about the sampling methodology followed in the Eurobarometer survey we direct the interested reader to the technical specification section located in the Annex of the Eurobarometer (European Commission, 2013). The participants in the Eurobarometer were asked to respond to 4 groups of questions. From these groups relevant to this study is the first group of questions which is the *"Generally speaking, please tell me if you are very satisfied, rather satisfied, rather unsatisfied or not at all satisfied with each of the following issues in [CITY NAME]?".* This group of questions asks the respondents to answer twelve topics. Relevant to this study and measurable using the proposed indicators are the following eight topics:

i. Public spaces such as markets, squares and pedestrian areas
ii. Green spaces such as parks and gardens
iii. Public transport, for example the bus, tram or metro
iv. Health care services, doctors and hospitals
v. Sports facilities such as sport fields and indoor sport halls
vi. Cultural facilities such as concert halls, theatres, museums and libraries
vii. The state of the streets and buildings in your neighbourhood
viii. Schools and other educational facilities

For all the above issues, the available Eurobarometer statistics include the percentage of respondents for each city that have replied one of the five possible answers. These answers are the "*Very satisfied*", "*Rather satisfied*", "*Rather unsatisfied*", "*Not at all satisfied*" and "*Don't Know or No Answer*". In this study we have used the percentage of "*Very satisfied*" citizens and the percentage of the "*Totally satisfied"* citizens which is the sum of the percentages of the "*Very satisfied*" and "*Rather satisfied*" citizens.

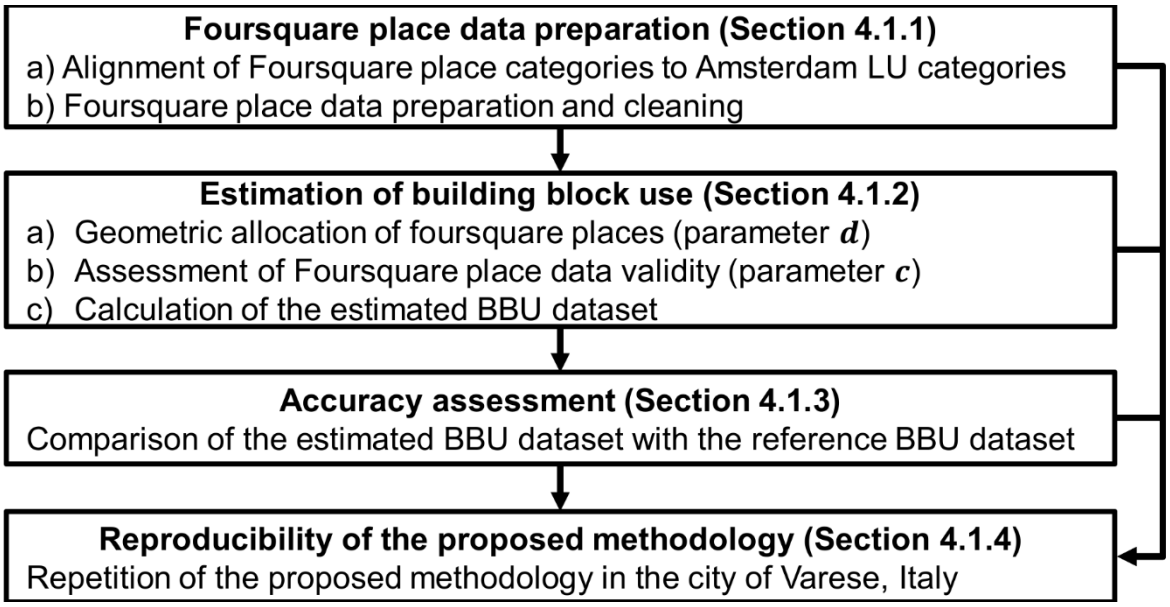
4. METHODOLOGY

In this Section, we present the methodology used in this research. First, in Section 4.1, we present the methodology used in the case study A (see Figure 10). In Section 4.2, we present the methodology used in the case study B (see Figure 18).

4.1. CASE STUDY A - ESTIMATING BUILDING BLOCK USE

In this Section we present the methodology followed for the case study A – "Estimating building block use". As shown in Figure 10, the methodology consists of four steps. The first step, described in Section 4.1.1, is the preparation of Foursquare place data. This step includes the alignment of the Foursquare place classification to the Amsterdam non-residential LU classification and the cleaning of the Foursquare place dataset. The second step, presented in Section 4.1.2, describes the generation of the estimated BBU dataset. For that purpose we have introduced two parameters, the $\boldsymbol{d}$ and the $\boldsymbol{c}$. The third step, presented in Section 4.1.3, includes the definition of the method used for the comparison of the estimated BBU dataset with the reference BBU dataset. Finally the last step, in Section 4.1.4, deals with the reproducibility of the proposed methodology in other urban areas.

Figure 10: Methodology followed for the preparation of Foursquare place data, for the estimation of BBU, for assessing the accuracy of these estimations and for testing the reproducibility of the proposed methodology.



4.1.1. FOURSQUARE PLACE DATA PREPARATION

In the Foursquare place data preparation step, we performed two operations so as to use the Foursquare place data for calculating the estimated BBU dataset. First, we aligned the Foursquare place classification (Foursquare, 2015a) to the Amsterdam non-residential LU classification (City of Amsterdam, 2011). Second, we cleaned the Foursquare place dataset by removing data that belong to Foursquare categories that were not aligned to the Amsterdam non-residential LU classification.

The Foursquare place classification serves different purposes from the Amsterdam non-residential LU classification. The former is used for classifying the type of places so as Foursquare users can easily discover them. As a result, activities that are on the main interest of Foursquare users are described in more detail. For example, as shown in Table 2 the Foursquare place classification is very detailed in place categories that belong to the "*Hotels, restaurants & cafes* (a_4)*"* and the "*Retail* (a_3)*"* LU categories. The Amsterdam non-residential classification is used for LU mapping purposes and, as a result, it reflects the interests of public authorities, urban planners and managers. For example the Amsterdam LU categories "*Societal*" and "*Leisure*" are very detailed in terms of subcategories, since they have a major impact on the quality of life of the citizens.

The Foursquare application uses a predefined classification for the description of the category of a place (Foursquare, 2015a). As of January 2015, there were in total 712 Foursquare place categories. From these categories, the 633 are relevant to this study since they describe non-residential functions that are hosted within buildings. These 633 categories were manually aligned to their corresponding 8 Amsterdam LU categories by taking into consideration the detailed Amsterdam LU subcategories. The detailed matching between these classifications is available in the Appendix 1. The alignment of the Foursquare and Amsterdam classifications was based on personal interpretation. This alignment was made difficult due to semantic and spatial granularity differences between the two classifications. For example, the Foursquare place categories "*College Cafeteria*" and "*College Stadium*" were aligned to the Amsterdam LU category "*Societal*", which describes Colleges and Universities among others, and not to the "*Hotels, restaurants & cafes*" and the "*Leisure*" LU category, which describe cafes and sport stadiums respectively.

Table 2: Number and percentage of Amsterdam LU subcategories for each of the 8 main Amsterdam LU categories, and number and percentage of Foursquare place categories that have been aligned to them.

	Amsterdam LU category							
	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
Number of Amsterdam LU subcategories	7	7	24	13	36	51	9	2
Percentage of Amsterdam LU subcategories	5%	5%	16%	9%	24%	34%	6%	1%
Number of Foursquare place categories	10	20	138	261	80	112	10	2
Percentage of Foursquare place categories	2%	3%	22%	41%	13%	18%	2%	0%

Source: Spyratos et al., 2016

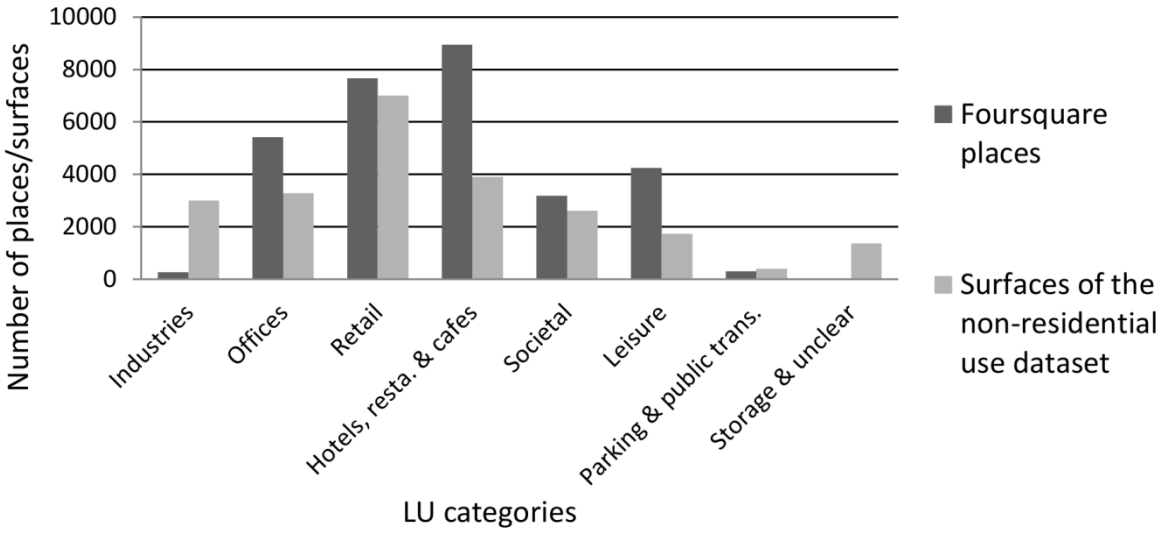
The second operation after aligning the Foursquare place classification to the Amsterdam one, was the cleaning of the Foursquare place dataset. As we mentioned in Section 3.2.1, in February 2015, there were 37,482 Foursquare places within the Amsterdam study area. From them, 30,036 were used for the estimation of the BBU, since they were referring to one of the 633 Foursquare place categories that were aligned to the 8 Amsterdam non-residential LU categories. As shown in Figure 11, places that belong to the "*Storage*" or "*Industries*" LU categories are underrepresented in the Foursquare dataset. On the contrary, places that belong to the rest categories such as the "*Leisure*" and the "*Hotels, restaurants & cafes"* are well represented. This fact shows that the Foursquare users are biased in favour of contributing content about commercial or recreational places.

The cleaned place dataset, represented by the set $\boldsymbol{p}$, has x= 30,036 members and is formally described in (5) (Spyratos et al., 2016).

$$\boldsymbol{p} = \{p_1, p_2, \dots., p_{x-1}, p_x\}, x = 30{,}036 \quad (5)$$

with p being a Foursquare place, and x the total number of Foursquare places within the Amsterdam study area.

Figure 11: Number of Foursquare places and number of surfaces of the non-residential use dataset for each Amsterdam LU category.



Source: Spyratos et al., 2016

Every place $p_j(f, a^p, c^p, d^p, b^p), j = 1 \dots x$, is a member of the $\boldsymbol{p}$ set and it has the attributes f, a^p, c^p, d^p, b^p (Spyratos et al., 2016):

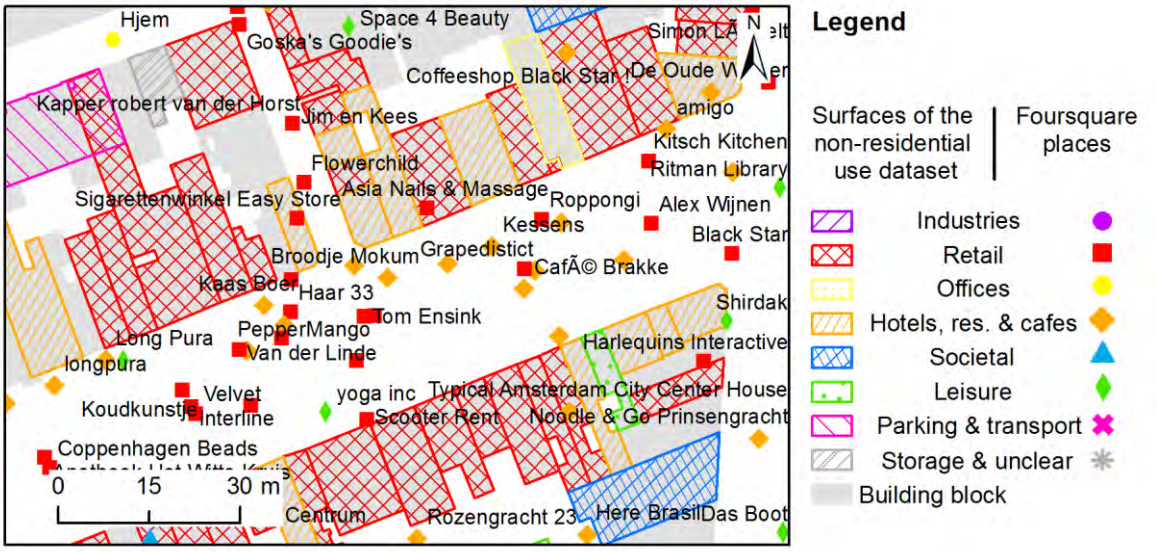
- The attribute f represents the Foursquare category of a place p_j. The value of this attribute represents one of the 633 Foursquare place categories that describe non-residential functions that are hosted within buildings.
- a^p is the Amsterdam non-residential LU category, which corresponds to the Foursquare place category f of a place p_j, as this was defined during the alignment of the two classifications (see Appendix 1). The a^p attribute is a member of the $\boldsymbol{a}$ set, defined in (3).
- c^p is the total number of Foursquare users that have declared a visit, or as is widely known, have checked in a place p_j.
- d^p is the shortest distance between a geometric point that represents the location of a place p_j, and the perimeter of the footprint of its nearest building block. In case a place is located within a building block, then $p_j(d^p) = 0$.
- b^p is the building block that either geometrically contains a place p_j or is the nearest building block to it. The b^p attribute is a member of the $\boldsymbol{b}$ set, defined in (1).

4.1.2. BUILDING BLOCK USE (BBU) ESTIMATION

In this section we present the method which was used for the generation of the estimated BBU dataset. In order to estimate the BBU, two parameters were taken into consideration, the parameters d and c, which are presented later in this Section. The estimation of the use of the built-up areas was performed at the spatial scale of urban building block level and not at the building level.

The reason for increasing the spatial scale of the LU estimation to building block level, is that in Amsterdam, and particularly in its historic centre, many buildings have narrow façade. This is because in the past their owners were taxed based on the building frontage (Farmer, 1993). The determination, of the position of places that are hosted in buildings with small frontage, by adding a marker on top of a web-map, requires high precision on behalf of Foursquare users and in addition web maps of high quality. Foursquare users are not aiming at describing the location of a place as accurately as possible but they simply require a level of accuracy that will allow them to efficiently share a geo-referenced piece of information (see Section 1.4.2). As a result many places are falsely described to be located in the nearby attached building. Due to that reason, an earlier attempt to allocate LU to buildings failed in terms of the accuracy of the estimations.

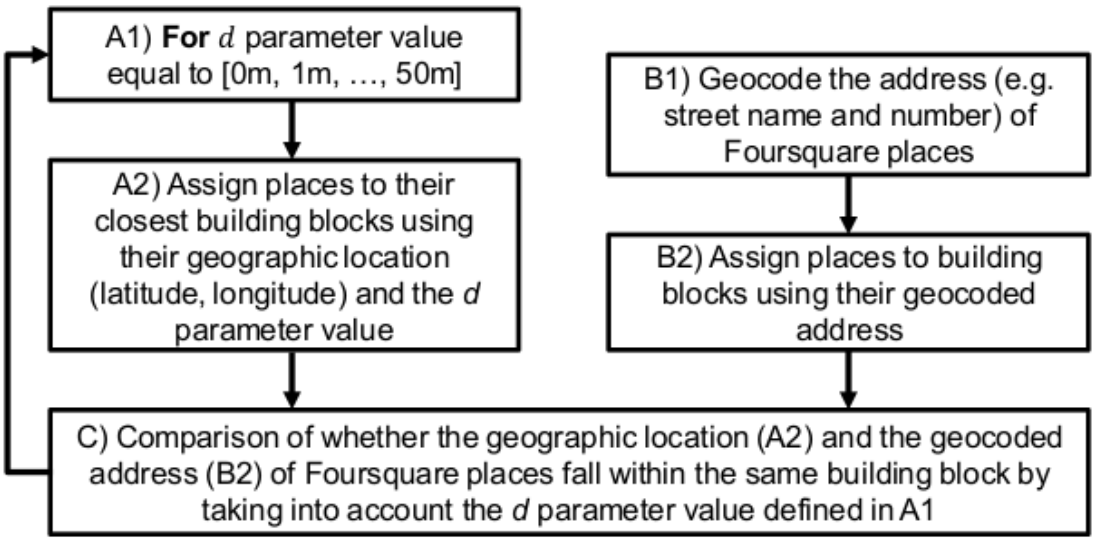
Figure 12: Foursquare places p_j and surfaces of the non-residential use dataset s_i located at the intersection of Rozengracht and Eerste Bloemdwarsstraat streets in Amsterdam. Both datasets are classified according to the Amsterdam LU categories.



Source: Spyratos et al., 2016

As shown in Figure 12 a considerable amount of Foursquare places that describe facilities hosted within buildings, are located outside the footprint of the building blocks. Nearly only one out of three places of the cleaned Foursquare place dataset is located within a building block. In order to use these places for the generation of the estimated BBU dataset, we need to allocate them into the building blocks they belong. For that purpose we have introduced the parameter d. This parameter describes the maximum value that the distance $p_j(d^p)$ from a place p_j to its closest building block b_k may take, in order to include p_j in b_k (Spyratos et al., 2016). The d parameter may take any integer value in the range [0…50]. When d parameter value is $d = 0m$, only places that are located within building blocks are taken into consideration for the generation of the estimated BBU dataset. The highest value, $d = 50m$, was determined empirically from the accuracy assessment of the estimated BBU dataset (see Section 5.1), with the rationale to examine all the possible d parameter values that are needed for the identification of the optimal estimated BBU classification.

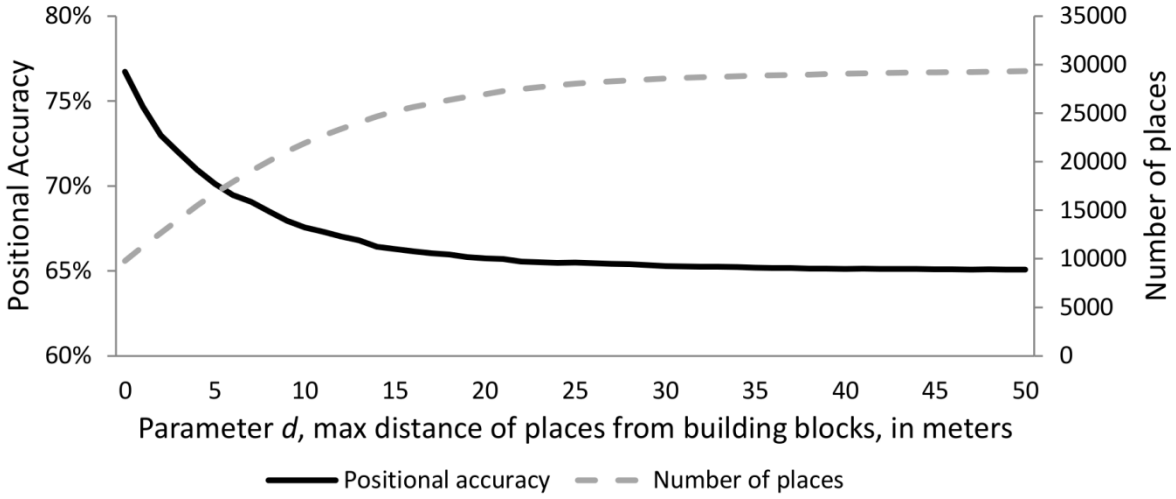
Figure 13: Method used for assessing the accuracy of the geometric allocation of Foursquare places to their closest building block based on the application of different d parameter values



Clearly, a place that is located within a building block has higher probabilities to be correctly allocated to the building block it belongs, than a place that is located many meters from its closest building block. For that reason, based on a method described in Figure 13, we measured how the accuracy of the geometric allocation of Foursquare places to their closest building block, varies based on the application of different d

parameter values. From the 30,036 Foursquare places used in the Amsterdam study area, 9,845 had complete address information in their description, meaning street name, street number and postcode. For obtaining the geographical coordinates that correspond to these addresses we used the OpenStreetMap's Nominatim geocoding service (OpenStreetMap, 2015). As true location of these places we used the geographical location that corresponds to their address. Finally, we tested for the 9.845 places with complete address information whether their address location derived from the geocoding service falls within the same building block as their geographic location which was defined by the Foursquare users. For assigning places to their closest building blocks based on their geographic location we used the d parameter. The results of this assessment are presented in Figure 14. As the d parameter value is increasing, the total positional accuracy of places decreases, while the number of places increases. Clearly, the selection of the d parameter value is a trade-off between the number of the places and their positional accuracy.

Figure 14: Percentage of Foursquare places for which both their address and their geographic location refer to the same building block (positional accuracy) vs number of Foursquare places, using the d parameter values in the range [0m … 50m]. For example, for the parameter value d =6m there are 21,000 places from which, based on the 9,845 place sample, 70% is positionally accurate.



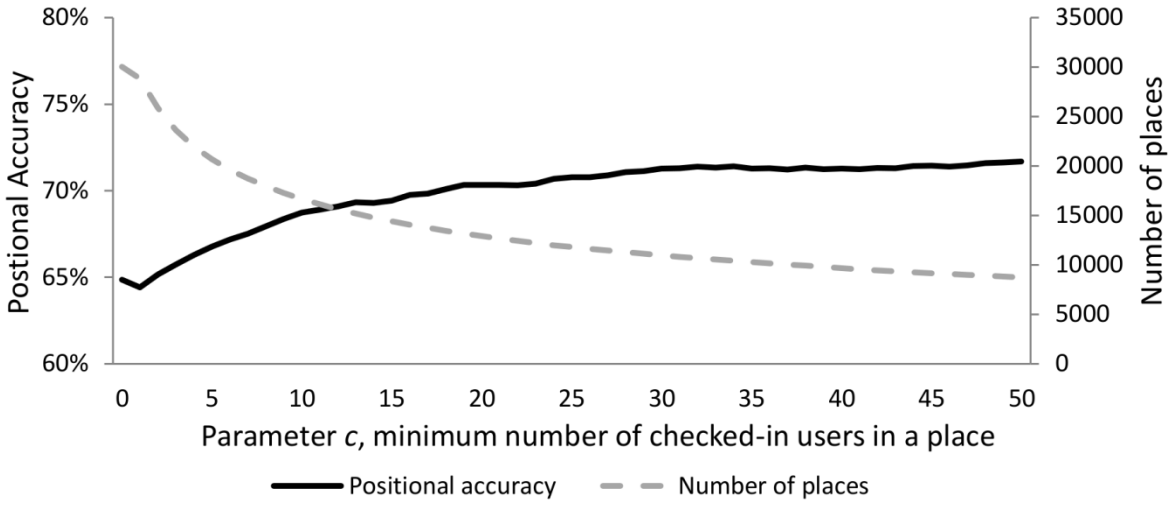
Source: Spyratos et al., 2016

To further investigate factors that affect the positional accuracy of Foursquare places we made use of the Linus' law (Raymond, 2001). Linus' law states that, the higher the number of users or contributors of a product is, the higher is the probability that an error

will be identified or fixed by someone. Haklay et al. (2010) demonstrated that Linus's law is valid for OpenStreetMap data, since it was found that there is a positive correlation between the number of contributors and the positional accuracy of the data. Linus' law applied on the Foursquare place dataset, would state that the higher the number of Foursquare users that have declared a visit in a place is, the higher is the probability that this place is accurately described (Spyratos et al., 2016).

To test whether Linus' law is applied on Foursquare place data we have introduced the parameter c. This parameter describes the minimum value that the number of Foursquare checked users in a place $p_j(c^p)$ may take, in order to include p_j in b_k.This parameter may take any integer value in the range [0…50], and its maximum parameter value was roughly determined as for the parameter d. To assess how the accuracy of Foursquare places varies based on the application of different c parameter values we have followed the same methodology as described above for the parameter d. First, we geocoded the address of 9,845 Foursquare places with complete address information. Second, we assessed whether the location of the geocoded address of these 9,845 places and their geographic location, expressed in latitude and longitude coordinates, refer to the same building block, by taking into account different values of the parameter c. As shown in Figure 15, while the c parameter value increases, the positional accuracy of Foursquare places also increases, with an exception when $c = 1$. For example, for the parameter value $c = 11$ there are about 16,000 places from which, based on the 9,845 place sample, the 69% is positionally accurate. This verifies that Linus' law applies to Foursquare place data. On the contrary, as the c parameter value increases, the number of Foursquare places that have equal or higher number of checked in users than the c parameter value, $p_j(c^p) \geq c$, is decreasing. As is also valid for the d parameter, the selection of the c parameter value is a trade-off between the number of the place data and their positional accuracy.

Figure 15: Percentage of Foursquare places for which both their address and their geographic location refer to the same building block (positional accuracy) vs number of places, using c parameter values in the range [0…50].

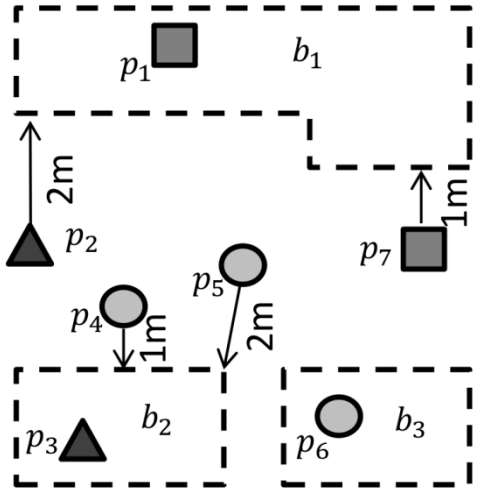


Source: Spyratos et al., 2016

Finally, taking into consideration the c and d parameters, we computed using formula (6 *)* the estimated BBU dataset $\boldsymbol{E}_{z,k,d,c}$. In detail, using formula (6) we calculated for each building block b_k whether places of each LU category a_z are hosted in it, taking into account all the possible combinations of the parameter values c and d (Spyratos et al., 2016). For instance, as shown in Figure 16, for $c^p \geq c = 0$ and within a distance $d^p \leq d = 2m$ from the building block b_2 there are 3 places: place p_3, which belongs to the a_5 LU category, and the places p_4 and p_5 that belong to the a_6 LU category. Therefore, in the building block b_2 the LU categories a_5 *"Societal"*, and a_6 *"Leisure"* are assigned.

$$\boldsymbol{E}_{z,k,d,c} \begin{cases} 1 \; if \; \exists \;\; p_j(a^p, b^p, d^p \leq d, c^p \geq c) \in \boldsymbol{p}, \\ where \; a^p = a_z, b^p = b_k, d = [0 \ldots 50] \; and \; c = [0 \ldots 50] \\ \\ 0 \; if \; \nexists \;\; p_j(a^p, b^p, d^p \leq d, c^p \geq c) \in \boldsymbol{p}, \\ where \;\; a^p = a_z, b^p = b_k, d = [0 \ldots 50] \; and \;\; c = [0 \ldots 50] \end{cases} \quad (6)$$

Figure 16: BBU estimation using Foursquare place data by taking into consideration the parameters c and d. In the building block b_2, using parameter values $c^p \geq c = 0$ and $d^p \leq d = 2m$, the a_5 and a_6 LU categories are assigned.



Parameter		Building Block	BBU of p_j		
d	c	b_k	a_3	a_5	a_6
d =0, $d^p \leq 0m$	c = 0, $c^p \geq 0$	b_1	p_1		
		b_2		p_3	
		b_3			p_6
d = 2, $d^p \leq 2m$	c = 0, $c^p \geq 0$	b_1	p_1, p_7	p_2	
		b_2		p_3	p_4, p_5
		b_3			p_6

Source: Spyratos et al., 2016

4.1.3. BBU ACCURACY ASSESSMENT

The accuracy assessment of the BBU estimations was performed for each LU category separately. This is because our aim is to evaluate the correctness of the assignment of each LU category on the building blocks. To this end, we compared the estimated BBU dataset, which was calculated using formula (6) in Section 4.1.2, with the reference BBU dataset which was calculated using formula (4) in Section 3.2.2.1. This comparison was performed 2,601 times in order to take into consideration any possible combination of the d and c parameter values. For the determination of the optimal BBU classification for each LU category, we used the Cohen's kappa coefficient (Cohen, 1960), which is presented in (7). In detail, the optimal BBU classification for each LU category is the one that was produced using the set of the parameter values d and c, for which the highest Cohen's kappa coefficient was achieved when compared to the reference BBU dataset.

The reason for using the Cohen's kappa coefficient is, first, that the LU categories have different sample size, second, it normalises for the expected chance of agreement (Carletta, 1996), and third it determines whether the classification results are significantly better than a random result (Congalton, 1991). When there is total agreement between the estimated and the reference BBU datasets, the kappa coefficient is one. On the contrary, when there is no agreement other than that which would be expected by chance, the kappa coefficient is zero. Negative kappa coefficient values can also occur and they indicate agreement less than that achieved by chance (Viera and Garrett, 2005).

$$k = \frac{p_0 - p_e}{1 - p_e} \qquad (7)$$

The p_o value in formula (7), known as the "observed" agreement, represents the proportion of times that the estimated and the reference BBU datasets agree for a given LU category. In detail, the p_o value is estimated using the formula (8) and the confusion matrix presented in Table 3. The p_e value in formula (7), known as the "expected" agreement, represents the proportion of times that the estimated and the reference BBU datasets are expected to agree by chance only (Spyratos et al., 2016). In detail, the p_e value is computed using formula (9).

$$p_o = \frac{TP + TN}{TP + FP + FN + TN} \qquad (8)$$

$$p_e = \left(\frac{TP + FP}{TP + FP + FN + TN} * \frac{TP + FN}{TP + FP + FN + TN}\right) + \left(\frac{FN + TN}{TP + FP + FN + TN} * \frac{FP + TN}{TP + FP + FN + TN}\right) \quad (9)$$

Table 3**:** Confusion matrix for the "*Retail*" LU category. TP represents the number of building blocks that have "*Retail*" LU in both the estimated and the reference BBU datasets. TN represents the number of building blocks that do not have "*Retail*" LU in both the estimated and the reference BBU datasets.

		Reference BBU dataset	
		Retail	Non-Retail
Estimated BBU dataset	Retail	True Positives (TP)	False Positives (FP)
	Non-Retail	False Negatives (FN)	True Negatives (TN)

Source: Spyratos et al., 2016

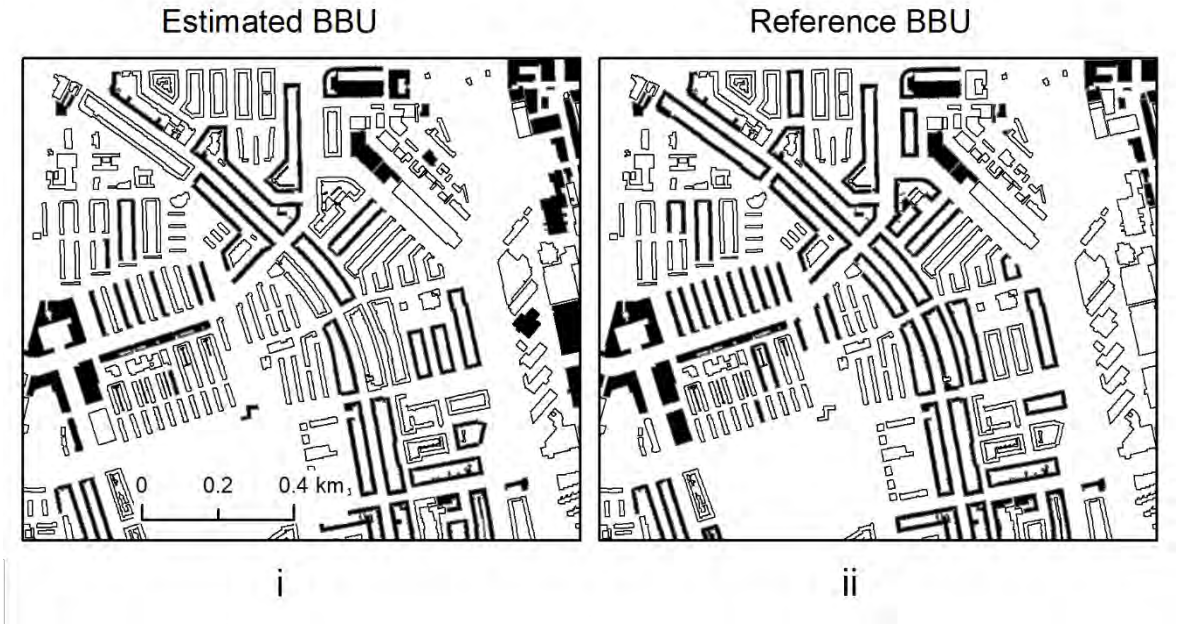
In addition to Cohen's kappa coefficient, we estimated the precision, the sensitivity and the specificity for each LU category, using formulas (10), (11) and (12) respectively. The precision refers to the probability that a building block classified, for example, to have the "*Retail*" LU in the estimated BBU dataset (see Figure 17-i), actually has that LU category in the reference BBU dataset (see Figure 17-ii). The sensitivity refers to the probability that a building block that has the "*Retail*" LU in the reference BBU dataset is correctly estimated as having the "*Retail*" LU. Finally, the specificity refers to the probability that a building block that does not have the "*Retail*" LU category in the reference BBU dataset, is correctly estimated as not having the "*Retail*" LU category. The accuracy assessment results are presented in the Section 5.1.

$$precision = \frac{TP}{TP + FP} \quad (10)$$

$$sensitivity = \frac{TP}{TP + FN} \quad (11)$$

$$specificity = \frac{TN}{FP + TN} \quad (12)$$

Figure 17: Building blocks estimated to have retail use (i) are shown in the left side of the figure and building blocks that have retail use in the reference dataset (ii) are shown in the right side of the figure. The building block are located in the "*Bos en Lommer*" neighbourhood, Amsterdam.



Source: Spyratos et al., 2016

Apart from estimating the existence or not of a LU category in a building block, we also estimated the number of facilities that belong to well predicted LU categories. To this end, a linear regression analysis was performed between the number of Foursquare places $p_j(b^p, a^p)$ for each LU category in each building block of the estimated BBU dataset and the number of surfaces $s_i(b^s, a^s)$ of each LU category in each building block of the reference BBU dataset. The results of this linear regression analysis are presented in Section 5.1.3. Finally, we repeated the proposed methodology in two subareas of the Amsterdam study. The reason for that repetition, is to assess how the accuracy of the estimations vary when the methodology is applied to subareas with different characteristics.

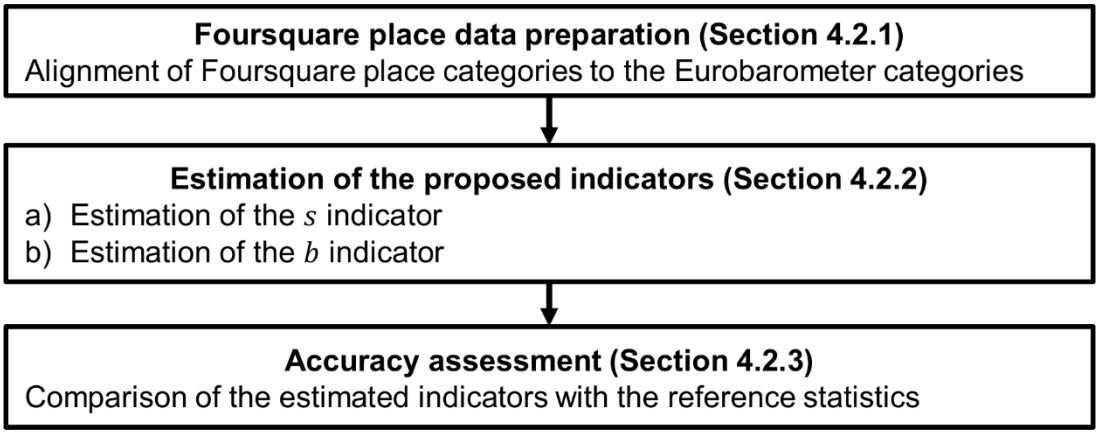
4.1.4. REPRODUCIBILITY OF THE PROPOSED METHODOLOGY

To test the reproducibility of the proposed methodology, we repeated it in the city of Varese, in Italy. To this end we collected Foursquare place data for the city of Varese and we processed it by using the method described in Section 4.1.1. Moreover, we produced the estimated BBU dataset of Varese, by using the same methodology as the one used for the Amsterdam study area (see Section 4.1.2). The accuracy assessment of the estimated BBU dataset of Varese, was performed on 150 randomly selected building blocks of the study area, for which reference data was available. The reference LU data for these 150 building blocks, shown in Figure 7, was collected through a ground survey as described in Section 3.2.3. The results of the accuracy assessment are presented in Section 5.1.5. Finally, in Section 5.1.6 we propose a cost efficient method for the selection of appropriate d and c parameter values. The reason is to facilitate the reproducibility of the proposed methodology by other researchers or interested parties.

4.2.CASE STUDY B- EVALUATING SERVICES AND FACILITIES

As show in Figure 18 the methodology of this case study consists of three parts. The first part, the data preparation phase, where Foursquare place categories are matched to their corresponding type of facilities and services as these are defined in the Eurobarometer. In the second part, we estimated, using Foursquare place data, the proposed indicators. In the third part, we compared these indicators with the reference statistics about citizens' satisfaction with regard to urban facilities and services.

Figure 18: Methodology followed for the preparation of Foursquare place data, for the estimation of the proposed indicators and for assessing their accuracy.



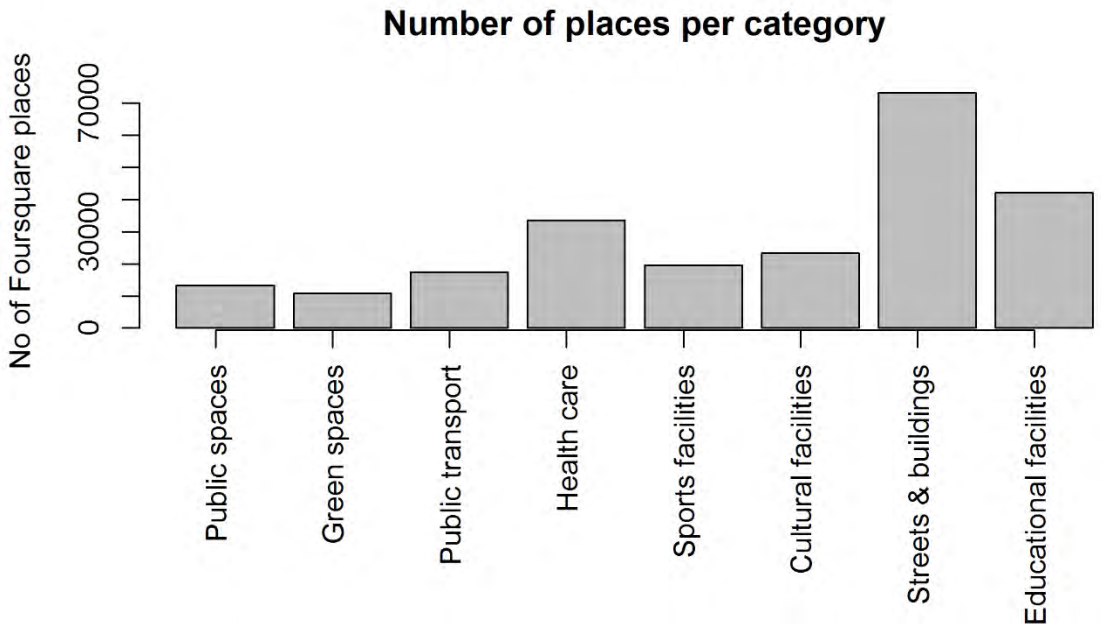
4.2.1. FOURQUARE PLACE DATA PREPARATION

The Foursquare places are classified based on the Foursquare place classification (Foursquare, 2015a). As of January 2015, the Foursquare place classification included 712 place type categories and subcategories. In the data preparation phase, we identified the Foursquare place categories, that refer to each Eurobarometer category. For instance, as shown in Table 4, 11 Foursquare place categories, which describe markets, squares and plazas among others, were matched to the *"Public spaces*" Eurobarometer category. Appendix 2 contains a link to a table that presents the detailed matching between the Eurobarometer categories with the Foursquare place categories. The Foursquare places, based on their Foursquare category, were assigned to Eurobarometer categories. In Figure 19, we present the total number of Foursquare places of the 17 cities, for each Eurobarometer category under investigation. In Appendix 3, we present the number of Foursquare places for each Eurobarometer category and each city. The category with the highest number of places is the "*Streets & buildings*".

Table 4: Alignment between the Eurobarometer categories and the Foursquare place categories

Eurobarometer Categories		Foursquare place Categories	
Name	Description	No	Name of Categories
Public spaces	Public spaces such as markets, squares and pedestrian areas	11	*Pedestrian Plaza; Playground; Road; Street; Plaza; Flea Market; Market; Christmas Market; Farmers Market; Fish Market; Night Market*
Green spaces	Green spaces such as parks and gardens	7	*Park; National Park; Sculpture Garden; Forest; Garden; Botanical Garden; Nature Preserve*
Public transport	Public transport, for example the bus, tram or metro	10	*Bus Line; Platform; Train; Bus Station; Bus Stop; Cable Car; Light Rail; Subway; Train Station; Tram*
Health care	Health care services, doctors and hospitals	10	*Medical Center; Dentist's Office; Eye Doctor; Hospital; Mental Health Office…*
Sports facilities	Sports facilities such as sport fields and indoor sport halls	49	*Mini Golf; Gym / Fitness Center; Hockey Arena; Tennis; Baseball Field; Basketball Court; Golf Course; Rugby Pitch; Soccer Field; Tennis Court; Swim School….*
Cultural facilities	Cultural facilities such as concert halls, theatres, museums and libraries	21	*Art Gallery; Comedy Club; Concert Hall; Movie Theater; Museum; Music Venue; Performing Arts Venue; Library; Multiplex; Art Museum; Theater…*
Streets & buildings	The state of the streets and buildings in your neighborhood	5	*Road; Street; Neighborhood; Building; Residential Building (Apartment / Condo)*
Educational facilities	Schools and other educational facilities	31	*College & University; Medical School; Elementary School; High School; Language School; Middle School;…*

Figure 19: Number of Foursquare places of the 17 cities, for each Eurobarometer category.



4.2.2. ESTIMATION OF THE PROPOSED INDICATORS

Our research hypothesis is that the higher the number of places is that belong to a facility or a service type on social media, then higher the citizen satisfaction rate with regard to this facility or service type will be. To test this hypothesis, we have introduced two indicators. The first is the s indicator which is estimated using equation (13), as follows:

$$s_{u,c,w} = \frac{n_{u,c}}{p_u^{1/w}} \quad (13)$$

where w is the proposed weight, p_u is the population of the urban area u, and the $n_{u,c}$ the number of places in an urban area u that belong to the Eurobarometer category c.

In practice, the s indicator measures the number of places that belong to a Eurobarometer category per citizen in a city. In line with our hypothesis, the higher the value of the s indicator is for a Eurobarometer category, then the higher the percentage of satisfied citizens will be with regard to this category. The number of facilities in a city is not increasing proportionally to its population. For instance, in a small town of 1,000 population, there might be one football field, while in a metropolitan area of 1,000,000 population there might be 1 football field per 5,000 citizens. The average football field of a metropolitan area will have a higher capacity and it will be used more intensively than a football court in a small town. To take into consideration the fact the number of facilities does not grow linearly with the population of an urban area, we have introduced a weight, denoted w, in the estimation of both indicators. This weight is used so as to logarithmically increase the denominators of the equations (13) and (14). It may take any value in the range [0.5 - 3.4] and its minimum and maximum value was determined empirically with the rational to include any possible value that optimizes the accuracy of the indicators.

As one would expect, not all of the facilities of a city are included in the Foursquare dataset. The completeness of the Foursquare place dataset varies across different place categories and across different areas. For example, as shown in Figure 11, recreational and commercial places are better represented in the Foursquare place dataset compared to places that belong to other categories. Thus, a low s indicator for a Eurobarometer category of an urban area might indicate either low number of facilities that belong to this

Eurobarometer category in reality, or low representativeness of those facilities in the Foursquare place dataset. Low representativeness is due to the fact that the Foursquare users are not contributing information about these facilities to Foursquare. This might be due to the combination of the following reasons:

a) Low or no usage of those facilities by Foursquare users. Low or no usage of particular facilities might be due to various reasons, such as, the high cost for accessing them, or the negative Foursquare users' attitude towards participation in activities that are hosted in these facilities.
b) Low attractiveness of that facilities. Foursquare users are using these facilities but they are not declaring their presence on Foursquare. The more attractive a place is, the more likely is to be added to Foursquare. This is because Foursquare users are motivated to add places and share their presence in interesting places that enhances their self-presentation or to places that they endorse (Cramer et al., 2011).
c) Low number of Foursquare users in a city. Thus, differences in the s indicator between cities might only be due to the low numerator of equation (13), since the denominator is independent to the level of Foursquare usage in a city

To overcome the limitations that the low use of Foursquare in a city might introduce in the estimation of the indicator s, we have additionally estimated the indicator b. The indicator b, which is described in equation (14), divides the total number of places of each Eurobarometer category by the total number of places that describe retail facilities. Since retail places are an integral part of Foursquare datasets, any bias introduced by the low usage of Foursquare in a city will be removed or drastically reduced. Additionally, the numerator and denominator of the equation (14) are independent variables, since the places that belong to the retail categories are not included in any of the 8 categories for which we estimated the indicators in this study. In total, 133 Foursquare categories describe retail facilities and the places of this category counts for the 19.82% of the total Foursquare places in the 17 urban cities under investigation.

$$b_{u,c,w} = \frac{n_{u,c}}{n_{u,retail}^{1/w}} \qquad (14)$$

Where $n_{u.retail}$ is the number of retail places in an urban area u, $n_{u,c}$ the number of places in an urban area u that belong to the category c, and w is the proposed weight.

In this case study we did not take the number of check-ins and likes into consideration in the estimation of the s and b indicators. The reason for this is that earlier attempts to take these attributes into consideration for the estimation of the indicators failed due to the low accuracy results when compared to the reference statistics. The number of check-ins and likes are highly affected by marketing strategies, for instance, paid advertisement, and importantly by the popularity of "Landmark places" which attract a high number of tourists. For example, the "Camp Nou" stadium in Barcelona had 3,340 likes and 69,293 check-ins as of February 2015, which account for the 46.5% and the 18.7% of all the likes and all the check-ins in sport facilities in Barcelona respectively. Obviously the popularity of "Camp Nou" does not have such a major effect on the day to day perception of Barcelona's residents with regard to sport facilities. The same applies to the famous public space and road junction "Piccadilly Circus" in London. As of March 2015, the "Piccadilly Circus" had 81,017 check-ins and 1,279 likes which count for the 12.3% and the 10.5% of all check-ins and likes in "public spaces" in London respectively. To overcome these limitations we have tested various methods for reducing the impact that very popular and touristic places might have in the estimation of the overall citizens' satisfaction with regard to urban facilities and services without any considerable result.

4.2.3. ACCURACY ASSESSMENT

To test the validity of our hypothesis, we compared the relative number of places of a city that belong to a Eurobarometer category, with the percentage of satisfied citizens for the same category in the same city. In practice, the higher the relative number of places is that belong to a Eurobarometer category, meaning the values of the s and the b indicators, the higher the percentage of satisfied citizens with regard to this category will be. More specifically, we used the least squares linear regression model to study any association between the s and b indicators and the percentage of "Very satisfied" and "Totally satisfied" citizens about particular urban facilities or services. The coefficient of determination, denoted R^2 and defined in (15), was used for the evaluation of the correspondence between these variables. Taylor (1990) defined the coefficient of determination as *"the percent of the variation in the values of the dependent variable (y) that can be "explained" by variations in the value of the independent variable (x)*". For example an $R^2 = 0.65$ would mean that the 65% of the total variation in the indicators s or b can be "explained" by the variation in the percentage of "Very satisfied" or "Totally satisfied" citizens about particular urban facility or service type. The coefficient of determination ranges from zero to one.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad (15)$$

Where SS_{tot} is the "total sum of squares" which is the sum of the squared deviations of all the observations from their mean (Everitt and Skrondal, 2010), and SS_{res} is the "sum of squares of residuals", which is the sum of squared differences between the actual values of the s and b indicators and the values predicted by the linear regression model. To protect against being misled by the random variation in the estimates we performed a test of significance and we tested the null hypothesis. We decided to use a very low significance criterion, $a = 0.01$, so as to reduce the possibility of falsely rejecting the null hypothesis (Cohen et al., 2013: 15). The null hypothesis is that the true value of the linear regression coefficient, i.e. the slope, is zero (Rawlings et al., 1998: 16). A zero slope means that the value of the indicators s and b show neither positive nor negative change as the percentage of satisfied citizens increases or decreases.

5. RESULTS

In this Section, we present the results of this research. First, in Section 5.1, we present the results of the case study A. In Section 5.2, we present the results of the case study B.

5.1. CASE STUDY A - ESTIMATING BUILDING BLOCK USE

In this Section we present the results of the case study A – *"Using Foursquare place data for estimating building block use*". In Section 5.1.1 we present the best accuracy results of the estimated BBU dataset. In Section 5.1.2 we present an assessment about the impact of the d and c parameter values on the accuracy of the estimated BBU dataset. In addition, in Section 5.1.3 we assess how the accuracy of the proposed methodology varies through space, by repeating it in two subareas of the Amsterdam study area. In Section 5.1.4 we analyse the density of places that belong to the two LU categories, for which we had robust estimations, the "*Retail*" and the "*Hotels, restaurants & cafes*". We present the results of the repetition of the proposed methodology in the city of Varese in Section 5.1.5, and finally, we describe a method for selecting appropriate d and c parameter values in Section 5.1.6.

5.1.1. BEST ACCURACY RESULTS

The comparison of the estimated and the reference BBU datasets was performed for each LU category 2,601 times in order to take into consideration any possible combination of the c and d parameter values. The best accuracy results (see Table 5) refer to the optimal BBU classification for each LU category that was produced using the set of the parameter values d and c, for which the highest kappa coefficient was achieved when compared to the reference BBU dataset (Spyratos et al., 2016). These optimal sets of parameter values c and d for each LU category are presented in the last two columns of Table 5.

The highest Cohen's kappa coefficient values, 0.76 and 0.65 were estimated for the "*Hotels, restaurants & cafes*" and the "*Retail*" LU categories respectively. According to a classification proposed by Landis and Koch (1977), these kappa coefficient values, since are higher than 0.61 and lower than 0.80, are indicating substantial agreement between the reference and the estimated BBU datasets. For the "*Offices*", the "*Societal*" and the "*Leisure"* LU categories the kappa coefficient values are indicating moderate

agreement since they range from 0.42 to 0.52 (Landis and Koch, 1977). Finally, the lowest kappa coefficient values, below 0.2, were achieved for the LU categories "*Industries*", *"Parking & public transport*" and "*Storage & unclear*".

The precision for the "*Retail*" and the "*Hotels, restaurants & cafes*" LU categories is 72% and 82% respectively. These percentages, are indicating that the majority of the building blocks which were estimated using Foursquare place data as having one of the above LU categories, actually have these LU categories in the reference BBU dataset. The sensitivity, for the "*Retail*" and the "*Hotels, restaurants & cafes*" LU categories is lower compared to their precision. The 75% of the building blocks with the "*Hotels, restaurants & cafes*" LU category and the 67% of the building blocks with the "*Retail*" LU category were correctly identified using Foursquare place data. The specificity for all the LU categories is high, above 90%. This is because in the study area there are many building blocks without any non-residential LU that were correctly estimated as not having such LU.

Table 5**:** Best accuracy results for each LU category of the Amsterdam study area. In the last two columns we present the d and c parameter values for which the highest Cohen's Kappa coefficient values were achieved.

	Kappa	Precision	Sensitivity	Specificity	Parameters	
					d	c
Industries	0.06	45.70%	5.72%	98.51%	43	0
Offices	0.50	56.23%	61.25%	90.64%	18	0
Retail	0.65	72.48%	67.14%	95.48%	29	5
Hotels, restaurants & cafes	0.76	82.41%	75.12%	97.63%	22	26
Societal	0.46	58.01%	52.29%	92.18%	21	1
Leisure	0.42	48.05%	49.58%	92.61%	20	3
Parking & public transport	0.17	30.38%	14.63%	98.84%	17	0
Storage & unclear	0.01	33.33%	0.46%	99.91%	30	0

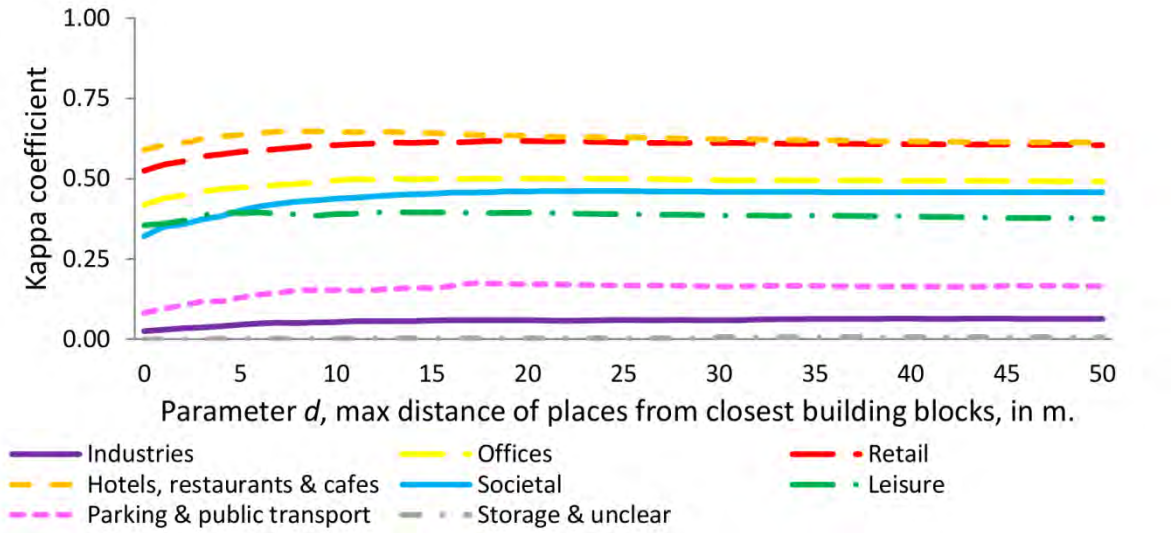
Source: Spyratos et al., 2016

5.1.2. IMPACT OF d AND c PARAMETER VALUES

The d and c parameter values have a major impact on the accuracy of the estimated BBU dataset. As shown in Figure 20, the kappa coefficient for each LU category is increasing as the d parameter value increases up to its optimal value. These optimal d parameter values range from $d = 17m$ for the "*Offices*" LU category to $d = 43m$ for the "*Industries*" LU category (see Table 5). This difference is due to the fact that most of the facilities of the former LU category are located in densely built-up areas while most of facilities of the latter are located in sparsely built-up areas. As shown in Figure 20, for d parameter values higher than $d > 20m$, the Cohen's Kappa coefficient changes insignificantly. This is because only the 10% of Foursquare places are located more that 20m from their closest building blocks (see Figure 14).

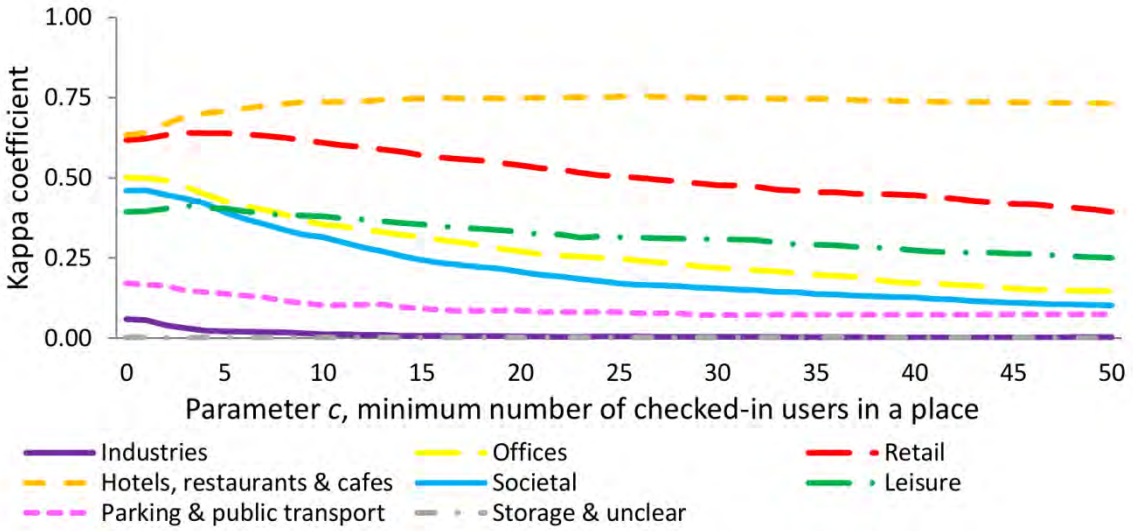
As regard the parameter c, for some LU categories the kappa coefficient value is increasing as the number of minimum checked in users increases up to the optimal c value (see Figure 21). This optimal value ranges among the LU categories. For example, for the category "*Hotels, restaurants & cafes*", the optimal c parameter value is 26 while for the LU category "*Industries*" is 0. As we mentioned in Section 4.1.2, the selection of the c parameter value is a trade-off between the number of the place data and their positional accuracy. The absence or the excess of place data for each individual LU category (see Figure 11) has an impact on the determination of its optimal c parameter value.

Figure 20: Cohen's Kappa coefficient of the estimated BBU dataset calculated using the parameter values $d = [0m \dots 50m]$ and $c = 0$.



Source: Spyratos et al., 2016

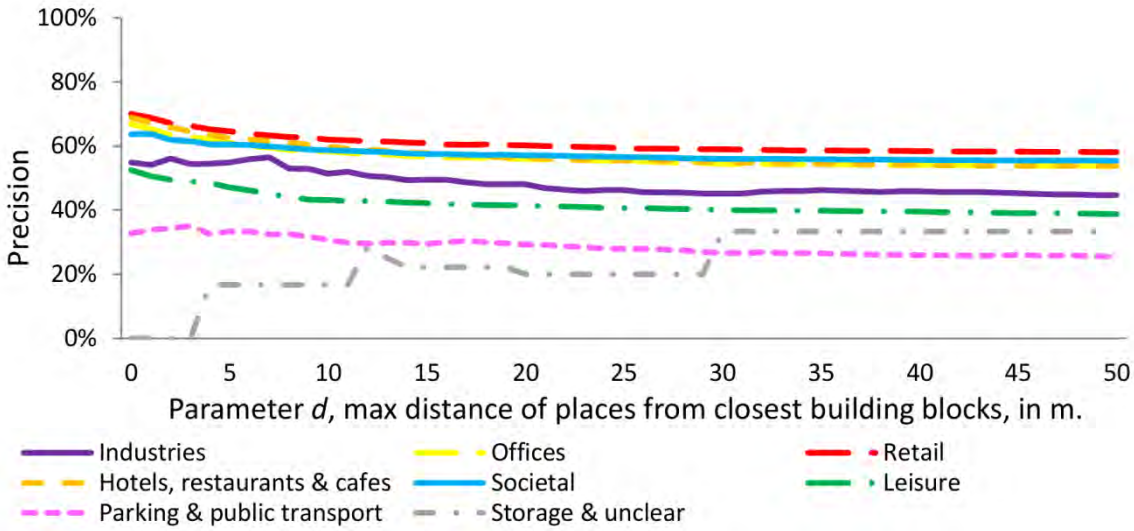
Figure 21: Cohen's Kappa coefficient of the estimated BBU dataset calculated using the parameter values $c = [0 \dots 50]$ and $d = 20m$.



Source: Spyratos et al., 2016

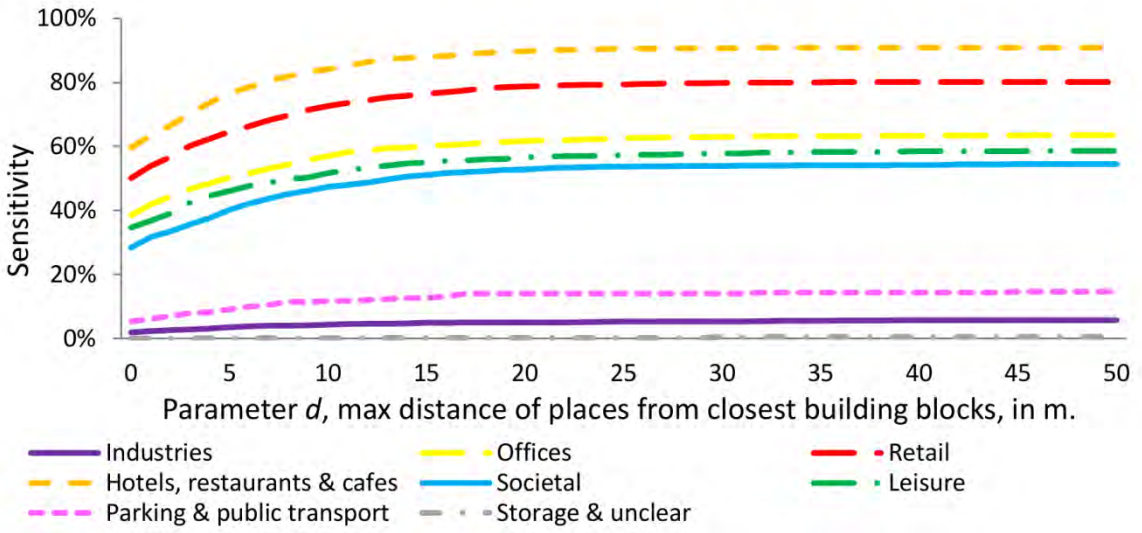
The precision and the sensitivity of the 8 LU categories is affected by the value of the parameter d. As shown in Figure 22, when the d parameter value is increasing, the precision of the estimations for the 8 non-residential LU categories is slightly decreasing. This is because, as shown in Figure 14, by taking into account for the generation of the estimated BBU dataset, Foursquare places that are located far from building blocks, the probability that these places are positionally accurate is decreasing. Inversely, as shown in Figure 23, as the d parameter value is increasing up to its optimal value, the sensitivity for each of the 8 non-residential LU categories is increasing as well. This is because, as the maximum distance between places and their closest building blocks is increasing, more Foursquare places are taken into account for the generation of the estimated BBU dataset. As a result, the probability that a building block with one or more non-residential LU in the reference dataset is being identified as having non-residential LU in the estimated dataset is increasing.

Figure 22: Precision per LU category of the estimated BBU dataset calculated using the parameter values $d = [0m \dots 50m]$ and $c = 0$.



Source: Spyratos et al., 2016

Figure 23: Sensitivity per LU category of the estimated BBU dataset calculated using the parameter values $d = [0m \dots 50m]$ and $c = 0$.



Source: Spyratos et al., 2016

As shown in Figure 24 and Figure 25 the c parameter value affects the precision and the sensitivity of the individual LU categories of the estimated BBU dataset. Due to the application of the Linus law (see Section 4.1.2), as the c parameter value is increasing, places of higher quality are used for the generation of the estimated BBU dataset. Thus

the estimations are more precise. Nevertheless, as the c parameter value is increasing less places are used for the generation of the BBU dataset (see Figure 15) and thus less building blocks with one or more non-residential LU are identified as having such LU. This fact decreases the sensitivity of the estimations.

Figure 24: Precision per LU category of the estimated BBU dataset calculated using the parameter values $c = [0 \dots 50]$ and $d = 20m$.

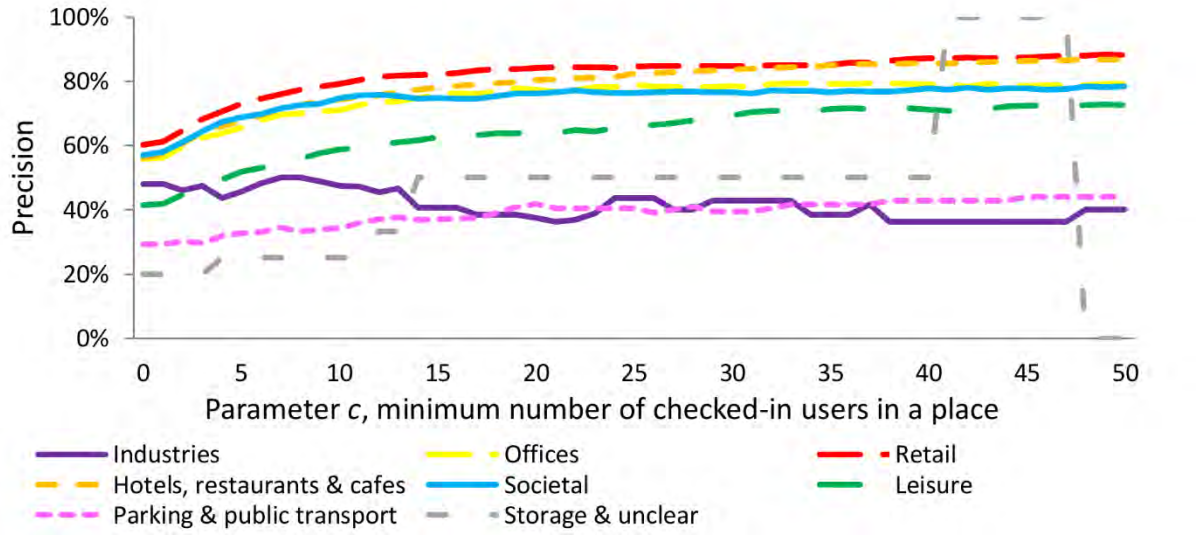
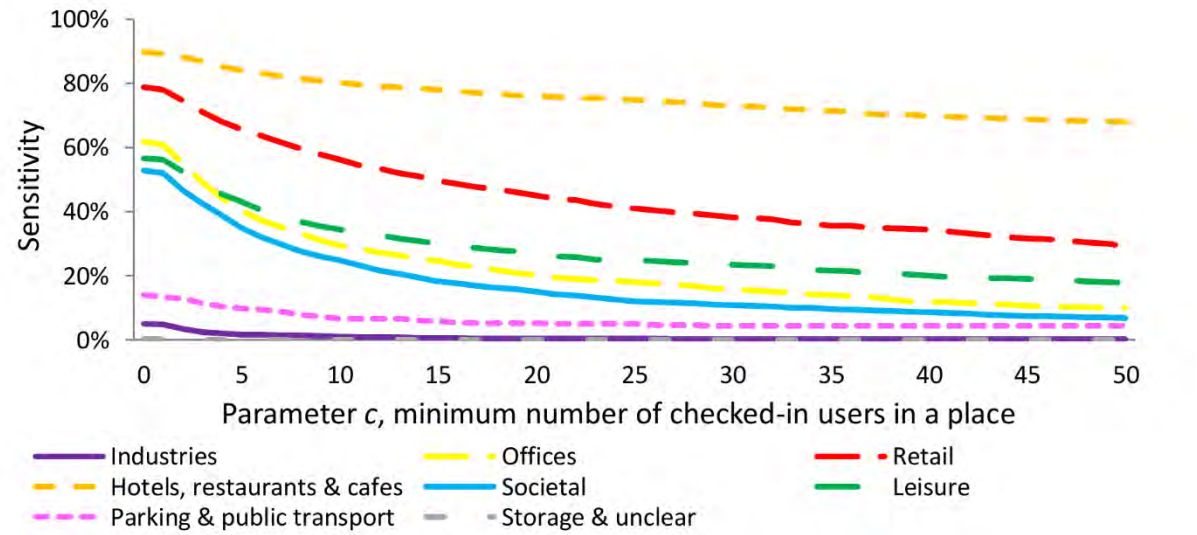


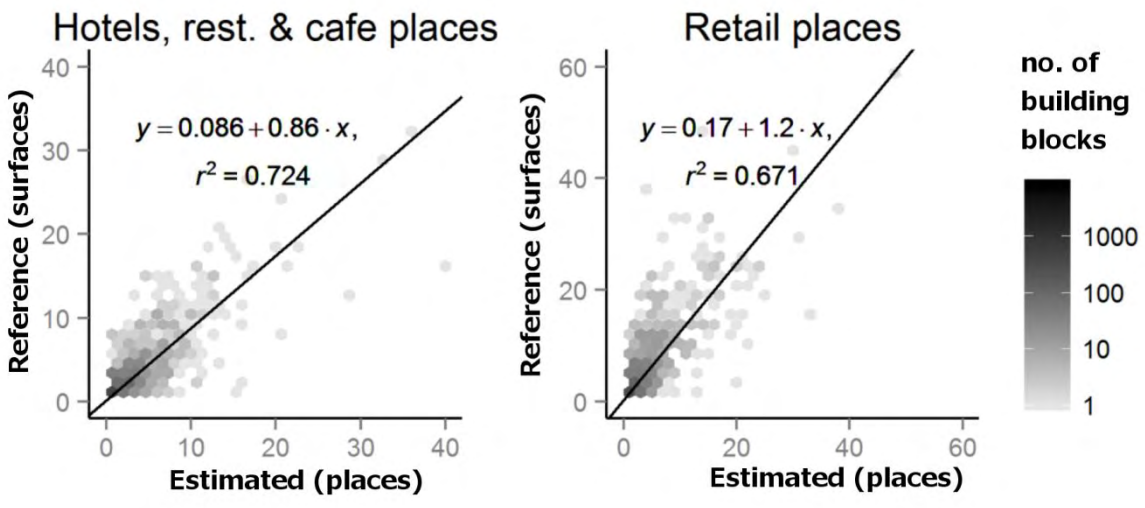
Figure 25: Sensitivity per LU category of the estimated BBU dataset calculated using the parameter values $c = [0 \dots 50]$ and $d = 20m$.



5.1.3. DENSITY OF "RETAIL" AND "HOTELS, RESTAURANTS AND CAFES" LU

In this Section we assess whether, apart from estimating the existence or not of a LU category in a building block, we are also able to estimate the number of facilities that belong to a particular LU category in a building block, meaning the density of that LU. This assessment was performed for the two LU categories, for which we had robust estimations, the "*Retail*" and the "*Hotels, restaurants & cafes*". To this end, a linear regression analysis was performed in order to assess whether the number of Foursquare places $p_j(b^p, a^p)$ for each LU category in each building block of the best estimated BBU dataset, is correlated to the number of surfaces $s_i(b^s, a^s)$ of each LU category of the reference BBU dataset. The coefficient of determination, denoted by R^2, and the slope of the fitted line are used for the evaluation of the correspondence between the two variables. As shown in Figure 26, the coefficient of determination is $R^2 = 0.72$ (p value < 0.01) for the "*Hotels, restaurants & cafes*" LU category and $R^2 = 0.67$ (p value < 0.01) for the "*Retail*" LU category, while the slope of the fitted line is 0.86 and 1.2 respectively.

Figure 26: Linear regression plots between the number of places of the best estimated BBU dataset and the number of surfaces in the reference BBU dataset of each building block for the LU categories "*Hotels, restaurants & cafes*" and "*Retail*". The number of building blocks for each combination of estimated places and reference surfaces is presented in logarithmic scale.

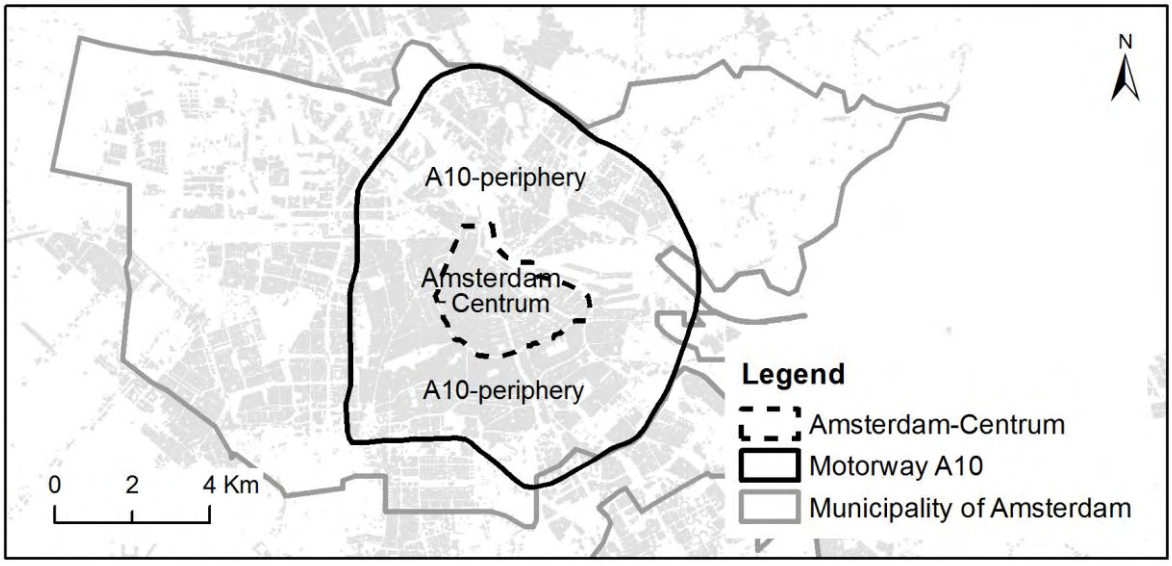


Source: Spyratos et al., 2016

5.1.4. SPATIAL VARIATION OF THE ACCURACY RESULTS

In this Section, we present the results of the repetition of the proposed methodology in two subareas of the Amsterdam study area. The reason for that repetition, is to assess how the accuracy of the estimations vary when the methodology is applied to subareas with different characteristics. The selected subareas are the Amsterdam-Centrum, and the A10-periphery. The A10-periphery is the area that remains if we exclude the Amsterdam-Centrum from the area that is enclosed by the A10 motorway (see Figure 27). Amsterdam-Centrum is the urban centre of Amsterdam, and in contrast to the A-10 periphery, is the area where most of the Foursquare places and most of the activity of Foursquare users are located (see Figure 4).

Figure 27: The two subareas of the Amsterdam study area: the Amsterdam-Centrum and the A10-periphery, which is the area that remains if we exclude the Amsterdam-Centrum from the area that is enclosed by the A10 motorway.



As shown in Table 6, the kappa coefficient, the precision and the sensitivity of the estimated BBU dataset in the Amsterdam-Centrum is higher than in the A10-periphery. The reason for that is that in Amsterdam-Centrum there is an excess of Foursquare places, while in A10-periphery there is an absence. On the contrary, the specificity in the A10-periphery is higher that it is in the Amsterdam-Centrum. This is because in the A10-periphery there is a high percentage of building blocks without any non-residential LU which were correctly estimated as such.in in The optimal *c* parameter values of the well-

estimated LU categories vary insignificantly between two subareas. Contrariwise, the optimal values of the d parameter vary significantly

Table 6: Best accuracy results for the two subareas of the Amsterdam study area.

	Amsterdam-Centrum						**A10 Periphery**					
	Kappa	Precision	Sensitivity	Specificity	d	c	Kappa	Precision	Sensitivity	Specificity	d	c
Industries	0.08	56%	8%	98%	35	0	0.06	43%	5%	99%	43	0
Offices	0.55	71%	80%	75%	35	0	0.45	50%	55%	92%	18	0
Retail	0.67	80%	81%	86%	47	5	0.60	67%	63%	96%	27	4
Hotels, restaurants & cafes	0.81	89%	89%	92%	17	26	0.70	80%	67%	98%	25	28
Societal	0.46	65%	56%	88%	21	3	0.46	59%	50%	94%	23	1
Leisure	0.43	65%	62%	81%	9	3	0.35	41%	41%	94%	20	3
Parking & public transport	0.15	40%	15%	96%	17	0	0.16	31%	12%	99%	17	4
Storage & unclear	0.00	NA	0%	100%	0	2	0.01	50%	1%	100%	30	2

Source: Spyratos et al., 2016

5.1.5. ASSESSMENT OF THE METHODOLOGY IN THE CITY OF VARESE, ITALY

To test the reproducibility of the proposed methodology, we repeated it in the city of Varese, in Italy. The accuracy assessment of the estimated BBU dataset of Varese was performed on 150 building blocks, for which LU data was collected through a ground survey (see Section 3.2.3). The results of the accuracy assessment are presented in Table 7 and are similar to a certain extent with the results of the Amsterdam study area (see Table 5). Compared to Amsterdam, the kappa coefficient values for all the LU categories are lower in Varese. In detail, for the "*Hotels, restaurants & cafes"* LU category the kappa coefficient values differs slightly and is 0.76 for Amsterdam and 0.73 for Varese. For the "*Retail*" LU category, the kappa coefficient values differ considerably and are 0.65 for Amsterdam and only 0.48 for Varese. The kappa coefficient values are highly affected by the low sensitivity. The low sensitivity reveals that there is an absence of Foursquare place data in Varese. As a results, the optimal c parameter values are low so as to include more Foursquare places in the estimation of the BBU dataset. Regarding, the optimal value of the parameter d, as for the Amsterdam study area, it varies significantly across the LU categories.

Table 7: Accuracy assessment results of the Varese study area.

	Kappa	Precision	Sensitivity	Specificity	Parameters	
					d	c
Industries	0	NA	0	1	0	0
Offices	0.39	75.00%	33.33%	97.56%	2	0
Retail	0.48	70.00%	52.50%	91.82%	42	0
Hotels, restaurants & cafes	0.73	85.00%	70.83%	97.62%	9	5
Societal	0.33	61.54%	30.77%	95.97%	10	0
Leisure	0.39	66.67%	30.77%	98.54%	5	0
Parking & public transport	0.00	NA	0.00%	100.00%	0	5
Storage & unclear	0.00	NA	0.00%	100.00%	0	0

Source: Spyratos et al., 2016

5.1.6. SELECTION OF APPROPRIATE PARAMETER VALUES

As revealed from the accuracy assessment (see Section 5.1.2), the selection of the d and c parameter values is vital for optimal BBU estimations. The selection of these values depends on the purpose of the application, and as shown in Figure 14 and Figure 15 it is a trade-off between place data quality and place data quantity. In this study, the Cohen's Kappa coefficient was used for the determination of the optimal parameter values, since it determines whether the classification results are significantly better than a random result (Congalton, 1991). For instance, applications that require high precision will have to select higher c parameter values and lower d parameter values compared to applications that require high sensitivity.

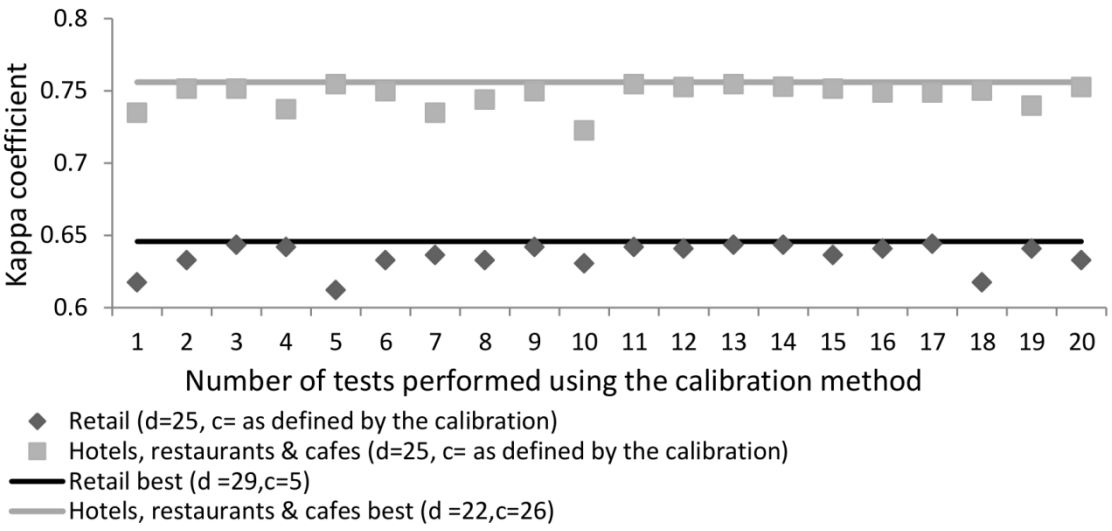
As revealed from the accuracy assessment of the Amsterdam and Varese study areas, the optimal sets of d and c parameter values are not consistent across different cities. This is because these values are highly affected by two area-dependent factors. First, by the characteristics of the study area such as its urban morphology, the size of building blocks, and the distance between building blocks. Second, by the characteristics of the Foursquare dataset in the study area such as the number of Foursquare users and the degree on which the Foursquare application is used by its users.

As regards the d parameter, the accuracy assessment revealed that values between $0m \leq d \leq 25\text{m}$, have a major impact on the accuracy of the estimations and especially in the sensitivity. Roughly estimated, for d parameter value equal to, $d = 25\text{m}$, the kappa coefficient reach on average for all the LU categories its highest value. For d parameter values higher than 25m, $d > 25\text{m}$, the positional accuracy of Foursquare places and the kappa coefficient value, the precision, and the sensitivity of the estimated BBU dataset remain nearly stable (see Figure 14, Figure 20, Figure 22 and Figure 23). As a result, d parameter values higher than, $d > 25\text{m}$, do not have any considerable effect on the estimations. Thus we suggest, as a general guidance for those willing to use Foursquare place data for LU mapping purposes, to use a fixed d parameter value of $d = 25\text{m}$ for all the LU categories (Spyratos et al., 2016).

As regards the c parameter, the accuracy assessment revealed that its optimal values differ considerable for each LU category. As a result, the c parameter values needs to be chosen separately for each LU category. For the selection of these values, a calibration is suggested. This calibration, is the application of the methodology in a sample of 5% of

randomly selected building blocks of the study area, for which there is ground truth data available. The optimal c parameter values that have been estimated during the accuracy assessment of the sample, are then selected to be applied to the whole study area. For testing the validity of the calibration method, we have generated 20 estimated BBU datasets of the Amsterdam study area using a fixed d parameter value d =25, and c parameter values as specified by performing 20 different calibrations. As shown in Figure 28, the majority of these 20 estimated BBU datasets had, when compared to the reference BBU dataset, kappa coefficient values slightly lower than the BBU dataset that was produced using the optimal set of the d and c parameter values. As a result, we conclude that the proposed calibration is a cost efficient method for estimating the c parameter values for each LU category.

Figure 28: The Kappa coefficients of the BBU datasets that were estimated using parameter values d =25 and c as defined by the calibrations on 20 randomly selected samples of the Amsterdam study area.



Source: Spyratos et al., 2016

5.2.CASE STUDY B - EVALUATING SERVICES AND FACILITIES

In this Section we present the results of the case study B – *"Evaluating the services and facilities of European cities using crowd-sourced place data*". In Section 5.2.1 we present the correspondence between the proposed indicators and the Eurobarometer statistics, and in Section 5.2.2 we present the impact of the weight on the accuracy of the estimations.

5.2.1. ACCURACY OF INDICATORS

In Table 8 we present the results of the comparison between the best s and b indicators and the percentage of "Very satisfied" and "Totally satisfied" citizens. The best indicators are those that were calculated using the optimal weight values (w), for which the highest R^2 values were achieved when compared to the percentage of "Totally satisfied" or "Very satisfied" citizens. The p-values for the 5 Eurobarometer categories are less than the significance criterion, $a = 0.01$, and thus we reject the null hypothesis. Therefore, we conclude that the two indicators and the percentages of satisfied citizens for these five categories are not unrelated.

As shown in Table 8, there exist strong linear relationships ($R^2 > 0.6$) between the s indicators of the categories *"Sport facilities"* and *"Streets & buildings"* and the percentage of "Very satisfied" citizens. As regard, the indicators b and the percentage of "Very satisfied" citizens, strong linear relationships exist for the categories *"Sport facilities"* and *"Cultural facilities"*. Coefficients of determination values between $0.4 < R^2 \leq 0.6$ indicate moderate linear relationships between the indicators and the percentage of satisfied citizens. As shown in the last columns of Table 8 moderate linear relationships exist between the indicators of the 5 well correlated Eurobarometer categories and the percentage of "Totally satisfied" citizens. Weak and not statistically significant linear relationships ($R^2 < 0.4$ and $p > a = 0.01$) exist between the two indicators and the percentage of "Very satisfied" or "Totally satisfied" citizens for the categories *"Public transport"*, *"Health care services"* and *"Educational facilities"*.

Table 8: Linear regression analysis results between the percentages of "Very satisfied" and "Totally satisfied" citizens as recorded in the Eurobarometer survey and the proposed indicators, where s and b are the proposed indicators, and w is the weight used for the estimation of the indicators. Significance levels are noted as follows: $***\ p \leq 0.001$, $**\ p \leq 0.01$, $*\ p \leq 0.05, ns\ p > 0.05$

	Very satisfied				**Totally (Very or Fairly) satisfied**			
	$s_{u,c,w}$ indicator		$b_{u,c,w}$ indicator		$s_{u,c,w}$ indicator		$b_{u,c,w}$ indicator	
	R^2	*w*	R^2	*w*	R^2	*w*	R^2	*w*
Public spaces	**0.422****	1.5	**0.433****	1.4	**0.562*****	2.3	**0.564*****	1.8
Green spaces	**0.448****	2.5	**0.578*****	1.5	**0.500****	1.9	**0.584*****	1.3
Public transport	0.125	3.4	0.141	0.5	0.180	2	0.349*	0.5
Health care services	0.216	2.3	0.280*	1.5	0.279*	2.1	0.234*	2
Sports facilities	**0.636*****	1.5	**0.665*****	1.2	**0.514****	1.6	**0.455****	1.4
Cultural facilities	**0.519****	1.6	**0.659*****	1.2	**0.517****	1.9	**0.515****	1.5
Streets & buildings	**0.640*****	1.7	**0.561*****	1.4	**0.571*****	1.9	**0.482****	1.6
Educational facilities	0.229	0.5	0.021	0.5	0.262*	0.5	0.018	0.5

In details, the scatter plots of the s and the b indicators versus the percentage of "Very satisfied" and "Totally satisfied" citizens for the "*Public spaces*" and the "*Sports facilities*" categories are presented in Figure 29 and Figure 30 respectively.

Figure 29: Correlation between the proposed indicators of the "*public spaces*" category and the percentage of "Very satisfied" and "Totally satisfied" citizens with regard to "*public spaces*" as recorded in the Eurobarometer survey.

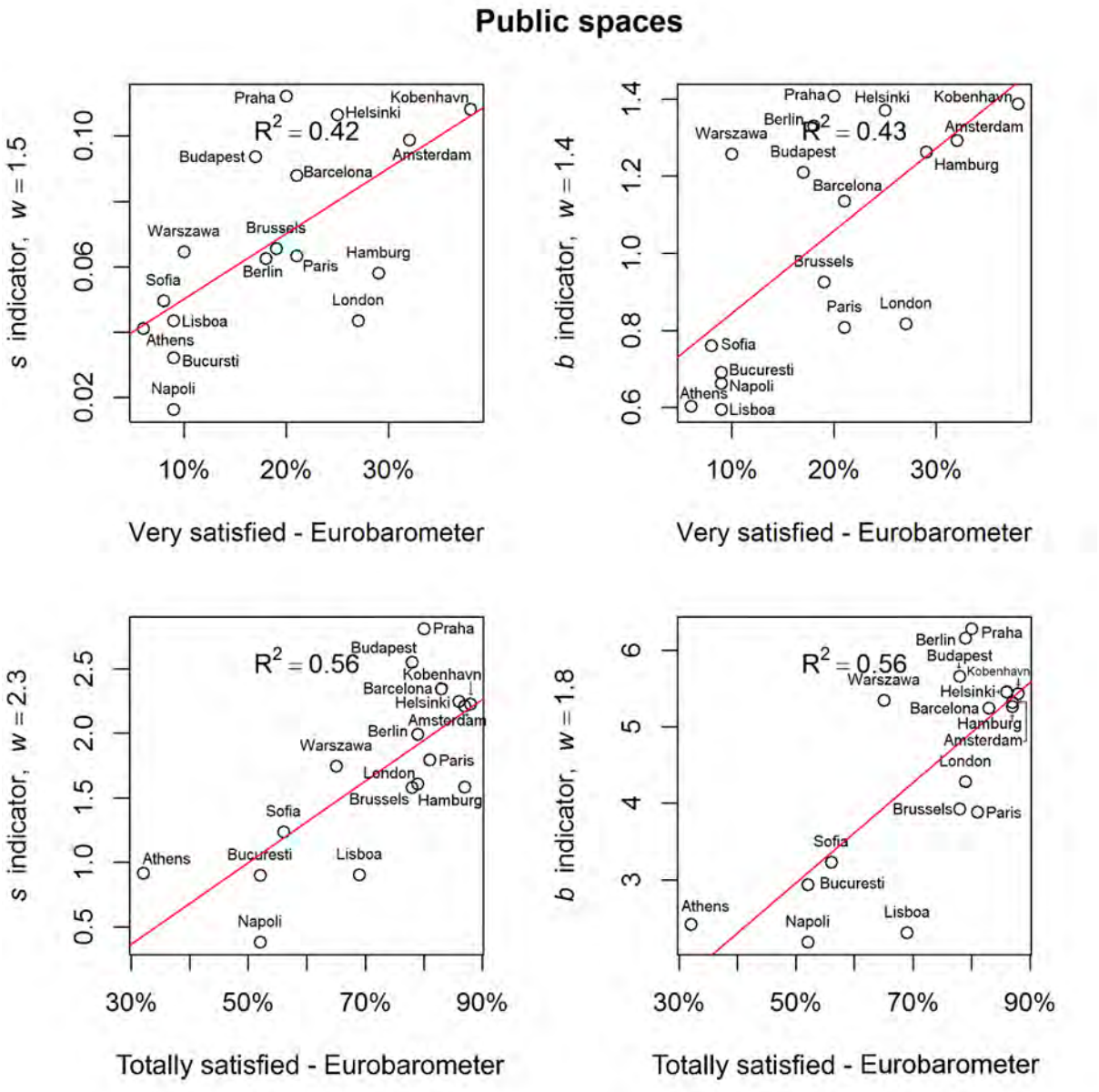
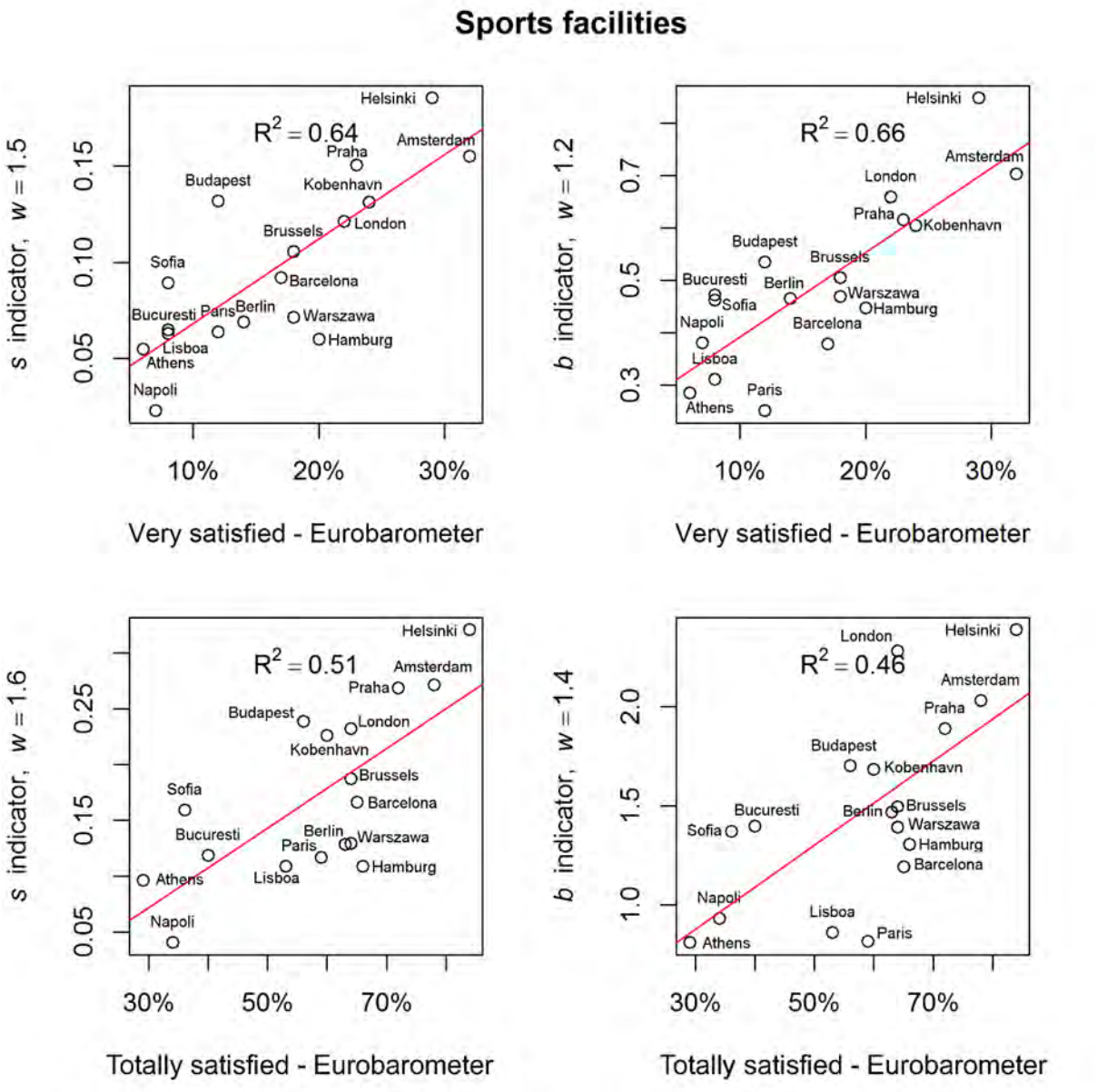


Figure 30: Correlation between the proposed indicators of the "*Sports facilities*" category and the percentage of "Very satisfied" and "Totally satisfied" citizens with regard to "*Sports facilities*" as recorded in the Eurobarometer survey.



5.2.2. IMPACT OF THE WEIGHT VALUES ON THE OF THE ACCURACY

In this Section, we assess the impact of the weight value (w) on the accuracy of the indicators. The optimal weight values are those which were used for the estimation of the best indicators. The best indicators are those for which the highest R^2 values were achieved when compared to the percentage of "Totally satisfied" or "Very satisfied" citizens. For example, the optimal weight value of the s indicator for the *"Sports facilities"* category when compared to the percentage of "Very satisfied" citizens is $w = 1.5$. This is because, as shown in Figure 31, for $w = 1.5$, the R^2 of the "*Sport facility*" category reaches its highest value. In theory, the weight value can also be used for keeping the citizens' satisfaction level in a city with regard to a Eurobarometer category stable, as its population increases. To do so, the places that this category represents should be increasing in relationship to the population increase by a growth rate that is derived by calculating the first derivative of the equation (16), which is defined in equation (17).

$$n_{u,c} = k + j * p_u^{1/w} \quad (16)$$

$$n'_{u,c} = f'(p_u) = \frac{j * p_u^{\frac{-w+1}{w}}}{w} \quad (17)$$

Where p_u is the population of the urban area u, and j is a variable that differs for each category and each weight value, k is a constant value, and w is the proposed weight.

As shown in Table 8, for the well correlated categories ($R^2 > 0.4$), the optimal weight values of the s indicators are higher that the optimal weight values of the b indicators. This is because the number of retail facilities is already logarithmically increasing as the population increases. The optimal weight values of both indicators for most of the well correlated categories, are higher when they are compared to the percentage of "Totally satisfied" citizens than to the percentage of "Very satisfied" citizens. In Figure 31 we present the impact of the weight value on the coefficient of determination (R^2) of the linear relationship between the s indicator and the percentage of "Very satisfied" citizens. These optimal weight values for all the well correlated categories ($R^2 > 0.4$), with the exception of the "*Green spaces*" category, range between $1.5 \leq w \leq 1.7$. In Figure 32 we present the impact of the weight value on the coefficient of determination (R^2) of the linear relationship between the b indicator and the percentage of "Totally satisfied"

citizens. These optimal weights for all the well correlated categories range between $1.3 \leq w \leq 1.8$. As shown in both figures the weight values of the well correlated categories, with the exception of the "*Green spaces*" category, follow a similar pattern. This fact verifies the robustness of the model and enables its reproduction. Since the optimal weight values are not case study dependent, we suggest to those who want to reproduce this study to reuse the weight values found in this study.

Figure 31: The impact of the weight value (w) on the coefficient of determination (R^2) between the s indicator and the percentage of "Very satisfied" citizens

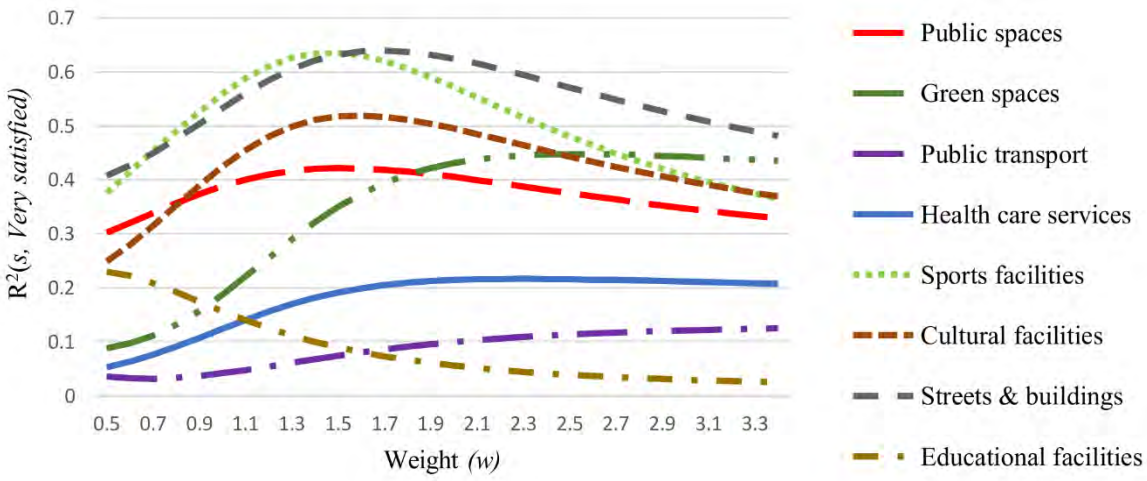
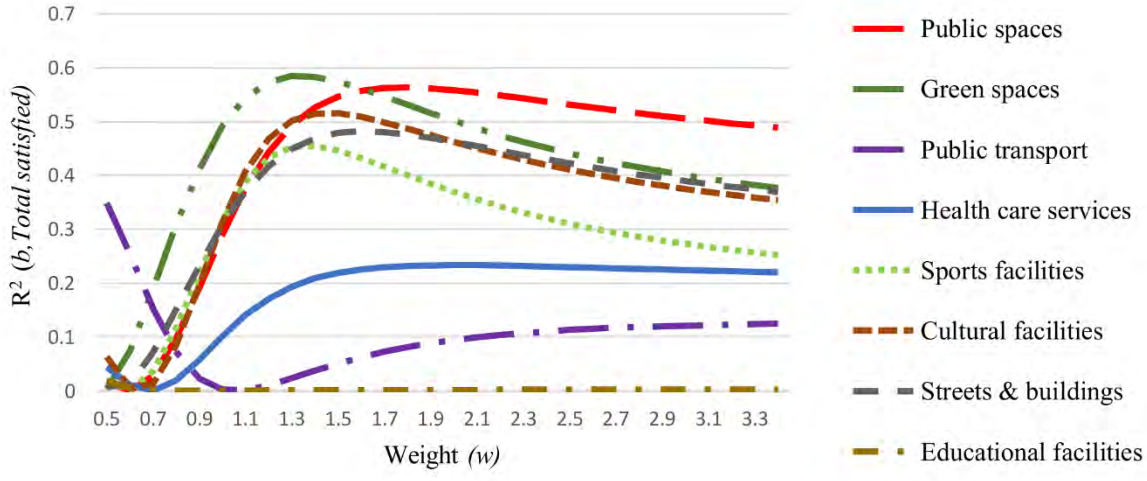


Figure 32: The impact of the weight value (w) on the coefficient of determination (R^2) between the b indicator and the percentage of "Totally satisfied" citizens



6. DISCUSSION & COCLUSIONS

In this Section we discuss the results, we present our conclusions, and we propose a generic framework for using crowd-sourced data. First, in Section 6.1 and Section 6.2 we discuss the results and we present our conclusions from the case study A and the case study B, respectively. In Section 6.3 we propose a generic framework for using crowd-sourced data for various applications. Finally, in Section 6.3, the major conclusions from this thesis are presented.

6.1. CASE STUDY A - ESTIMATING BUILDING BLOCK USE

6.1.1. DISCUSSION

The estimated and the reference BBU datasets differ, since they consist of data which has been collected for different purposes and by data collectors with different characteristics. The reference BBU dataset, consists of LU data collected by LU surveyors (see PGI category in Section 1.3) and it reflects the interests of public authorities, urban planners and managers. The variety of how people perceive and experience space is not expressed in the reference BBU dataset, since in traditional LU surveys, such the one performed in Amsterdam, the users of buildings are not participating through questionnaires or interviews. The estimated BBU dataset, on the other hand, consists of citizen-contributed place data which belong to the SGD category (see Section 1.3). Thus, as discussed in Section 3.1, it mostly reflects, first, the interests of Foursquare users who are mostly young smartphone owners, and second, how space is experienced and subjectively perceived by them. According to Tuan (1979, p. 387) *"the space that we perceive and construct, the space that provides cues for our behaviour, varies with the individual and cultural group".* For instance, the Starbucks stores are perceived by the majority of their users as coffee shops, but some individuals are additionally experience them as co-working spaces (Spyratos et al., 2016).

Despite the conceptual differences between the estimated and the reference BBU datasets, described in the previous paragraph, their comparison revealed interesting scientific outcomes. A major outcome is that by reusing Foursquare place data for LU mapping purposes we identified with high confidence the building blocks of the Amsterdam study area with at least one "*Retail*" or "*Hotels, restaurants & cafes*" LU and the building

blocks of the Varese study area with at least one "*Hotels, restaurants & cafes*" LU (see Table 5 and Table 7). On the contrary, using the proposed methodology we failed to reliably identify building blocks with "*Industries*", "*Storage & unclear*" or "*Parking & public transport*" LU. The reason for the above is that many Foursquare places that belong to these categories are missing from the Foursquare place dataset. The completeness of the Foursquare place dataset varies depending on the LU category (see Figure 11). Since Foursquare users decide which places will be added to the Foursquare application, places of their main interest, such as recreational and commercial places, are the are better represented in the Foursquare place dataset compared to others. For example, in the Amsterdam study area, for the LU category "*Hotels, restaurants & cafes*" there were 8,948 foursquare places compared to the 3,905 surfaces of the non-residential use dataset, while for the LU category "*Industries*" there were only 260 foursquare places compared to the 3,004 surfaces.

An additional scientific outcome of this case study is that Linus' Law (Raymond, 2001) is valid for Foursquare place data. As shown in Figure 14, the positional accuracy of the Foursquare places' location is low. Many foursquare places that describe indoor facilities are falsely described to be located outside the footprint of the building blocks that they belong (see Figure 12). The positional and thematic accuracy of the Foursquare place data is not assured and in-situ controlled through an established quality assurance and quality control mechanism (Spyratos et al., 2016). As is common practice with many crowd-sourced data, the identification and correction of errors in the description and the location of places rely on the volunteering efforts of the Foursquare users. This study proved, that the higher the number of the Foursquare users that have declared a visit in a place is, the higher is the probability that this place is accurately described (see Figure 15).

Finally, it worth mentioning, that limitations are introduced to the accuracy results of the Amsterdam study area, by the use of the non-residential use dataset. As is described in Section 3.2.2.1, this dataset is used for the generation of the reference BBU dataset. These limitations are the following:

- First, the LU survey was limited to the exterior of the buildings and thus some non-residential surfaces might have not been surveyed.
- Second, the non-residential use dataset is outdated in some areas. This is because building surfaces with non-residential use had been surveyed during the time period March 2010 - November 2014. As a result, some new facilities or existing

non-residential facilities that have changed LU might not be included in the non-residential use dataset.

- Third, the *"Storage & unclear"* category includes under a single non differentiable LU category surfaces that have storage, or unclear, or no use. On the Foursquare dataset there are no places with unclear or no use and thus the comparison of this category is problematic.
- Fourth and last, the field surveyors in many cases did not register minor or secondary uses, for instance, the retail shops in the Amsterdam central train station.

These limitations introduce an error in the accuracy assessment of the Amsterdam's estimated BBU dataset, which cannot be quantified.

6.1.2. CONCLUSIONS

In this study, we developed a methodology for the production of low cost LU datasets based on place data contributed by citizens to the Foursquare application. Based on the type of Foursquare places, for example universities and hospitals, we assigned to building blocks, LU categories that describe the types of the activities that are hosted in them. For improving the quality of the LU estimations we selected and filtered the Foursquare places by introducing two parameters. First, the parameter d, which describes the maximum value that the distance from a place to its closest building block may take, in order to take that place into consideration for the estimation of the LU categories of the building block it belongs. Second, the parameter c, which describes the minimum value that the number of Foursquare checked users in a place may take, in order to take that place into consideration for the estimation of the LU categories of the building block it belongs. Since Linus' law applies to Foursquare place data, the c parameter is used for assessing the probability of a Foursquare place to be accurately described. Moreover to evaluate the quality of the estimations we compared the estimated BBU dataset, which represents how Foursquare users perceive space, with the reference BBU dataset, which represents how LU surveyors observe space.

The main study has been conducted in the city of Amsterdam, the Netherlands. To test the reproducibility of the proposed methodology we conducted an additional study in the city of Varese, Italy. Our evaluation metric is the kappa coefficient, which determines if the classification results are significantly better than a random result (Congalton, 1991).

The highest kappa coefficient values, which are indicating substantial agreement between the reference and the estimated BBU datasets for both the Amsterdam and the Varese case studies, were achieved for the "*Hotels, restaurants & cafes*" LU category. The second highest kappa coefficient values for the Amsterdam and the Varese case studies, which are indicating substantial and moderated agreement between the reference and the estimated BBU datasets respectively, were achieved for the "*Retail*" LU category. This is because places of the above two LU categories are of the Fourquare users main interests, and as a results are well represented in the Foursquare place dataset. On the contrary, the proposed methodology failed to identify building blocks with "*Industries*", "*Storage & unclear*" and "*Parking & public transport*" use. The reason for this is that for the above three LU categories the Foursquare dataset is incomplete.

The developed methodology can be utilized for the production of up-to-date, low cost, and globally harmonized datasets about building blocks with "*Hotels, restaurants & cafes*" or "*Retail*" LU. In the future, due to technological and social developments, the quality of citizen-contributed data is expected to be improved. The increasing educational attainment and the wide availability of mobile devices is expected to further increase the use of location-based social media applications. The emergence of indoor position systems is expected to improve the positional accuracy of citizen-contributed geo-referenced data. As a result, data from social media applications can proved to be a valuable source of data for estimating LU. This source of data is especially relevant for cities, where the scarcity of LU data remains a challenge. As future work, one may consider the combine use of foursquare place data with business directories. Such combination might improve the identification rate of LU categories that are underrepresented in crowd-sourced datasets.

6.2.CASE STUDY B - EVALUATING SERVICES AND FACILITIES

6.2.1. DISCUSSION

A facility of a city e.g. a sports complex or a single sport facility, might be represented by none, one or more places of that category. Tuan (2001: 54) defined place as "*enclosed and humanized space*". He additionally supports that "*if we think of space as that which allows movement, then place is pause; each pause in movement makes it possible for location to be transformed into place"* (Tuan, 2001: 6).The number of places that belong to each facility or group of facilities rely on the subjective perception of the Foursquare users. The Foursquare users might experience the space where those facilities are located differently, and as a result they might use diverse places for representing the 'real facility'.

As presented in the Section 5.2.1, strong associations were identified between the percentages of citizens who are very satisfied with regard to "*Sport facilities*", "*Cultural facilities*", and the "S*treets & buildings*", and the two indicators that are proposed by this study. Moderate associations were identified between the percentages of citizens who are satisfied with regard to "*Public spaces", "Green spaces"* and the proposed indicators. Obviously, the correspondence-correlation between these variables does not mean causality. Using the results of this study, we cannot presume that the reason why only a small percentage of citizens are satisfied with a particular facility in a city is the low number of facilities that belong to that category for two reasons. First, as mentioned in the previous paragraph a place does not necessarily correspond to a facility and vice versa. Second, as discussed in Section 4.2.2, low number of Foursquare places that belong to a Eurobarometer category in social media applications might be due to several factors that cannot be identified using the proposed indicators.

The proposed indicators were successful in estimating citizens' satisfaction with regard to Eurobarometer categories that describe facility-oriented uses while were unsuccessful in estimating service-oriented uses. This is because, for example, for the "*Health care*", "*Public transport*", and "*Educational facilities*" categories the quality of service provision plays a major role in the citizens' evaluation compared to the quality of the facility where they are hosted. Another outcome of this study is that the indicators of the three Eurobarometer categories with the highest R^2 values were more successful in estimating the percentage of "Very satisfied" citizens compared to the percentage of "Totally satisfied" citizens. The better detection of the more extreme situations using

crowd-sourced data was also demonstrated by the study of Venerandi et al. (2015). This study revealed that the extremes, the 10% most deprived and the 10% least deprived areas are better detected compared to middle cases using Foursquare and OpenStreetMap data.

The main limitations of this case study are the following. First, as discussed in Section 3.1, the demographic profile of Foursquare users is not representative of society, and thus, it is expected that the place data will represent mostly the activity space of the young people who use Foursquare the most. Second, there is a 2.5-year temporal difference between the two datasets that are used in this study. The Eurobarometer statistics represents the public opinion at the time that the survey took place, between the 15th of November and the 7th of December 2012, while Foursquare place data was harvested between 26th of January and the 4th of May 2015. Last, only Foursquare place data that are located within the administrative boundaries of the urban cities were harvested and analyzed. Thus, services and facilities that are located outside the administrative boundary of an urban city but serve its population were not taken into consideration.

6.2.2. CONCLUSIONS

The case study B demonstrated that citizens' satisfaction with regard to urban facilities and services can be effectively surveyed using crowd-sourced datasets. Two indicators, proposed by this study, are associated with the percentage of citizens' who are according to the Eurobarometer survey satisfied with regard to "*Sport facilities*", "*Cultural facilities*", "*Streets & buildings*", "*Public spaces", and "Green spaces"*. The proposed indicators are based on openly available Foursquare place data, and thus the cost for their estimation is very low. Since Foursquare place data is globally available, these indicators enable the comparison of citizens' satisfaction levels across various cities around the world.

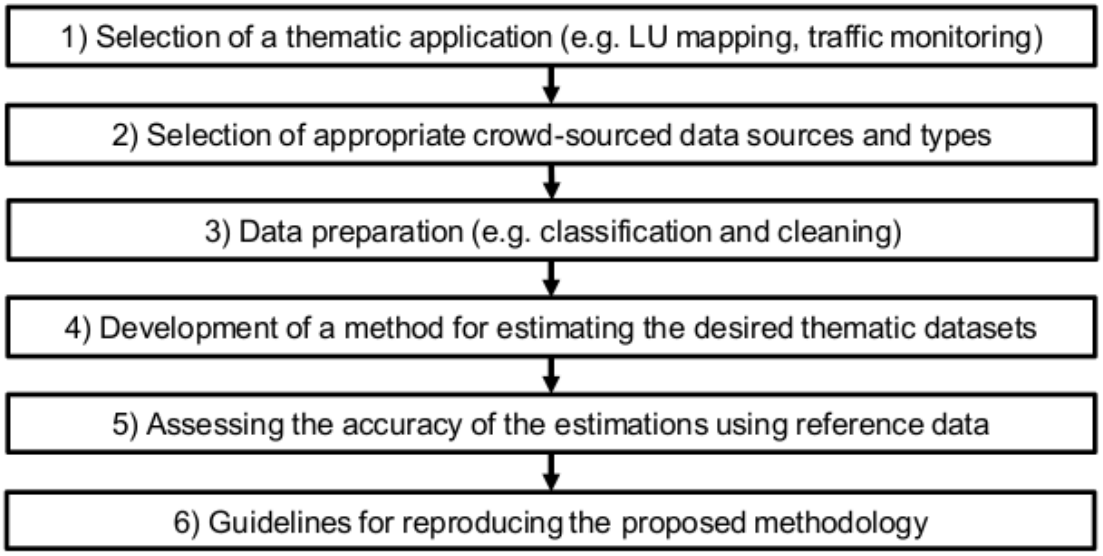
The proposed indicators provide estimates of citizens' satisfaction, but they cannot replace traditional public opinion surveys. They can provide important information for designing public opinion surveys more efficiently and effective by indicating what questions, when and where might be relevant to the public. These indications can minimize the cost and maximize the profit of public opinion surveys, by helping researchers to target them: a) first, spatially into problematic urban areas; b) second, thematically into problematic categories of facilities; and c) last, timely, in time periods that there are enough indications that citizens satisfaction levels have changed. Better

design of public opinion surveys can lead into results that can ultimately facilitate evidence-based decision making.

As future work, one may consider the application of the proposed indicators at neighborhood level. Such analysis might provide more insights, first, about whether is possible to detect inequalities in citizens' satisfaction levels within cities, and second, about the reasons why some facilities are underrepresented or overrepresented in crowd-sourced place datasets.

6.3. GENERIC FRAMEWORK FOR USING CROWD-SOURCED DATA

Figure 33: A generic framework for using crowd-sourced data for multi-thematic applications



In Figure 33, we present a theoretical framework which purpose is to assist researchers and readers in the development of methods that reuse crowd-sourced data for multi-thematic applications. This framework consists of six steps. The first step is the selection of a thematic application. For example, in this thesis we selected two applications, the estimation of building block use (case study A) and the evaluation of urban service and facility types (case study B). The second step is the selection of the appropriate crowd-sourced data sources and data types. For example in both case studies of this thesis we used place data from the Foursquare. The third step is the data preparation, where the crowd-sourced data is classified and cleaned in order to be used as input in the fourth step. The fourth step is the development of a method for estimating the desired thematic dataset from crowd-sourced data observations. In this step, attention need to be given in the temporal, thematic and spatial cohesion of the crowd-sourced data observations which will form the estimated thematic dataset. In the fifth step the accuracy of the estimated dataset will be assessed by using a reference dataset which corresponds temporally, thematically and spatially to the estimated. Finally, guidelines for reproducing the proposed methodology in other case study areas need to be given. For example, for reproducing the case study A guidelines are given in Section 5.1.6 and for the case study B in the last paragraph of Section 5.2.2.

6.4.OVERALL DISCUSSION & CONCLUSIONS

Nowadays, due to the widespread use mobile devices with numerous embedded sensors, an increasing number of citizens is collecting various types of crowd-sourced geographical, environmental and geo-referenced data. This data is collected for both science-oriented and socially-oriented purposes and is freely contributed by citizens to Internet-based applications. Part of this massive amount of citizen-contributed data is accessible to third parties at no or limited cost. In order to assess the usefulness of this new data source for urban planning we performed two case studies which are answering to the two research questions of this PhD.

The two case studies demonstrated that by using appropriate methods such as those developed in this PhD research, crowd-sourced data can be reused for spatial and urban planning purposes. The case study A, entitled "*Estimating building block use*" demonstrated that crowd-sourced data can be utilized for the production of up-to-date, low cost, and globally harmonized datasets about building blocks with "*Hotels, restaurants & cafes*" or "*Retail*" land use. Such data, due to its dynamic nature, is useful, for example, to urban planners and real estate professionals who are interested on studying spatiotemporal changes on the distribution of recreational and retail facilities.

The case study B, entitled "*Evaluating the services and facilities*" demonstrated that Foursquare place data can be used for generating indicators which are estimating citizens' satisfaction with regard to specific urban facilities. The proposed indicators can provide low cost, up-to-date and globally harmonized estimates about citizens' satisfaction with regard to "*Sport facilities*", "*Cultural facilities*" and "*Streets & buildings*". The proposed indicators can be used for identifying changes in citizens' satisfaction levels, for highlighting problematic areas or facility types, and for comparing citizens' satisfaction levels across multiple cities. All these estimates can assist researchers to design public opinion surveys more efficiently and effective by indicating what questions, when and where might be relevant to the public.

Crowd-sourced data represents how the data contributors subjectively experience and perceive space. The availability of such data, which is expected to be increased in the near future, can facilitate evidence-based decision making processes. The use of that data can be proved to be a step towards citizens' participation in the governance of urban areas. Its analysis can highlight major environmental, economic and social processes and issues

in the areas under investigation that require further investigation. However, the use of crowd-sourced data must not been used as a form of frictionless participation of citizens in decision making processes. Such use cannot replace traditional forms of citizens' participation in decision making, such as questioners and deliberation processes.

A major drawback of the use of crowd-sourced data for spatial and urban planning is that social media applications are used by a non-representative subset of the society. Less populated and less popular areas, as well as the activity space of citizens that are not contributing data to Internet applications, are underrepresented in such datasets. Additionally, since the production of crowd-sourced data is often controlled by "closed" organizations who store and process that data on their servers, there exists the risk that these organizations might manipulate these data for supporting their interests (Tenney and Sieber, 2016). To minimize all these limitations, it is necessary to analyse the characteristics of crowd-source data in detail before using them. Otherwise, a non-well examined use of crowd sourced data might lead to the partial understanding of the environment and to decisions that harm or ignore some parts of the society.

REFERENCES

Antoniou V, Morley J and Haklay M (2010) Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon. *Geomatica* 64(1): 99–110.

Benkler Y (2002) Coase's Penguin, or, Linux and' The Nature of the Firm'. *Yale Law Journal* 112(3): 369–446.

Benkler Y (2006) *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. *Yale University Press*, New Haven, CT.

BOP Consulting (2013) *The Economic , Social and Cultural Impact of the City Arts and Culture Cluster*. London: City of London Corporation

Brabham DC (2008) Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence: The International Journal of Research into New Media Technologies* 14(1): 75–90.

Budhathoki N, Haklay M and Nedovic-budic Z (2010) Who map in OpenStreetMap and why? In: *Presentation at the State of the Map conference, Atlanta, USA, 2010*.

Burrough PA and Frank A (1996) *Geographic objects with indeterminate boundaries.* London: Taylor and Francis

Büttner G and Barbara K (2007) *CLC2006 technical guidelines*. *European Environment Agency, Technical Report*, EEA Technical report, Copenhagen.

Carletta J (1996) Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, Computation and Language 22(2): 249–254.

Cattell V, Dines N, Gesler W, et al. (2008) Mingling, observing, and lingering: everyday public spaces and their implications for well-being and social relations. *Health & place* 14(3): 544–561.

Cheshire P and Sheppard S (1995) On the price of land and the value of amenities. *Economica* 62: 247–267.

City of Amsterdam (2011) *Functiekaart-classificatie 2011-6.* Amsterdam. Available from: https://drive.google.com/file/d/0BxRbFXteeEQjSHViaThaWUxSNGc/view?usp=sharing.

City of Amsterdam (2015) Non-residential functions (Function Map). Available from: http://maps.amsterdam.nl/open_geodata/ (accessed 28 July 2015).

Cohen J (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20(1): 37–46.

Cohen J, Cohen P, West SG, et al. (2013) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Third Edit. Mahwah, NJ: Lawrence Erlbaum Associates.

Congalton RG (1991) A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment* 37(1): 35–46.

Cormode G and Krishnamurthy B (2008) Key Differences between Web 1.0 and Web 2.0. *First Monday* 13(6). doi:10.5210/fm.v13i6.2125

Couclelis H (1992) Location, place, region, and space. In: Abler, R. F., Marcus, M. G. and Olson JM (ed.), *Geography's inner worlds Pervasive themes in contemporary American geography*, New Brunswick NJ: Rutgers University Press, pp. 215–233.

Craglia M, Ostermann F and Spinsanti L (2012) Digital Earth from vision to practice: making sense of citizen-generated content. *International Journal of Digital Earth* 5(5): 398–416.

Cramer H, Rost M and Holmquist LE (2011) Performing a Check-in: Emerging Practices, Norms and 'Conflicts' in Location-Sharing Using Foursquare. In: *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, Aug 30 – Sept 2, 2011, Stockholm, Sweden: New York: ACM, pp. 57–66.

Delaney DG, Sperling CD, Adams CS, et al. (2008) Marine invasive species: validation of citizen science and implications for national monitoring networks. *Biological Invasions* 10(1): 117–128.

Elwood S (2008) Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal* 72(3): 173–183.

Elwood S, Goodchild MF and Sui DZ (2012) Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers* 102(3): 571–590.

EPA (2014) Corine Land Cover Mapping. *Environmental Protection Agency*. Available from: http://www.epa.ie/soilandbiodiversity/soils/land/corine (accessed 8 October 2015).

Estellés-arolas E and González-ladrón-de-guevara F (2012) Towards an integrated crowdsourcing definition. *Journal of Information Science* 38(2): 1–14.

European Commission (2013) *Flash Eurobarometer 336 - Quality of life in cities: Perception survey in 79 European cities*. Luxembourg.

European Union (2011) *Mapping Guide for a European Urban Atlas*. Available from: https://cws-download.eea.europa.eu/local/ua2006/Urban_Atlas_2006_mapping_guide_v2_final.pdf (accessed 20 July 2016).

Eurostat (2015) Land cover and land use (LUCAS) statistics. Available from: http://ec.europa.eu/eurostat/statistics-explained/index.php/Land_cover_and_land_use_(LUCAS)_statistics (accessed 29 July 2016).

Everitt B and Skrondal A (2010) *The Cambridge dictionary of statistics*. fourth edi. Cambridge: Cambridge University Press.

Facebook (2015) The Graph API. Available from: https://developers.facebook.com/docs/graph-api (accessed 12 September 2015).

Facebook (2016) About Facebook. Available from: https://www.facebook.com/about/ (accessed 22 March 2016).

Fan H, Zipf A and Qing F (2014) Estimation of Building Types on OpenStreetMap Based on Urban Morphology Analysis. In: Huerta J, Schade S, and Granell C (eds), *Connecting a Digital Europe Through Location and Place*, Springer International Publishing, pp. 19–35.

Farmer B (1993) Needs and means. In: Farmer B and Louw HJ (eds), *Companion to contemporary architectural thought*, London: Routledge, pp. 21–28.

Felstiner A (2011) Working the crowd: employment and labor law in the crowdsourcing industry. *Berkeley Journal of Employment and Labor Law* 32(1): 143–204.

Floris R and Zoppi C (2015) Social Media-Related Geographic Information in the Context of Strategic Environmental Assessment of Municipal Masterplans: A Case Study Concerning Sardinia (Italy). *Future Internet* 7(3): 276–293.

Fort K, Adda G and Cohen K (2011) Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics* 37(2): 413–420.

Foursquare (2015a) Foursquare for Developers - Foursquare Category Hierarchy. Available from: https://developer.foursquare.com/categorytree (accessed 20 January 2015).

Foursquare (2015b) Foursquare for Developers - Venues Service. Available from: https://developer.foursquare.com/overview/venues.html (accessed 12 February 2015).

Foursquare (2016a) About Foursquare. Available from: https://foursquare.com/about (accessed 8 January 2016).

Foursquare (2016b) Venues Service. Available from: https://developer.foursquare.com/overview/venues.html (accessed 22 February 2016).

Frank LD, Sallis JF, Conway TL, et al. (2006) Many Pathways from Land Use to Health and Air Quality. *Journal of the American Planning Association* 72(1): 75–87.

Frias-Martinez V, Soto V, Hohwald H, et al. (2012) Characterizing Urban Landscapes Using Geolocated Tweets. In: *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 3-5 September 2012 Amsterdam, Netherlands: California: IEEE, pp. 239–248.

Fritz S, See L, van der Velde M, et al. (2013) Downgrading recent estimates of land available for biofuel production. *Environmental science & technology* 47(3): 1688–94.

Geo-Wiki (2016) The Geo-Wiki Project. Available from: http://www.geo-wiki.org/ (accessed 8 February 2016).

Giles-Corti B, Broomhall MH, Knuiman M, et al. (2005) Increasing walking: how important is distance to, attractiveness, and size of public open space? *American journal of preventive medicine* 28(2): 169–176.

Girres J-F and Touya G (2010) Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS* 14(4): 435–459.

Goodchild M (2008) Assertion and authority: the science of user-generated geographic content. In: *Proceedings of the Colloquium for Andrew U. Frank's 60th Birthday . GeoInfo 39. Department of Geoinformation and Cartography, Vienna University of Technology*, Vienna, Austria.

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4): 211–221.

Goodchild MF and Li L (2012) Assuring the quality of volunteered geographic information. *Spatial Statistics* 1: 110–120.

Google (2016) What is Google Map Maker? Available from: https://support.google.com/mapmaker/answer/157176?hl=en&ref_topic=1093469&rd=1 (accessed 29 July 2016).

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.

Haklay M and Weber P (2008) OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing* 7(4): 12–18.

Haklay M, Basiouka S, Antoniou V, et al. (2010) How many volunteers does it take to map an area well? The validity of Linus' law to volunteered geographic information. *Cartographic Journal* 47(4): 315–322.

Howe J (2006a) Crowdsourcing: A Definition. *Crowdsourcing: Tracking the Rise of the Amateur*. Available from: http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html (accessed 26 July 2016).

Howe J (2006b) The Rise of Crowdsourcing. *Wired Magazine* 14.06: 1–5. Available from: http://www.wired.com/wired/archive/14.06/crowds.html.

Huang H, Kieler B and Sester M (2013) Urban building usage labeling by geometric and context analyses of the footprint data. In: *Proceeding of 26th international cartographic conference (ICC)*, Dresden, Germany.

Isaac M (2016) Foursquare Raises $45 Million, Cutting Its Valuation Nearly in Half. *The New York Times*. Available from: http://www.nytimes.com/2016/01/15/technology/foursquare-raises-45-million-cutting-its-valuation-nearly-in-half.html?_r=0.

ISPEX (2016) About iSPEX. Available from: http://ispex-eu.org/ (accessed 8 February 2016).

Jokar Arsanjani J, Helbich M, Bakillah M, et al. (2013) Toward mapping land-use

patterns from volunteered geographic information. *International Journal of Geographical Information Science* 27(12): 2264–2278.

Koukoletsos T, Haklay M and Ellul C (2012) Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS* 16(4): 477–498.

Landis JR and Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159–174.

Lee V (1994) Volunteer monitoring: a brief history. *The Volunteer Monitor* 6(1): 29–33.

Litman T and Steele R (2014) *Land Use Impacts on Transport: How Land Use Factors Affect Travel Behavior*. Available from: http://www.vtpi.org/landtravel.pdf.

NBN (2013) New NBN Record Cleaner Rules now available. Available from: http://www.nbn.org.uk/News/Latest-news/New-Record-Cleaner-Rules-now-available.aspx (accessed 21 January 2016).

NOAA (2014) What is the Coop Program? Available from: http://www.nws.noaa.gov/om/coop/what-is-coop.html (accessed 29 July 2016).

NoiseWatch (2013) About NoiseWatch platform. *European Environment Agency*. Available from: https://drive.google.com/file/d/0BxRbFXteeEQjVm5KZXBWN1Rfczg/view?usp=sharing (accessed 18 April 2014).

Noulas A, Mascolo C and Frias-Martinez E (2013) Exploiting Foursquare and Cellular Data to Infer User Activity in Urban Environments. In: *14th International Conference on Mobile Data Management*, 3–6 June 2013 Milan, Italy: California :IEEE, pp. 167–176.

Oort P van (2006) Spatial data quality: from description to application. *Wageningen: Wageningen Universiteit. PhD thesis*, Wageningen Universiteit.

OpenStreetMap (2015) Nominatim geocoding service. Available from: http://wiki.openstreetmap.org/wiki/Nominatim (accessed 19 July 2015).

OpenStreetMap (2016) About OpenStreetMap. Available from: https://www.openstreetmap.org/about (accessed 22 February 2016).

Parker CJ, May A and Mitchell V (2012) Understanding Design with VGI using an Information Relevance Framework. *Transactions in GIS* 16(4): 545–560.

Pei T, Sobolevsky S, Ratti C, et al. (2014) A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science* 28(9): 1988–2007.

Pérez-Lombard L, Ortiz J and Pout C (2008) A review on buildings energy consumption information. *Energy and Buildings* 40(3): 394–398.

Powell LM, Slater S, Chaloupka FJ, et al. (2006) Availability of physical activity-related facilities and neighborhood demographic and socioeconomic characteristics: A national study. *American Journal of Public Health* 96(9): 1676–1680.

Preotiuc-Pietro D, Cranshaw J and Yano T (2013) Exploring venue-based city-to-city similarity measures. In: *2nd ACM SIGKDD International Workshop on Urban Computing*, Chicago, Illinois, USA, August 11 - 14, 2013: New York: ACM, pp. 1–4.

Prins RG, Van Empelen P, Te Velde SJ, et al. (2010) Availability of sports facilities as moderator of the intention-sports participation relationship among adolescents. *Health Education Research* 25(3): 489–497.

Quercia D and Saez D (2014) Mining Urban Deprivation from Foursquare: Implicit Crowdsourcing of City Land Use. *Pervasive Computing, IEEE* 13(2): 30–36.

Rawlings JO, Pantula SG and Dickey DA (1998) *Applied Regression Analysis : A Research Tool ,Texts in Statistics*. Second Edi. New York: Springer-Verlag.

Raymond E (2001) *The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. Revised Ed. O'Reilly Media, Inc.

Salesses P, Schechtner K and Hidalgo CA (2013) The Collaborative Image of The City : Mapping the Inequality of Urban Perception. *PLoS ONE* 8(7): e68400.

See L, Comber A, Salk C, et al. (2013) Comparing the quality of crowdsourced data contributed by expert and non-experts. *PloS ONE* 8(7): e69958.

Soto V and Frias-Martinez E (2011) Robust land use characterization of urban landscapes using cell phone data. In: *First Workshop on Pervasive Urban Applications (PURBA)*, San Francisco, CA.

Spyratos S, Lutz M and Pantisano F (2014) Characteristics of Citizen - contributed Geographic Information. In: *AGILE'2014 International Conference on Geographic Information Science, June, 3-6*, Castellón, Spain.

Spyratos S, Stathakis D, Lutz M, et al. (2016) Using Foursquare place data for estimating building block use. *Environment and Planning B: Planning and Design*. First Published 27 Jul 2016. doi: 10.1177/0265813516637607.

Stephens M (2013) Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal* 78(6): 981–996.

Stucky TD and Ottensmann JR (2009) Land Use And Violent Crime. *Criminology* 47(4): 1223–1264.

Taylor R (1990) Interpretation of the Correlation Coefficient: A Basic Review. *Journal of Diagnostic Medical Sonography* 6(1): 35–39.

Tenney M and Sieber R (2016) Data-Driven Participation : Algorithms , Cities , Citizens , and Corporate Control. *Urban Planning* 1(2): 101–113.

Toole JL, Ulm M, González MC, et al. (2012) Inferring land use from mobile phone activity. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp '12*, New York, New York, USA: ACM Press, pp. 1–8.

Tuan Y (1979) Space and Place: Humanistic Perspective. In: Gale S and Olsson G (eds), *Philosophy in Geography*, Springer Netherlands, pp. 387–427.

Tuan Y (2001) *Space and place: The perspective of experience*. 5th ed. Minneapolis, USA: University of Minnesota Press.

Twitter (2016) About public and protected Tweets. Available from: https://support.twitter.com/articles/14016# (accessed 22 February 2016).

United Nations (2015) *World Urbanization Prospects The 2014 Revision*. New York, USA.

USGS (2013) Crowd-Sourcing the Nation: Now a National Effort. Available from: https://drive.google.com/file/d/0BxRbFXteeEQjM005emZRd0JwY2M/view?usp=sharing (accessed 15 September 2014).

Van Lenthe FJ, Brug J and MacKenbach JP (2005) Neighbourhood inequalities in physical inactivity: The role of neighbourhood attractiveness, proximity to local facilities and safety in the Netherlands. *Social Science and Medicine* 60(4): 763–775.

Venerandi A, Quattrone G, Capra L, et al. (2015) Measuring Urban Deprivation from User Generated Content. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, Vancouver, BC, Canada — March 14 - 18, 2015: New York: ACM, pp. 254–264.

Vickery G and Wunsch-Vincent S (2007) *Participative Web and User-Created Content: Web 2.0 wikis and social networking*. Paris: OECD Publishing.

Viera AJ and Garrett JM (2005) Understanding interobserver agreement: The kappa statistic. *Family Medicine* 37(5): 360–363.

VU geoplaza (2015) Basisregistraties Adressen en Gebouwen (BAG) 2013. Available from: http://geoplaza.vu.nl/data/dataset/bag (accessed 9 February 205AD).

Waze (2015) Waze, About Us. Available from: https://www.waze.com/about (accessed 15 January 2016).

Zandbergen P a. and Barbeau SJ (2011) Positional Accuracy of Assisted GPS Data from High-Sensitivity GPS-enabled Mobile Phones. *Journal of Navigation* 64(3): 381–399.

Appendix 1: A table that presents the matching between the 712 Foursquare place categories (as of 01/2015) and the Amsterdam LU subcategories and categories.

Url to file:

https://drive.google.com/file/d/0BxRbFXteeEQjTGt1N2JsMmVUMms/view?usp=sharing

Appendix 2: A table that presents the matching between the Eurobarometer categories plus the Retail category with the Foursquare place categories (as of 01/2015)

Url to file:

https://docs.google.com/spreadsheets/d/1oooO6UK_XyCtLaBMAo_MgnVpan83HfkHGS5Glvt2-l4/edit?usp=sharing

Appendix 3: Number of Foursquare places for each Eurobarometer category and each city

	Public space	Green spaces	Public transport	Health care services	Sports facilities	Cultural facilities	Streets & buildings	Educational facilities	Retail shops
Lisboa	266	177	472	1001	385	496	1979	955	5145
Praha	1178	890	650	2794	1582	1602	4088	3236	12359
Helsinki	684	719	799	808	1191	858	5254	3136	5985
Brussels	618	505	1633	2410	998	1448	5815	2971	9011
Amsterdam	750	444	864	1543	1178	1619	3935	1560	7398
Barcelona	1109	481	1571	1972	1164	1467	3842	2034	15343
Athens	311	125	332	1308	417	677	1065	1085	6280
Kobenhavn	649	494	601	1196	788	1082	3067	1071	5474
Sofia	514	407	703	1205	927	796	2783	1544	9176
Napoli	141	59	303	228	198	198	528	409	1816
Warszawa	847	612	1867	1664	937	941	3651	2973	9121
Paris	952	608	671	2840	961	2854	7094	2878	19966
London	1405	2435	3721	4059	3921	3671	14739	5119	33820
Berlin	1311	946	959	3432	1446	2299	4273	2025	15517
Hamburg	780	568	765	2169	806	972	2439	1238	8071
Budapest	1254	872	520	2964	1766	1531	4689	3776	16676
Bucuresti	462	551	1000	1882	934	844	4086	6151	9023