UNIVERSITY OF SÃO PAULO
POLYTECHNIC SCHOOL

JULIANA SIQUERA GAY

Learning spatial inequalities: an approach to support transportation planning

São Paulo
2018

JULIANA SIQUEIRA GAY

Learning spatial inequalities: an approach to support transportation planning

Dissertation submitted to Polytechnic School at University of São Paulo for the grant of Master of Science degree.

Supervisor: Prof. Dr. Mariana Abrantes Giannotti

São Paulo
2018

Este exemplar foi revisado e alterado em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, ... de abril de 2018

Assinatura do autor

Assinatura do orientador

Catalogação-na-publicação

Siqueira-Gay, Juliana
Aprendizagem sobre desigualdades espaciais: uma abordagem para suporte ao planejamento de transportes / J. Siqueira-Gay -- versão corr. -- São Paulo, 2018.
95 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Transportes.

1.Sistemas de Transportes 2.Planejamento de transportes 3.Desigualdades 4.Aprendizagem computacional 5.Acessibilidade I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Transportes II.t.

Nome:
Título:

Dissertação apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção do
título de Mestre em Ciências

Aprovado em:

Banca examinadora

Prof. Dr. _____
Instituição: _____
Julgamento: _____

Prof. Dr. _____
Instituição: _____
Julgamento: _____

Prof. Dr. _____
Instituição: _____
Julgamento: _____

Ao Alfredo Gay Neto, meu amado esposo e companheiro

This work was developed in collaboration with Professor Monika Sester, head from the *Institut für Kartographie und Geoinformatik* at *Leibniz Universität Hannover*

Acknowledgments

I would like to acknowledge the financial support of Coordination for the Improvement of Higher Education Personnel (CAPES) during this master research.

I would like to thank Professor Mariana Abrantes Giannotti for the opportunity and liberty to conduct this research. It was considerably valuable for me to build up my own paths as a beginner in the research area. I acknowledge Professor Monika Sester for receiving me at the IKG and giving me all support in this work. I also thank to Professor Flavia Feitosa and Professor Felipe Loureiro for the valuable comments in the previous version of this work.

I also wish to express my appreciation to the professors of the disciplines that I made, especially Alfredo Pereira de Queiroz Filho, for being very didactic and clarifying some important research concepts.

A warm thank you to Professor Amarilis Lucia Figueiredo Gallardo and Professor Luis Enrique Sánchez for always guiding me and foremost, for encouraging me to go further.

Many thanks go to my colleagues from LabGeo, especially Diego Tomasiello and Pedro Logiodice for the data support and very good talks. Thanks to the IKG staff, especially Stefania Zourlidou for our long and deep conversations about research and life. Thanks to my loved friends and coworkers, Ana Paula Dibo, Barbara Souza, Carla Grigoletto Duarte and Josianne Rosa. Thank you for being with me, without friends and colleagues I would not achieve this task.

Finally, some words to whom I am completely grateful. To Nilce Siqueira, Mara Regina Garcia Gay and Alfredo Gay Junior for all love and care.

My deepest gratitude to Alfredo, for my daily smiles and for being with me every time that I needed. As well, for all technical support during this research, guiding and encouraging me in the difficult moments.

“The art of writing is the art of discovering what you believe”

(Gustave Flaubert)

Resumo

Título: Aprendizagem sobre desigualdades espaciais: uma abordagem para suporte ao planejamento de transportes. Dissertação de Mestrado.

Parte da literatura de planejamento de transportes conceitua a infraestrutura de transportes como uma forma de distribuir pessoas e oportunidades no território. Portanto, as desigualdades espaciais tornaram-se uma questão relevante a ser endereçada no planejamento de transportes e uso do solo. De maneira a contribuir com o desafio de avaliar desigualdades e sua heterogeneidade no ambiente urbano, esse trabalho tem como objetivo identificar e descrever padrões existentes na distribuição acessibilidade a diferentes equipamentos urbanos e dados socioeconômicos por meio de técnicas de Aprendizagem de Máquina (AM) para informar a tomada de decisão em planos de transportes. De forma a caracterizar a atual consideração de métricas de desigualdades espaciais na prática do planejamento de transportes no Brasil, nove planos de mobilidade foram revisados. Para investigar as potencialidades e restrições da aplicação de AM, análises supervisionadas e não supervisionadas de indicadores de renda e acessibilidade a saúde, educação e lazer foram realizadas. Os dados do município de São Paulo dos anos de 2000 e 2010 foram explorados. Os Planos de Mobilidade analisados não apresentam medidas para avaliação de desigualdades espaciais. Além disso, é possível identificar que a população de baixa renda tem baixa acessibilidade a todos os equipamentos urbanos, especialmente hospitais e centros culturais. A zona leste da cidade apresenta um grupo de baixa renda com nível intermediário de acessibilidade a escolas públicas e centros esportivos, evidenciando a heterogeneidade nas regiões periféricas da cidade. Finalmente, um quadro de referência é proposto para incorporação de técnicas de AM no planejamento de transportes.

Palavras-chave: Sistemas de Transportes; Planos de Mobilidade; Acessibilidade; Aprendizado computacional; São Paulo; Planejamento urbano.

Abstract

Part of the literature of transportation planning understand transportation infrastructure as a mean of distributing people and opportunities across the territory. Therefore, the spatial inequalities become a relevant issue in transportation and land use planning. To meet the challenge of evaluating the heterogeneity of transportation provision and land use in the urban environment, this work aims at identifying and describing patterns hidden the distribution of accessibility to different urban facilities and socioeconomic information using Machine Learning (ML) techniques to inform the decision making of transportation plans. To feature the current consideration of spatial inequalities measures in the practice of transportation planning in Brazil, nine mobility plans were reviewed. For investigating the potentialities and restrictions of ML application, unsupervised and supervised analysis of income and accessibility indicators to health, education and leisure were performed. The data of the São Paulo municipality from the years of 2000 and 2010 was explored. The analyzed plans do not present measures for evaluating spatial inequalities. It is possible to identify that the low-income population has low accessibility to all facilities, especially, hospital and cultural centers. The east zone of the city presents a low-income group with intermediate level to public schools and sports centers, revealing the heterogeneity in regions out of the city center. Finally, a framework is proposed to incorporate spatial inequalities by using ML techniques in transportation plans.

Keywords: Transport systems; Mobility Plans; Accessibility; Machine learning; São Paulo; Brazil

List of Figures

Figure 1 – Dissertation outline 20
Figure 2 - Municipalities with analyzed mobility plans and the São Paulo study area 22
Figure 3 - Census tracts 2000 (left) and 2010 (right) 25
Figure 4 – Spatial patterns and values of income (m.w.) in 2000 and 2010 with the division method used was natural breaks 26
Figure 5 – Metro and train transit lines in the São Paulo municipality: yellow and lilac lines were built after 2000 28
Figure 6 - Data used to calculate the accessibility measures 30
Figure 7 - Main steps of accessibility measures 30
Figure 8 – Research framework of unsupervised analysis 32
Figure 9 - Research framework of supervised analysis 33
Figure 10 – Urban Mobility Plans contents 38
Figure 11 – Machine learning techniques 43
Figure 12 – Principal components analysis 44
Figure 13 – Distances measures between clusters 45
Figure 14 - Main clustering algorithm classes 46
Figure 15 – K-means algorithm 47
Figure 16 – Main classification algorithms classes 50
Figure 17 – Main regression algorithms classes 51
Figure 18 - Accessibility measures of 2000 and 2010 55
Figure 19 – The proportion of the original variables in each eigenvector for 2000 dataset 58
Figure 20 – Principal components of 2000 dataset 59
Figure 21 - The proportion of the original variables in each eigenvector for 2010 dataset 61
Figure 22 – Principal Components of 2010 dataset 62
Figure 23 – Elbow curve 63
Figure 24 - Histogram of 2000 income 64
Figure 25 – Histogram of 2010 income 65
Figure 26 – Map and cluster's composition of 2000 dataset 67
Figure 27 - Map and clusters composition of 2010 dataset 68
Figure 28 – Surroundings of metro stations of 2000 (left) and 2010 (right) 71
Figure 29 – Check list for unsupervised analysis 80

Figure 30 – Check list for supervised analysis 81
Figure 31 – Framework of the contribution of ML techniques to transportation planning..... 83

List of Tables

Table 1 – Research questions and hypothesis of quantitative analysis 19
Table 2 – Mobility Plans reviewed 23
Table 3 – Variable of 2000 and 2010 Census used in the analysis 26
Table 4 – Metro and train stations implemented after 2000 27
Table 5 - Accessibility measures 31
Table 6 – Main findings in the analyzed documents 39
Table 7 – Summary of methods according to practical criteria 48
Table 8 – Descriptive statistic of 2000 dataset 56
Table 9 – Descriptive statistic of 2010 dataset 56
Table 10 – Correlation matrix of the 2000 dataset 57
Table 11 – Cumulative percentage of variance of the 2000 dataset 58
Table 12 – Correlation matrix of 2010 dataset 60
Table 13 - Cumulative percentage of variance of the 2010 dataset 60
Table 14 - Models parameters 64
Table 15 – Quartile limits of income in minimum wages distribution – all data 65
Table 16 – Descriptive values of income in minimum wages of each cluster of 2000 65
Table 17 Descriptive values of income in minimum wages of each cluster of 2010 66
Table 18 – The summary of clusters composition - 2000 dataset 69
Table 19- The summary of clusters composition – 2010 dataset 70
Table 20 – Model's parameter of the regression of income considering 2000 72
Table 21 - Model's parameter of the regression of income considering 2010 72
Table 22 – Summary of potentialities and risks of the application of ML techniques to explore inequitable distribution of opportunities 79

Contents

1 Introduction 15

1.1 Objectives 17

1.2 Research questions 18

1.3 Dissertation outline 19

2 Materials and methods 22

2.1 Mobility Plans review 22

2.2 Spatial inequalities analysis 24

2.2.1 Census data 24

2.2.2 Accessibility measures 27

2.2.3 Data analysis 31

3 Spatial inequalities in transportation planning 34

3.1 The rationality of spatial inequalities in transportation planning 34

3.2. Brazilian Mobility Plans 36

4 Machine learning 42

4.1 Unsupervised learning 43

4.1.1 Dimensionality reduction 44

4.1.2 Clustering 45

4.2 Supervised learning 49

4.3 Machine learning techniques for spatial data 52

5 Spatial inequalities in São Paulo 54

5.1 The indicators formulation 54

5.2 Unsupervised analysis 56

5.2.1 Dimensionality reduction 57

5.2.2 Clustering 63

5.3 Supervised analysis 71

5.3.1 Linear regression 71

5.4 Discussion 73

6 Outcomes for transportation planning 77

6.1 How could spatial inequalities be explored by using ML techniques to inform transportation planning? 78

6.2 What is the relation of ML contribution with the existing transportation planning structure? 82

7 Conclusions 85

References 87

1 Introduction

(In)equality is related to the distribution of goods, income, and opportunities (WEE; GEURS, 2011). The territory is the ground where the inequalities take place and the relations become concrete. It is in the territory that citizens interact with services and community, evidencing the existing unequal distribution of opportunities (KOGA, 2011). Because of this, the spatial dimension of inequalities is an intrinsic attribute to be investigated regarding the outcomes of public policies. Wei (2015) points out the need for studies that explore the effects on the spatiality of regional income/economic inequalities in a context of broad issues of social justice and sustainability. In the transportation literature, Manaugh; Badami; El-Geneidy (2015) encourages works that investigate the effects of transportation planning on different neighborhoods and groups for better understanding how equity is distributed.

In recent years, the importance of the role played by transport provision in distribution of goods and opportunities creates a paradigm shift in transportation planning (LUCAS, 2012). Grounded on this view, transportation infrastructure is a way to distribute people and resources in the urban environment. This rationality created an awareness to improve transportation plans to a more comprehensive approach to better integrate land use plans and incorporate such inequitable effects.

A suitable concept to frame such integration is the accessibility (BERTOLINI; CLERCQ; KAPOEN, 2005). Accessibility can be defined as the potential opportunities for interaction (HANSEN, 1959). About this concept, the literature points out: a considerable number of definitions and components (GEURS; VAN WEE, 2004); reviews of indicators (PÁEZ; SCOTT; MORENCY, 2012; VAN WEE, 2016); investigation about the changes in the access level over the years (FOTH et al., 2013); identification of the current consideration of such issues into transportation plans (BOISJOLY; EL-GENEIDY, 2017a) and how it is perceived by practitioners (BOISJOLY; EL-GENEIDY, 2017a) among other studies assessing cities of different size and transit infrastructure.

Accessibility measures are developed to inform, above all, the decision makers, about the number, quality, and availability of spatially distributed potential opportunities to be reached given a cost of travel. For that, different indicators have been calculated and their performance was compared by literature (NEUTENS et al., 2010;

WEE; GEURS, 2011). Especially, regarding equity analysis, Neutens et al. (2010) highlight the cumulative opportunities as a relevant metric to evaluate the distribution of opportunities across the territory. Other current studies refer, for instance, the Gini Index and Lorenz curve to evaluate the cumulative percentage of access of a specific group (DELBOSC; CURRIE, 2011; LUCAS; VAN WEE; MAAT, 2015).

Some recent works assess the current accessibility measures considered in transportation plans. Boisjoly; El-Geneidy (2017a) analyze plans in North America and highlight the weak consideration of such indicators. According to Boisjoly; El-Geneidy (2017b), the most part of practitioners are familiar with the concept and the half of them use accessibility metrics in their work. In the Brazilian reality, the Mobility Plans have the role of organizing and determining the future interventions on the transportation infrastructure in the major municipalities. The National Mobility Policy highlights the need of addressing social exclusion and inequalities, however, few plans present at least cost benefits analysis regarding the condition of low income population.

The main hurdles identified by practitioners are the lack of available data, practitioner's knowledge (BOISJOLY; EL-GENEIDY, 2017b), organizational barriers and lack of institutionalization of accessibility instruments (SILVA et al., 2017). Guidelines, frameworks and technical reports can bring useful and clear information about accessibility concepts and inequalities assessment to overcome the exiting gap between literature and planning practice. Specially those related to more comprehensive and innovative approaches for capturing the complexity in the opportunities distribution (LUCAS, 2012).

For addressing data analysis in a comprehensive way, new developments in the computational literature field present data mining techniques, which focus on extracting information from a large and complex dataset. This field encompasses techniques that relate ideas of knowledge acquisition, namely Machine Learning (ML) techniques. They are useful to explore high dimensional data for investigating two main hypotheses: (i) identify and describe hidden patterns in the dataset with unsupervised learning and (ii) predict and infer relation of values of continuous and categorical variables with supervised learning. The ML techniques model different data type, with or without spatial information. The algorithms can be adapted to consider geographical information in its formulation (BRUNSDON et al., 1996; RUß; KRUSE, 2011).

Many applications of ML may be found in different scholarly fields, such as: (i) biology, with animal behavior studies (JOSEPH et al., 2016), genome classification (REMITA et al., 2016) and microbiology ecology (MILLER-COLEMAN et al., 2012); (ii) medical, with specific symptoms as epileptic seizure prediction (SUBASI; KEVRIC; ABDULLAH CANBAZ, 2017), suicide occurrences prediction (RIBEIRO, 2017), classification of cancer cells (YAMAMOTO et al., 2017) and developments in medical imaging (WONG; WANG; WANG, 2017); (iv) remote sensing, especially for soil classification (HEUNG et al., 2016; FORKUOR et al., 2017), carbon mapping (MASCARO et al., 2014), topography mapping (SESTER, 2000) and analysis of city structures (WERDER et al, 2010); (v) interdisciplinary area with mosquitoes habitat detection through image classification (WIELAND et al., 2017) and in the forensic field to identify demographic characteristics based on hand dimensions (MIGUEL-HURTADO et al., 2016).

In transportation area, ML techniques are applied mainly to: (i) explore traffic and transit big data (FUSCO; COLOMBARONI; ISAENKO, 2016; MAHRSI et al., 2017); (ii) prediction of travel model choice (HAGENAUER; HELBICH, 2017; ZHU et al., 2017) and travel time (GAL et al., 2014); (iii) to quantify interdependence between land use and transport delivery (HU et al., 2016). Therefore, the most applications deal with high complexity data to better understand the object of study and extract knowledge from it. However, few applications are focused on better understanding inequalities and the condition of deprived groups.

To test the application of such techniques, recent works explore different aspects of social issues across the São Paulo territory describing different: (i) socioeconomic distribution (MARQUES, 2005); (ii) levels of segregation (FEITOSA; CÂMARA; MONTEIRO, 2007); (iii) degree of social exclusion (SPOSATI; MONTEIRO; 2017). However, few of them explore different level of transportation provision and land use distribution with income indicators.

1.1 Objectives

Grounded on ethics perspectives for delivering transportation infrastructure, Garcia (2016) proposes a deeper analysis in the understanding of decision problems and stated a problem-oriented approach for transportation planning. Four categories of problems

are summarized based on the differences in accessibility and mobility levels across space, modes, social groups and over time. Following this structure of problem-oriented analysis, this investigation focuses on discovering hidden patterns and acquire information about a specific spatial distribution of inequalities, mainly regarding the low-income population.

As main concern, this work aims at identifying and describing patterns hidden in the distribution of accessibility to different urban facilities and socioeconomic information using ML techniques to inform the decision making of transportation plans.

Two level of analysis are proposed: (i) an investigation about inequalities measures already considered in transportation plans is conducted. Built on literature and documents review, the current consideration of spatial inequalities measures in transportation plans is outlined. The objective is to feature the current state of Brazilian Mobility Plans consideration; (ii) discovering hidden information about income and accessibility data. Supervised learning is applied mainly to explore relations between the income, as the target variable, and the accessibility indicators, the explanatory variables. With the loads obtained as coefficient of the regression functions to estimates income, it is possible to determine the degree of influence of each variable in the income values. The hidden relation can be identified by testing the behavior of the target feature when the explanatory variable changes.

1.2 Research questions

The research is conducted by the main question: how spatial inequalities could be investigated by using ML techniques to inform the current transportation planning practice?

The ML analysis focus on exploring the São Paulo data. Two questions are stated: one with methodological focus, to explore the potentialities of the ML techniques, and the other, focused on investigating the phenomenon of inequalities distribution. For both aspects and for each technique, the hypothesis is presented for providing temporary affirmative sentence to be tested by the analysis (GIL, 2008).

Table 1 – Research questions and hypothesis of quantitative analysis

ML technique	Methodological aspects		Phenomenological aspects	
	Question	Hypothesis	Question	Hypothesis
Clustering (unsupervised)	How could spatial inequalities be explored using unsupervised techniques?	The clustering can capture distinguished condition of inequalities distribution	Has the low income low accessibility to all opportunities across São Paulo municipality?	The low-income population has low access to all facilities
Regression (supervised)	How could spatial inequalities be explored using supervised techniques?	The regression can estimate the relations between variables	What is the relation between income and other accessibility?	The income is influenced by all accessibilities

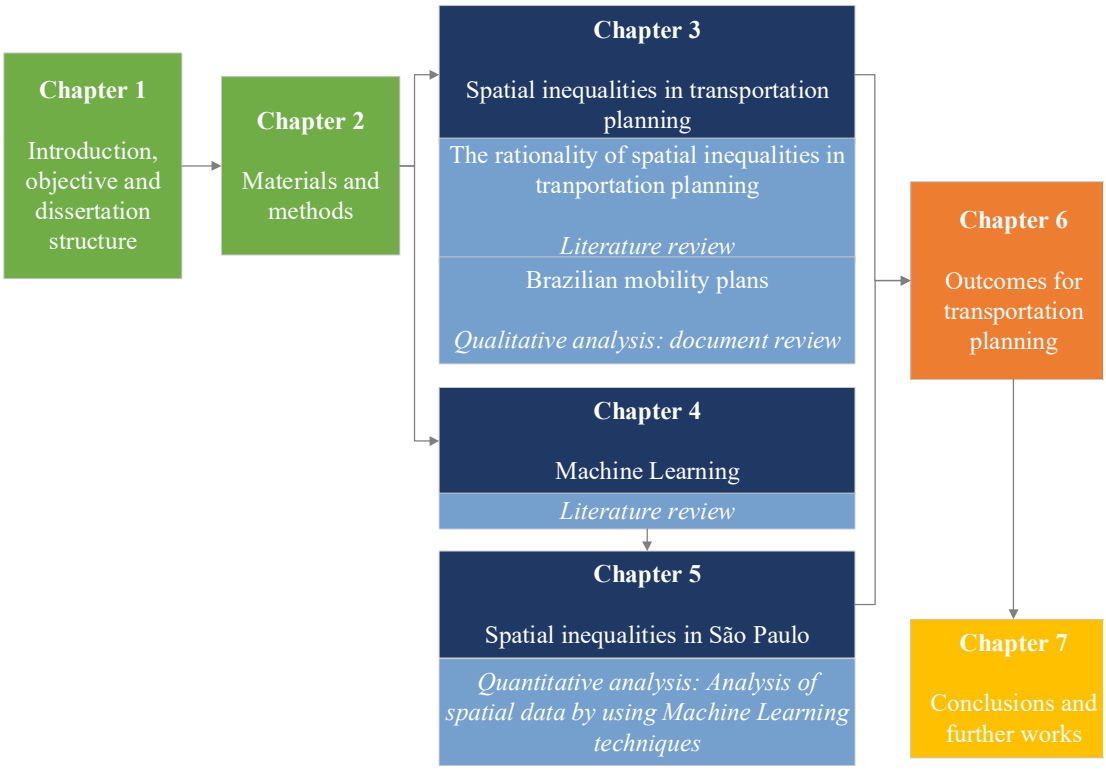
Source: the author

Finally, a framework is proposed for further application to evaluate spatial inequalities by using ML techniques in transportation planning practice. This work contributes to the literature discussing the implications of the using ML techniques for evaluating inequalities and proposing a framework to integrate such approach into transportation plans.

1.3 Dissertation outline

This work is presented in nine chapters. For achieving the proposal, the literature review and the qualitative analysis provide a background of the current international and national developments in the field. To test the potentialities of ML techniques, spatial inequalities of opportunities in the São Paulo municipality were explored. Finally, the outcomes were summarized for ML application in the context of transportation planning (Figure 1).

Figure 1 – Dissertation outline



Source: the author

Each chapter comprises the following contents:

Chapter 2. Materials and methods

This chapter presents two main parts of the procedures and materials used in the qualitative and quantitative analysis. The first part comprises the Mobility Plans review with information about the plans analysed. The second part presents the Census data used to feature the socioeconomic condition of the population and the accessibility indicators. The data analysis involves two main sections: (i) the unsupervised with dimensionality reduction and clustering; (ii) the supervised with linear regression.

Chapter 3. Spatial inequalities in transportation planning

This chapter provides an overview of the consideration of spatial inequalities in transportation plans. It also presents the Brazilian current legislation and contents of the Mobility Plans review.

Chapter 4. Machine learning

This chapter presents a literature review of machine learning techniques. The first part outlines the main concepts of unsupervised learning, involving dimensionality reduction and clustering techniques. The second part presents the review of supervised learning with classification, regression, and related algorithms. The last part presents a review about the adaptation of ML algorithms for spatial analysis and establishes the scope of the techniques applied in this work.

Chapter 5. Spatial inequalities in São Paulo

This chapter analyzed the spatial data of the São Paulo municipality by using ML techniques. First, the indicators of several opportunities to be reached within a travel time are presented. The measures of cumulative opportunities of hospitals, health centers, culture facilities, sports centers, public and private schools from 2000 and 2010 are presented. The Principal Component Analysis and K-Means clustering algorithm are applied to identify the heterogeneity in the dataset, especially different accessibility levels for the low-income population. Then the regression analysis of income reveals the dependence with accessibility measures for identifying hidden relations in the dataset.

Chapter 6. Outcomes for transportation planning

This chapter aims at summarizing the contributions of using ML techniques in transportation practice. The discussion based on the ethics questions are firstly presented. Then, the potentialities and constraints of application of ML techniques to explore the inequitable distribution of opportunities are depicted. Besides that, a checklist for the implementing such techniques is presented. A framework with the general stages of transportation decision making and respective contributions of ML is outlined.

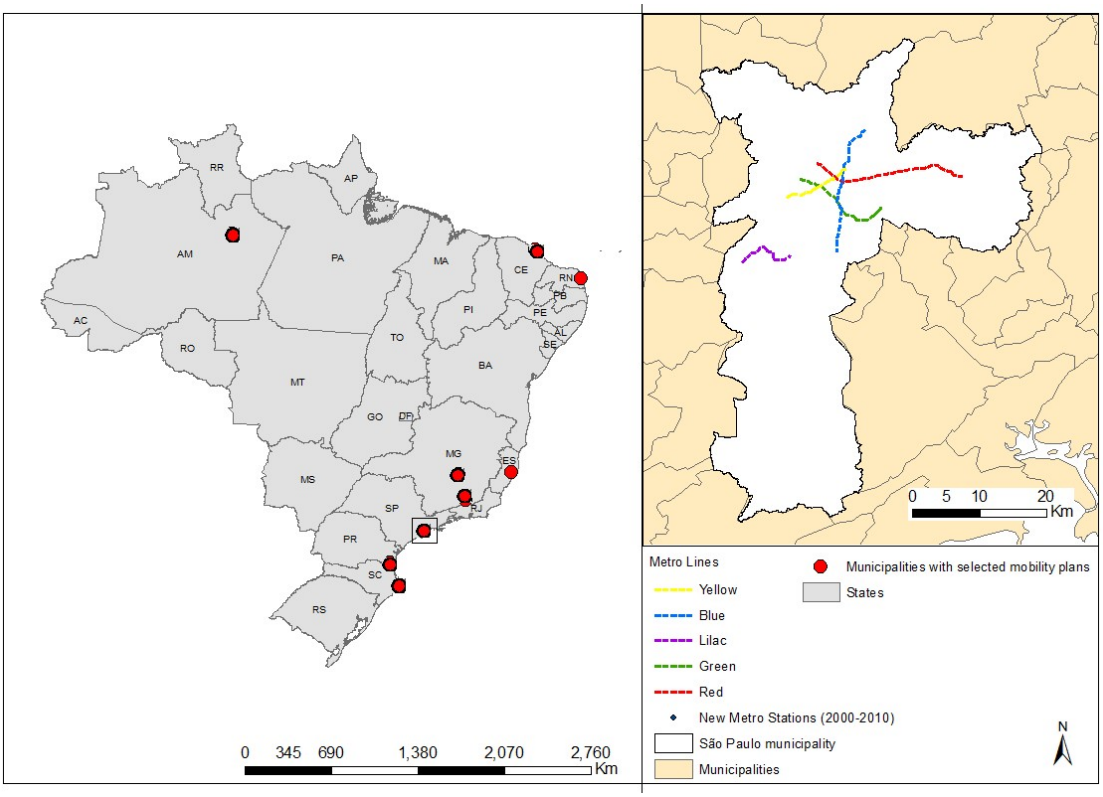
Chapter 7. Conclusions and further works

Finally, the trajectory of the study is resumed, the main conclusions are presented, and the further works outlined.

2 Materials and methods

This investigation comprises qualitative and quantitative methods. To feature the current state of spatial inequalities consideration in transportation plans in Brazil, nine mobility plans of different cities were reviewed (Figure 2). In the quantitative analysis, the data from the years of 2000 and 2010 of the São Paulo municipality were analyzed by using ML techniques. The database is composed by socioeconomic and accessibility indicators. The main changes in the transportation network between the 10 years analyzed can be seen in Figure 2. The following subsections present the procedures and data used.

Figure 2 - Municipalities with analyzed mobility plans and the São Paulo study area



Source: the author

2.1 Mobility Plans review

Brazilian Federal Law 12.587/2012 established the National Policy of Urban Mobility. It requires a mobility plan for the municipalities with more than 20,000

inhabitants. To support the decision making, the Ministry of Cities elaborated a brochure with principles and guidelines regarding the mobility plans development by the municipalities (SEMOB; MINISTÉRIO DAS CIDADES, 2015). The free online documents consulted are from cities of different size and Brazilian regions (Table 2).

Table 2 – Mobility Plans reviewed

Name	City	Year of publication	City inhabitants (IBGE - 2017)	Number of pages	Responsable
Sustainable Urban Mobility Plan of the great Florianópolis	Florianópolis (SC)	2015	285,838	263	Logit Engenharia Consultiva/ Strategy/ Machado Meyer Sendacz Opice Advogados
Urban Mobility Plan of Belo Horizonte	Belo Horizonte (MG)	2012	2,523,794	144	Logit/ BH Trans/ Prefeitura de Belo Horizonte
Fortaleza Mobility Plan	Fortaleza (CE)	2015	2,627,482	116	Prefeitura de Fortaleza/ Instituto de Planejamento de Fortaleza (IPLANFOR)
Urban Mobility Plan of Manaus	Manaus (AM)	2015	2,130,264	Vol I: 312/Vol II: 116	Oficina Engenheiros e consultores associados Ltda
São Paulo Mobility Plan	São Paulo (SP)	2015	12,106,920	201	SPTrans/ CET
Urban Mobility Plan of Juiz de Fora	Juiz de Fora (MG)	2016	563,769	377	Secretaria de Transportes e Trânsito - SETTRA/ Prefeitura Municipal de Juiz de Fora/ ML consultoria planejamento e gestão
Urban Mobility Plan of Joinville	Joinville (SC)	2016	577,077	Vol I - 164/Vol II -183	Fundação instituto de pesquisa e planejamento para o desenvolvimento sustentável de Joinville/ WRI Brasil/ UFSC
Mobility Plan of the municipality of Aracruz	Aracruz (ES)	2014	98,393	83	Logit/ Prefeitura de Aracruz
Master Plan of Urban Mobility of Natal	Natal (RN)	2015	885,180	Executive summary – 28	Prefeitura de Natal/ Tectran -Sytra group

Source: the author

The review was guided by keywords such as “acessibilidade” (accessibility), “exclusão social” (social exclusion); “desigualdades” (inequalities). They were selected considering the guidelines for the plans developments for identifying the common words

used in the planning practice (SEMOB; MINISTÉRIO DAS CIDADES, 2015). The results are presented in Chapter 4.

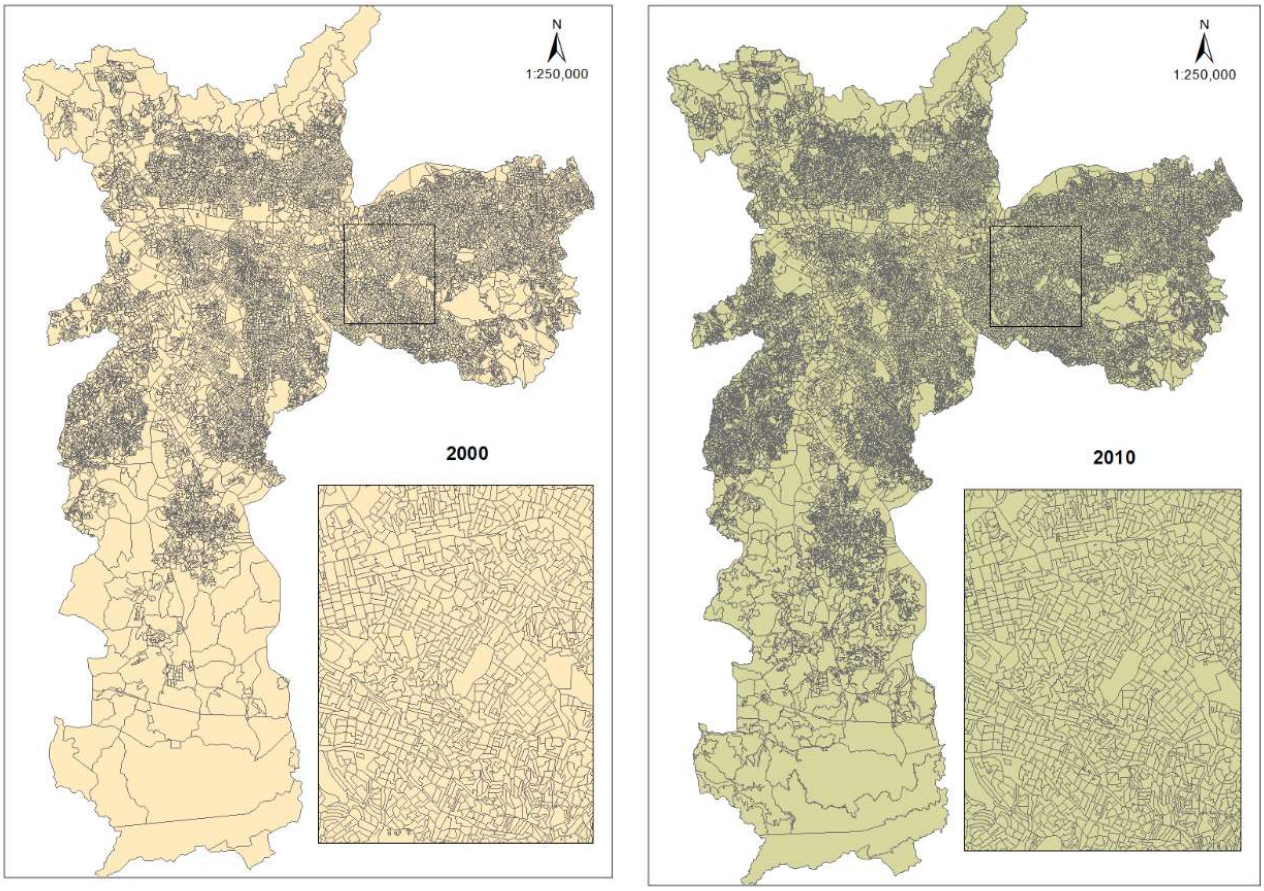
2.2 Spatial inequalities analysis

This work evaluates spatial inequalities by using ML techniques. First, the indicators of income and accessibility were calculated, and the data is then analyzed using supervised and unsupervised techniques.

2.2.1 Census data

For composing the socioeconomic features of the groups, the census data was taken as reference. The Brazilian Census (IBGE, 2003, 2011) is divided in two main surveys: the Universe and Sample. The first presents more basic variables about income, gender and age of households. The level of aggregation is on census tracts, the smallest division of territory with census data. The Sample comprises more detailed variables and the data are grouped into larger polygons to guarantee the confidentiality of the interviewed. In this work, Universe results and the division of census tracts are used herein (Figure 3). Both measures, accessibility and income indicators, were calculated for each census tracts. Further works can explore the other official divisions of the São Paulo territory to capture scale effects in accessibility and income indicators (See Section 6.4).

Figure 3 - Census tracts 2000 (left) and 2010 (right)

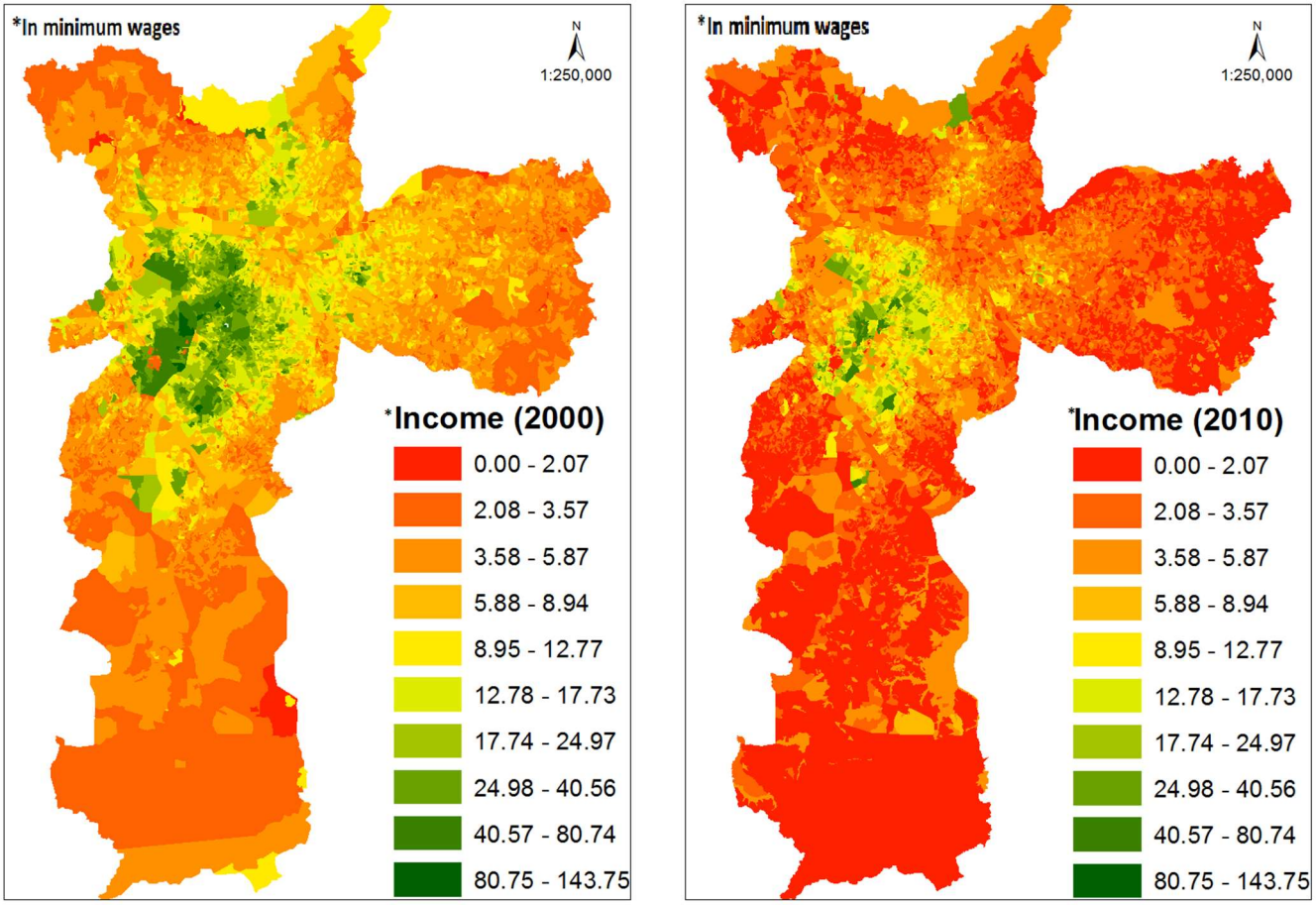


Source: the author

The census variable selected was the average monthly income of the householder, depicted in Figure 4. For excluding economic inflation from the analysis, the value was divided by the minimum wages (m.w.) (in 2000 it was R$ 151,00 and in 2010, R$ 510,00). Even after normalizing the income by m.w., the purchasing power may have changed over the years. In this context, we decided to keep this approach for quantifying income, rather simplistic, and leave the adoption of a more enhanced normalization for income for future works, which may require a complex discussion.

Another experiment with clusters and income were presented in Siqueira-Gay; Giannotti; Sester, 2017. The income variable was transformed into categorical values of low, medium and high income. In the present work, it was preferred the value of income in m.w. to test the clustering assignment with this continuous variable.

Figure 4 – Spatial patterns and values of income (m.w.) in 2000 and 2010 with the division method used was natural breaks



Source: modified from Siqueira-Gay; Giannotti; Sester (2017)

In the 2000 database, there are 13278 census instances and in 2010, 18953. The difference in the number of census tracts is due to the changes in the urban area and population growth over the years. Because of these changes, the interviewer changes his/her reachable area for interviewing the population. The missing values represent less than 1% of all the data in 2000 and 3% in 2010.

Table 3 – Variable of 2000 and 2010 Census used in the analysis

Census Variables		Indicators
Income	Average monthly income of householder	Average number of minimum wages earned by the householder

Source: the author

2.2.2 Accessibility measures

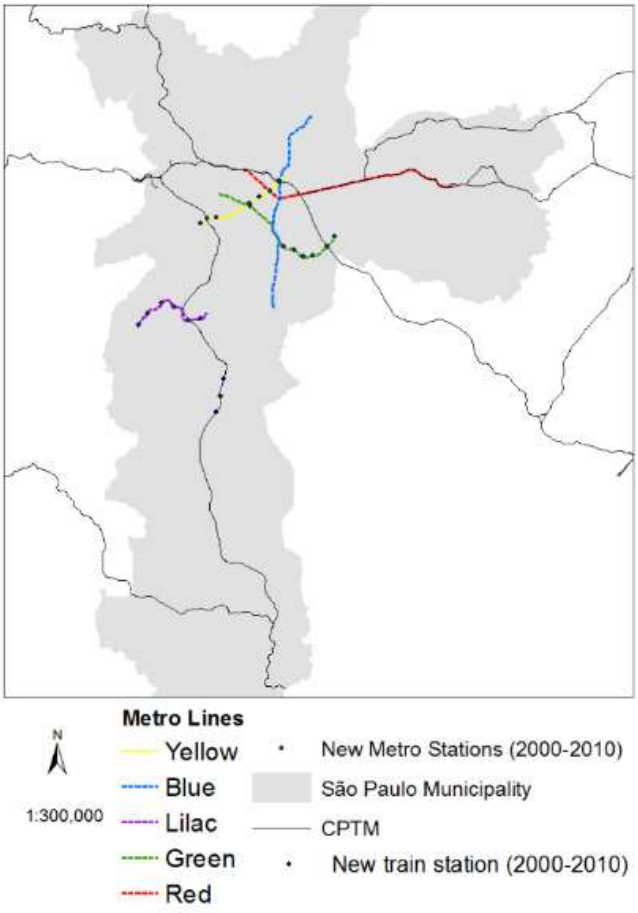
Based on the configurations of metro network (LOGIODICE, 2017; TOMASIELLO, 2016), in the two-time periods, 2000 and 2010, the inference of transit travel time was used to estimate the travel cost for the accessibility indicators. The network was built firstly for 2010 (TOMASIELLO, 2016) and regressed with the increased impedance of the travel time in the new metro lines extensions. The network changes comprise the new metro and train stations constructed after 2000 (Table 4). The main differences between the two years are new stations in yellow, lilac and green metro lines, available in 2010, and new train stations shown in Figure 5. For the accessibility indicators, the same urban equipment (e.g. hospitals and others) was used in accessibility measures for both years, therefore, the changes in accessibility levels reflect the changes only in the transportation network.

Table 4 – Metro and train stations implemented after 2000

New metro stations	Line	Year
Autódromo	CPTM (Train) Line 9-Esmeralda	2007
Primavera-Interlagos		2008
Grajaú		2008
Santos-Imigrantes	Metro - Linha Verde	2006
Chácara Klabin		2006
Alto do Ipiranga		2007
Sacomã		2010
Vila Prudente		2010
Tamanduateí		2010
Capão Redondo-Largo Treze	Metro - Linha lilás	2002 (reduced operation)-2008 (complete operation)
Adolfo Pinheiro		2014
Paulista-Faria Lima	Metro - Linha amarela	2010
Butantã		2011
Pinheiros		2011
República		2011
Luz		2011
Fradique Coutinho		2014

Source: the author

Figure 5 – Metro and train transit lines in the São Paulo municipality: yellow and lilac lines were built after 2000



Source: from Siqueira-Gay; Giannotti; Sester (2017)

A vast majority uses accessibility measures to compare equity, service provision and distribution effects (VAN WEE, 2016). Especially for equity analysis, Neutens et al. (2010) highlight the cumulative opportunities as a relevant metric to evaluating the distribution of opportunities across the territory. This indicator comprises two main dimensions: the number and availability of facilities and the cost of travel, which can be travel time or distance. In this work, the cumulative opportunities (2.1) aims at evaluating the potential number of urban equipment to be reached given a travel time (NEUTENS et al., 2010; PÁEZ; SCOTT; MORENCY, 2012):

$$A_{ik}^{p} = \sum_{j} W_{jk} I(c_{ij} \leq \gamma_{i}^{p}) \tag{2.1}$$

Where:

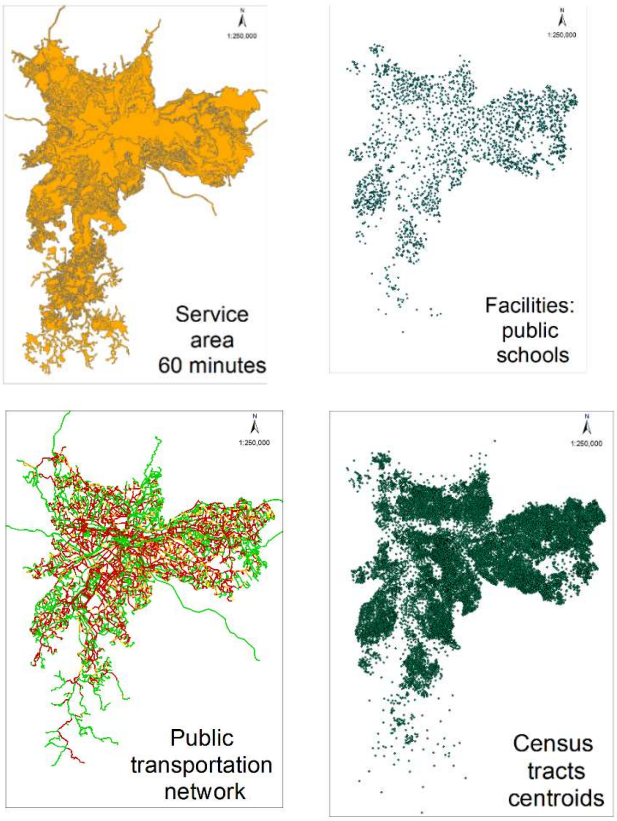
W_{jk} = facility of type k at location j

c_{ij}= cost of travel, here the travel time measured in the public transportation network from the location "i" to reference "j".

γ_i^p= threshold value

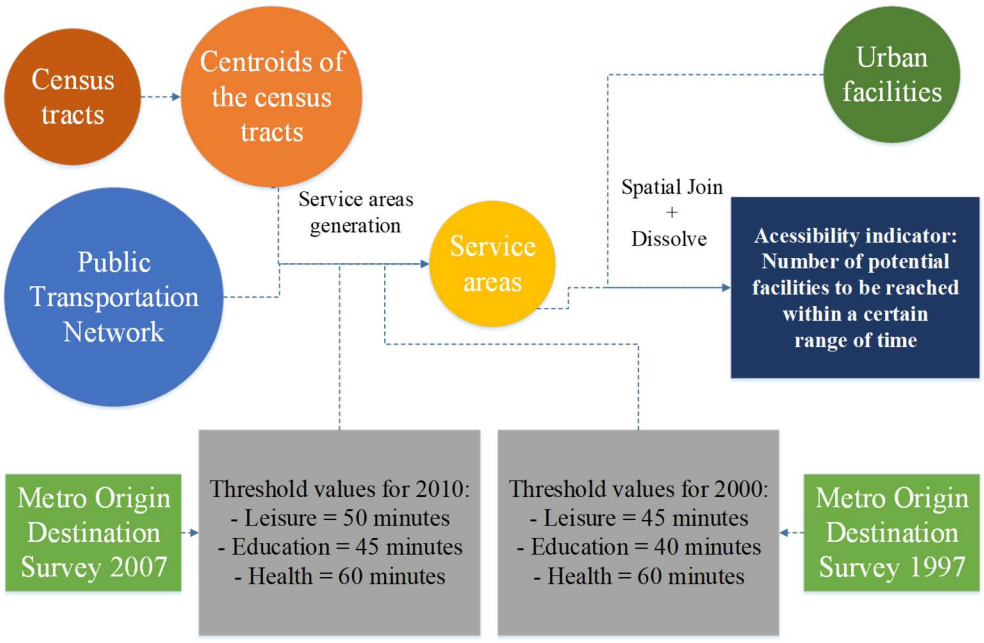
The data used for calculating the indicators were: (i) the network with public transport travel time to calculate the travel time and generate the service areas, (ii) the centroids of census tracts to be the center of services areas and (iii) the facilities (Figure 6). First, the service area is generated based on a given travel time with the network and then, the number of urban facilities in each service area is counted. Then, the sum of the number of facilities to be reached is associated to the census polygon of the respective centroid (Figure 7). The time at which people are willing to travel depends on the trip purpose. The threshold value is calculated based on the guideline of the Department for Transport Business Plan (2012) from the UK and represents the median of all travel using public transportation with specific purpose: education for accessibility to public and private schools; health for accessibility to hospitals and health centers and; leisure for accessibility to cultural facilities and sport centers. In 2000 and 2010 the values were similar, and the decision was to use the larger one to maintain the most conservative threshold. The steps constructing the indicator are summarized in Figure 7 and the final measures in Table 5. For calculating the travel time with the network, it was considered one working day (Wednesday) at the peak hour (8 a.m.) and the urban facilities are the same for both years. The general differences between 2000 and 2010 are only resulted by the simulated metro changes.

Figure 6 - Data used to calculate the accessibility measures



Source: the author

Figure 7 - Main steps of accessibility measures



Calculation of the thresholds values for each travel purpose (education, health and leisure)

Source: adapted from Siqueira-Gay; Giannotti; Sester (2017)

Table 5 - Accessibility measures

Type of accessibility measure	Urban facilities	Indicator
Cumulative opportunities	Hospitals	Number of hospitals to be reached within 60 minutes of travel time by transit
	Health centers	Number of health centers to be reached within 60 minutes of travel time by transit
	Public schools	Number of public schools to be reached within 45 minutes of travel time by transit
	Private schools	Number of private schools to be reached within 45 minutes of travel time by transit
	Sports centers	Number of sports centers to be reached within 50 minutes of travel time by transit
	Museums and public libraries	Number of museums and libraries to be reached within 50 minutes of travel time by transit

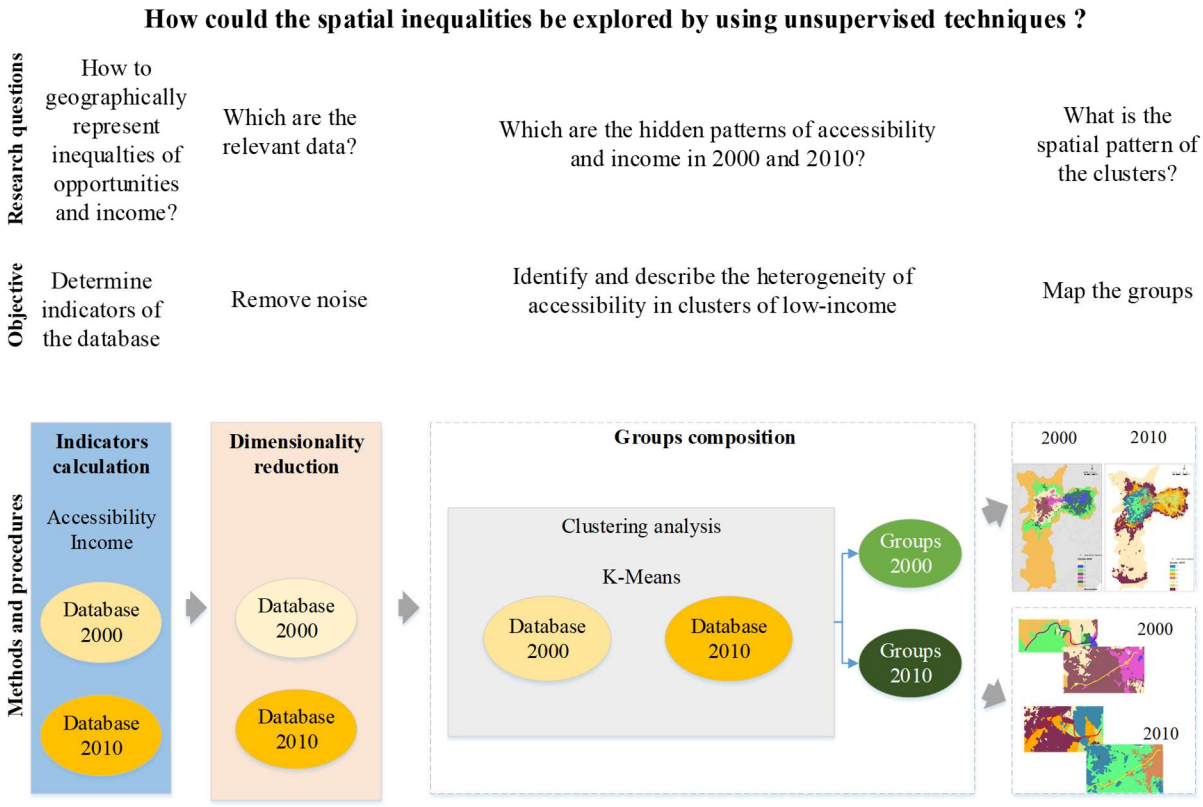
Source: from Siqueira-Gay; Giannotti; Sester (2017)

2.2.3 Data analysis

The analyses were divided in two groups, unsupervised and supervised. The algorithms selection was based on ML scholarly network and those implemented on Weka software. The analysis aims at exploring different algorithms already identified in the literature instead of implementing our own code.

For the unsupervised analysis, the objective is to explore hidden patterns and visualize the heterogeneity of the groups formed by the clustering algorithms. The details of the methodology are depicted in Figure 8.

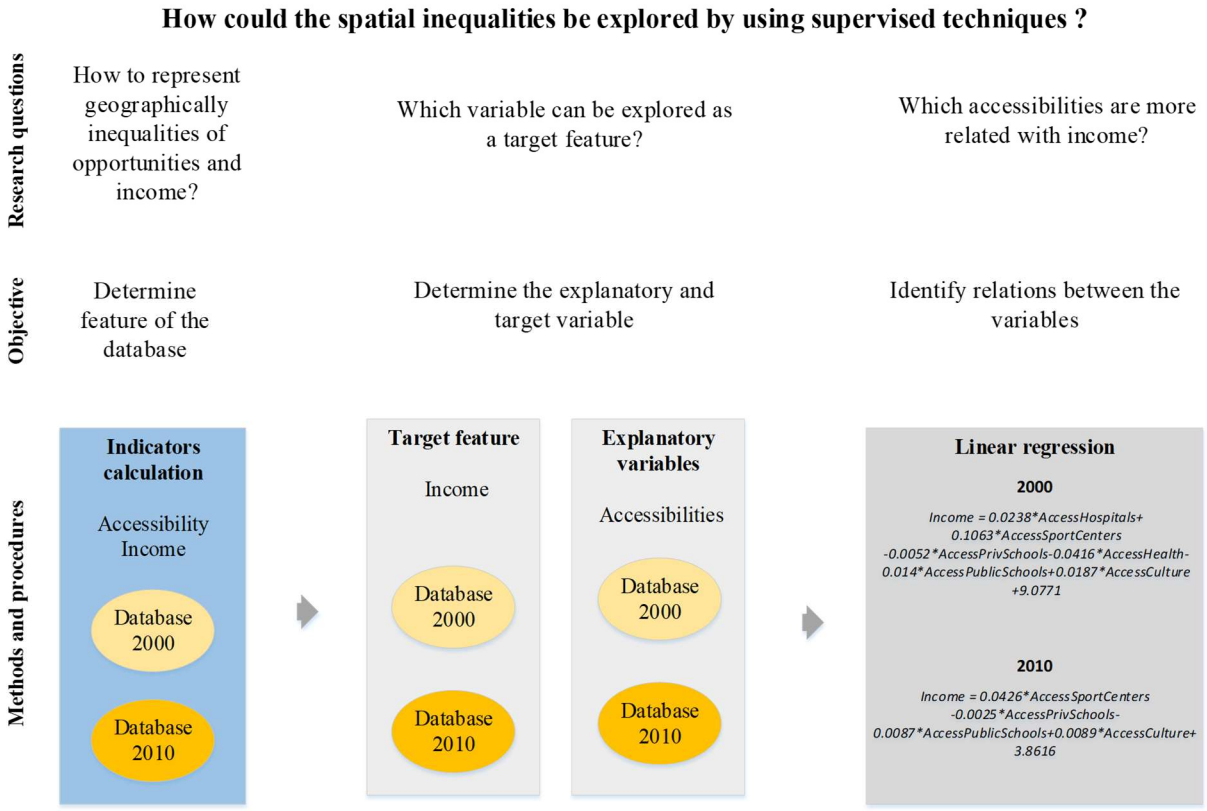
Figure 8 – Research framework of unsupervised analysis



Source: adapted from Siqueira-Gay; Giannotti; Sester (2017)

For the supervised analysis, the focus is describing the relations between the variables in the dataset. The first step is to compose the database and then to choose the desired target variable. The rest of the dataset is considered as explanatory variables. In this analysis, the income was chosen as a reference and the dependence between the socioeconomic data and accessibility can be explored (Figure 9). The estimation of income values based on accessibility could be useful for determining the influence of each accessibility change in the income level. For the planners practice, this estimation can bring insights about the relation between information of two different natures, socioeconomic and transportation data and how one influence in the other.

Figure 9 - Research framework of supervised analysis



Source: the author

The ArcGis 10.5 software was used for calculating the accessibility measure and graphic representation of geographical information. The explorer of Weka 3.8 (FRANK; HALL, 2016) was employed for preprocessing, dimensionality reduction, clustering and regression.

3 Spatial inequalities in transportation planning

This chapter aims at presenting a brief review about the literature of inequalities in transportation planning. Besides that, the qualitative analysis of nine Brazilian mobility plans is presented to feature how is the current consideration of spatial inequalities measures in Brazilian transportation planning practice.

3.1 The rationality of spatial inequalities in transportation planning

In the evolution of growing cities, the transport infrastructure plays an important role connecting regions and people. Thus, the evolution of transportation planning rationality presents a paradigm shift for strengthening the integration with other urban policies as a mean to address issues related to the urban complexity (BANISTER, 2008).

Some authors already identified, especially after 1970's, transportation plans considering some sustainability issues and incorporating the concept of "3Es" (Environment, Economic and Equity) (MANAUGH; BADAMI; EL-GENEIDY, 2015). Bertolini; Clercq; Kapoen (2005) argue that to achieve the sustainable development, i.e. the balance between economic, social and environmental perspectives, the integration between land use and transportation plans should consider economic, social and environmental inequalities. In short, the plan's outcomes related to sustainability can be "tangible", such as: (i) reducing congestion and GHG emissions; (ii) improving the air quality and safety; (iii) increasing coverage and use of transit as well as cycling and walking. Others outcomes, "less tangible" are related to social equity or exclusion (MANAUGH; BADAMI; EL-GENEIDY, 2015). Given this, to promote the sustainable development in the urban environment, the literature highlights the importance of value inequalities in distributions of opportunities into different social groups (EL-GENEIDY et al., 2015; VAN WEE, 2016).

Both land use and transportation planning play a role in distributing goods, resources and in connecting the territory (BANISTER, 2002). In this sense, the accessibility concept can be useful to frame the integration between land use and

transportation plans (BERTOLINI; CLERCQ; KAPOEN, 2005). The definition of accessibility concept across the literature (GEURS; VAN WEE, 2004). In a simple way, it can be simply defined as the potential opportunities for interaction (HANSEN, 1959).

The indicator comprises two main aspects: (i) the travel cost, which depends on the transportation infrastructure and can be measured as time or distance; (ii) the distribution of opportunities as well as its quantity and quality (PÁEZ; SCOTT; MORENCY, 2012). Therefore, as more urban equipment, jobs, parks or, generally, opportunities are available, the higher will be the accessibility level. Also, if the opportunities stay fixed but it is easier to reach them, the higher will be the accessibility.

The analysis of inequalities can explore relations between accessibilities measures and other indicators, such as segregation and social vulnerability, or consider the travel behavior of a deprived group in the indicator formulation. The first approach encompasses techniques such as statistic distribution measures as the range, variance, coefficient of variation, relative mean deviation, logarithmic variance, the variance of logarithms, Gini, Theil's entropy or welfare as Atkinson and Kolm measures (RAMJERDI, 2006). One approach broadly explored is the Gini Index formulation (NEUTENS et al., 2010; WEE; GEURS, 2011; LUCAS; VAN WEE; MAAT, 2015; GUZMAN; OVIEDO; RIVERA, 2017). The distribution of the accessibility level of a group is evaluated with an equality curve and the effects on transportation policies can be analyzed comparatively.

Although the literature presents robust assessments to evaluate accessibility and inequalities distributions, such measures are poorly addressed by transportation plans, even in a developed context (BOISJOLY; EL-GENEIDY, 2017a). Whilst the most part of practitioners are familiar with the concept, the half of them use accessibility metrics in their work (BOISJOLY; EL-GENEIDY, 2017b). Some barriers are identified such as the lack of available data (BOISJOLY; EL-GENEIDY, 2017b), practitioner's knowledge (BOISJOLY; EL-GENEIDY, 2017b), organizational barriers (SILVA et al., 2017) and lack of institutionalization of accessibility instruments (SILVA et al., 2017). Guidelines, frameworks and technical reports can bring useful and clear information about accessibility concepts and inequalities assessment to overcome the exiting gap between literature and planning practice.

3.2. Brazilian Mobility Plans

Some mandatory initiatives require the consideration of social issues into transportation plans. Especially in 2002-2003, the UK government published the Social Exclusion Unit (SEU) study about transport and social exclusion, which highlight the importance of identifying the relation between transport disadvantages, social groups and places with the potential attention of transportation policies (LUCAS, 2012). Since 2006, it is mandatory for the UK transport authorities to undertake strategic and local accessibility assessments as part of their local transport plans.

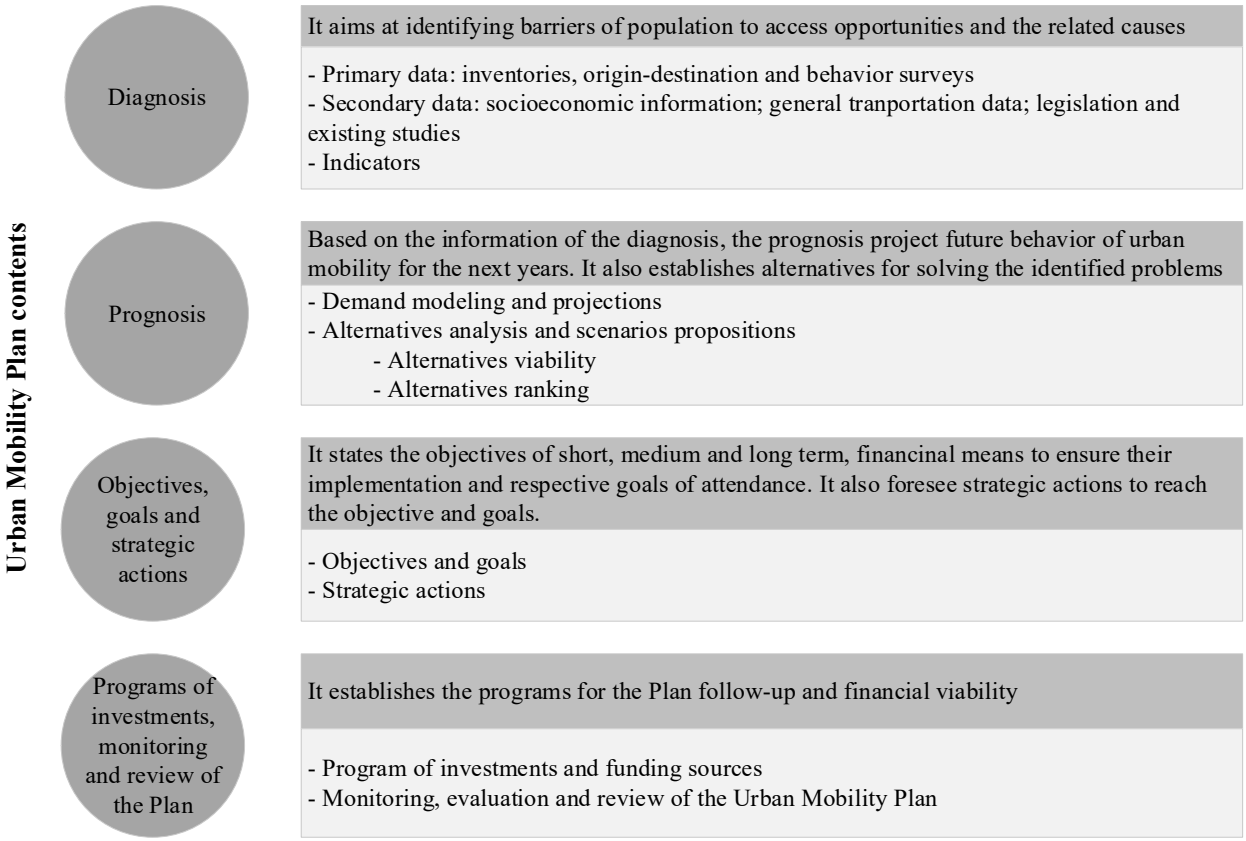
In Brazil, the National Policy of Urban Mobility (Federal Law 12.587/2012) is one instrument to support the urban policies, which aims at integrating different transport modes and improvement of accessibility, people mobility and cargo in the municipality's territory. Two main principles related to inequalities are established by the law: the fair distribution of benefits and onus resulting from the use of different modes and services (Art 5º - VII) and equity in the use of the public space circulation and places (Art 5º - VIII). The main objectives of the National Policy are to reduce the inequalities and promote the social inclusion; provide access to the basic services and social equipment; improve the accessibility and mobility condition of the population that lives in cities; promote the sustainable development with the mitigation of environmental and social costs resulted from the commuting of people and cargo transportation; consolidate the democratic urban management as an instrument and promote the continuous improvement of urban mobility. It also requires the Mobility Plan compatible with the master and other urban plans. The plan document must comprise the principles, objective and guidelines of the Federal Law. It must present (Art 24):

- Services of public transportation;
- Circulation in the roads;
- Infrastructure of the urban mobility system;
- Accessibility to people with disabilities and restriction of mobility;
- Integration of public transport modes with private and non-private modes;
- Operation of the cargo transportation in the road infrastructure;
- Traffic generating center;

- Public and private parking areas;
- Areas and schedule of restrict access or controlled circulation;
- Financing mechanisms and instruments of public transportation and urban mobility infrastructure;
- Systematic evaluation, periodic review and updating of the Urban Mobility Plan in maximum ten years.

The federal law requires to municipalities with more than 20,000 inhabitants to elaborate an Urban Mobility Plan. Given this, the City Ministry presented a guide for mobility plans elaborations (SEMOB; MINISTÉRIO DAS CIDADES, 2015). The document introduces the main concepts and guidelines for the plans elaboration by the municipalities. It also brings concepts related to the socioeconomic aspects of urban mobility and social exclusion, as income, age, and gender restrictions. Case studies aspects of the Urban Mobility Plan of Belo Horizonte and Curitiba are presented. The brochure indicates the minimum contents of an Urban Mobility Plan as depicts in Figure 10.

Figure 10 – Urban Mobility Plans contents



Source: adapted from SEMOB; Ministério das Cidades. (2015)

In the analyses of Brazilian mobility plans, nine cities were taken as reference. The review aims at identifying the data, measures, indicators used to assess spatial inequalities and social exclusion. The main results are shown in Table 6.

Table 6 – Main findings in the analyzed documents

Name	City	Objective	Measures of spatial inequalities, accessibility and equity
Sustainable Urban Mobility Plan of the great Florianópolis (2015)	Florianópolis (SC)	The plan aims at supporting the conception and implementation of urban mobility solutions in the region (p.25) by providing technical recommendations for future decision making (p.38)	It presents the socioeconomic impacts quantified by using economic gains and loss in the number of travels, time of travel, cost and externalities (p.39). In the alternatives assessment, the social impact and social inclusion were ranked and quantified by using the benefits for the class which earn up to 2 minimum wages in relation of the total (p.102).
Urban Mobility Plan of Belo Horizonte (2012)	Belo Horizonte (MG)	The plan aims at proposing physical and operational interventions to maximize the benefits for all society and exploring the potentialities of each component of the mobility system (p.13)	It presents analysis of benefits of the economic aspects generated for the society in each alternatives assessment (p.111). No measure is presented for low-income population.
Fortaleza Mobility Plan (2015)	Fortaleza (CE)	The plan aims at articulating all community in a participatory way to reduce the social inequalities, ensuring accessibility, optimizing the movement of people and assets (p.100).	No measures of spatial inequalities are presented.
Urban Mobility Plan of Manaus (2015)	Manaus (AM)	The plan aims at attending the need for population and urban development, considering the current situation e and challenges foreseen in the future of the city. (p. 12 Vol. I)	It presents accessibility[1] index of road network: balance of the number of intersections, blocks.
São Paulo Mobility Plan (2015)	São Paulo (SP)	The plan aims at: promoting universal accessibility on public sidewalks and roads; promoting accessibility to components of municipal urban mobility systems; optimizing the use of the road system; implementing appropriate environment to the movement of the active modes; improving freights logistics; consolidating the democratic management in improving the urban mobility; reducing the number of accidents and deaths caused by traffic; reducing the average travel time; increasing the use of collective mode; promoting the use of active modes; reducing atmospheric emissions; contributing to the policy of reducing social inequalities; making the macro accessibility in the city more homogeneous (p.54)	No measures of spatial inequalities are presented.

[1] Term used by the plan presents a different meaning from this work that considers accessibility indicator as the number opportunities to be reached at a given travel time.

Urban Mobility Plan of Juiz de Fora (2016)	Juiz de Fora (MG)	The plan aims at thinking and proposing how it will be the displacement of people and goods in the city. It guarantees the rights of all, favoring collective transportation and non-motorized modes (p.6)	No measures of spatial inequalities are presented.
Urban Mobility Plan of Joinville (2016)	Joinville (SC)	The plan aims at establishing strategies and action regarding sustainable mobility in the city (p.21)	No measures of spatial inequalities are presented.
Mobility Plan of the municipality of Aracruz (2014)	Aracruz (ES)	The plan aims at: providing a broad and democratic access to urban space, prioritizing collective and non-motorized means of transport, inclusive and sustainable; contributing to the reduction of inequalities and to the social inclusion; promoting access to basic services and social facilities; providing improvement of urban conditions with regard to accessibility and mobility; promoting sustainable development with cost mitigation environmental and socioeconomic aspects of the movement of people and; consolidating democratic management as an instrument to guarantee continuous improvement of urban mobility. (p.14)	It establishes the demand for public transportation but without considering accessibility, opportunities or income inequalities
Master Plan of Urban Mobility of Natal (2015)	Natal (RN)	The plan aims at developing short, medium and long-term proposals and action plans by the year 2025. (p.6)	It presents maps of income, demographic density, urban equipment. In the baseline, it presents the mobility rate (number of trips) for the low-income class (p.62). It presents maps with number of educational establishments, number of health facilities and number of establishments for other purposes per census tract.

Source: the author

The metrics presented are related to the estimation of future trends. The plans of Fortaleza, São Paulo, Juiz de Fora, Joinville do not present metrics related to spatial inequalities. Although the plans of Belo Horizonte, Manaus and Aracruz present measures and maps (Aracruz), none of them are related to spatial inequalities of a deprived group or accessibilities. The plan of Florianópolis presents the positive socioeconomic impacts as a proxy for the benefits of the alternatives assessment. Notwithstanding, there is no detail about the methodology used to determine the proportional benefits for the low-income population, there is only about the Analytic Hierarchical Process used to rank the alternatives. There is no spatial reference or geographical information about opportunities to be reached.

Manaugh; Badami; El-Geneidy (2015) analyzed the consideration of social equity in transportation plans in North America. Although the clear advance stage in relation the Brazilian reality, the analyzed plans also do not present appropriate measures for assessing the social equity goals. The analysis of stratified groups and accessibility measures are identified in some plans of Canada and USA. Whilst the National Policy and Brazilian plans stress the terms of "social exclusion/inclusion" as an important issue to be considered in transportation development, few comprehensive assessments are shown in the plans analyzed.

Especially considering the spatial dimension of opportunities and the interaction between land use and transportation plans, Boisjoly; El-Geneidy (2017b) performed the analysis of accessibility objective and indicators consideration in transportation plans. They took as a reference a comprehensive database of 32 metropolitan transport plans from North America, Europe, Australia and Asia. Although there is a trend in integrating accessibility objective, few plans have accessibility indicators. None of the analyzed Brazilian plans present accessibility measures to quantify the opportunities for interaction to be reached. Towards more robust plans and comprehensive accessibility assessment, Boisjoly; El-Geneidy (2017b) indicates to planners define clear accessibility goals with a distinction between accessibility and mobility.

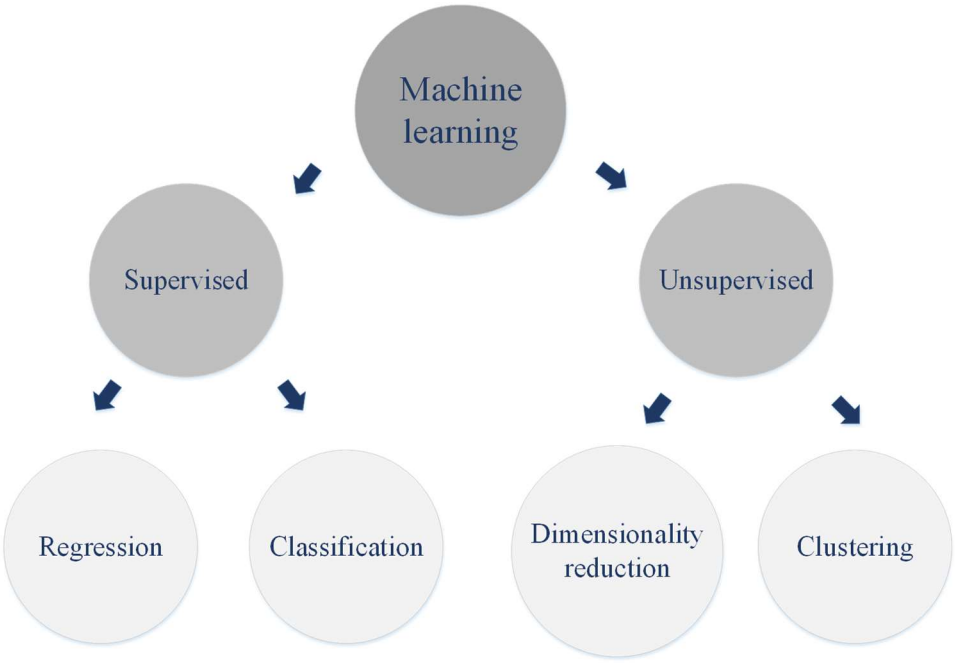
4 Machine learning

Data mining aims at discovering hidden patterns in the dataset. The word “mining” is related to the process of extracting minerals from the ground. In a comparative sense, the field of high complexity data analysis presents the same concept but related to gathering useful information from a dataset. ML, an especial class of data mining techniques, is based on automatic knowledge acquisition. Therefore, the general idea is that machine (computers) learning (acquiring knowledge) techniques serve to inform the undiscovered patterns and information in a complex dataset (WITTEN; FRANK; HALL, 2011).

Data mining represents a complementary approach to conventional statistics. Data analysis using the classical descriptive and inference character focuses on modeling data with information about its distribution (e.g. Poisson, Gaussian) and estimates the parameters. In contrast, the main goal of ML techniques is predicting and identifying hidden relations in the dataset, with or without a target variable. While the traditional statistic aims at validating a hypothesis, which is accepted/rejected depending on how consistent the measured observations are, ML techniques extract knowledge directly from the training data set, without necessarily assuming the previous function of data distribution. Given the potentialities of revealing heterogeneities in the dataset as well as relations between the variables, ML represents an attractive potential to deal with high complexity data (JOSEPH et al., 2016).

ML techniques can be divided into unsupervised and supervised learning techniques (Figure 11). Unsupervised learning works without a target variable, discovering hidden patterns in the dataset (HARRINGTON, 2016). The most well-known techniques are dimensionality reduction and clustering (JOSEPH et al., 2016). In contrast, supervised learning involves prediction of a target feature. It can be related to continuous variables, thus called regression, or to discrete values, named classification (Figure 11). Especially for the regression algorithms, the relations between the variables can be defined in a way to determine the degree of influence of each variable based on the function coefficients.

Figure 11 – Machine learning techniques



Source: the author

The next sections focus on the most traditional and consolidated ML algorithms. Recent studies propose new algorithms and approaches such as deep and reinforcement learning. For further discussion, some definitions are introduced to guide concepts mentioned hereafter. “Instance” is the input information to be analyzed. Each instance is considered as a vector with sample measures. The attributes are features associated with the class and its values.

4.1 Unsupervised learning

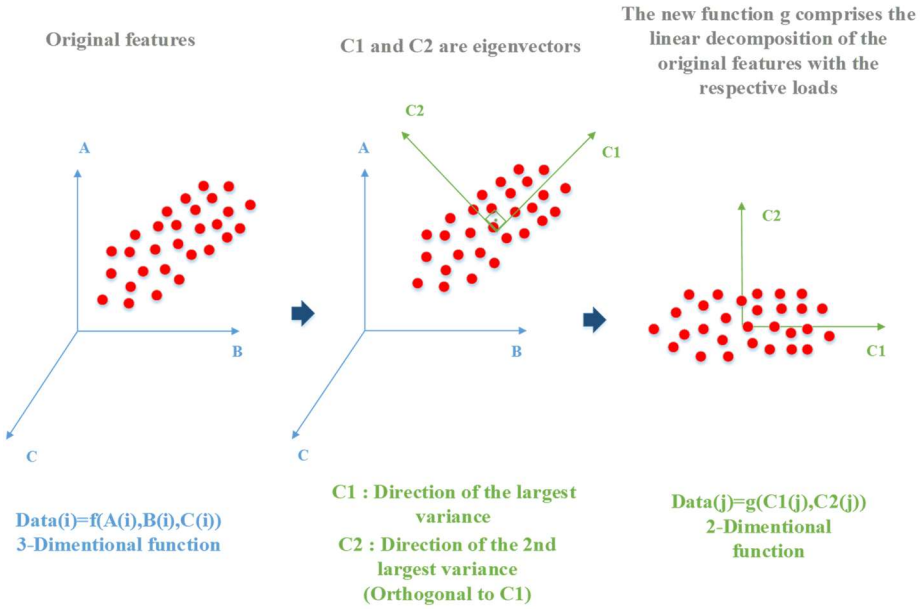
Unsupervised learning aims at identifying and describing hidden patterns. Consequently, it enhances the knowledge by identifying the aggregation level of data and its heterogeneity. In the following sequence, we present some techniques for each purpose.

4.1.1 Dimensionality reduction

Dimensionality reduction is useful to identify relevant features before applying ML algorithms. This procedure reduces the computational costs, removes noise and makes the dataset easier to use (HARRINGTON, 2016).

The Principal Component Analysis (PCA) is a well-known technique for dimensionality reduction. Briefly, it transforms high dimensional data into a new basis components, which explains partially the variance in the original dataset (GAN; MA; WU, 2007). Based on the correlation matrix, the eigenvalues and eigenvectors are calculated to discover the main structure of the data. The first main direction explains the largest variability in the original dataset. The second, necessarily perpendicular to the first, explains the second largest variability and so on. All of them represent a transformed basis with every component perpendicular to each other (ZAKI; MEIRA, 2013). The eigenvalues are the variance explained by each eigenvector. Therefore, the latter is the principal direction. Each principal component can be written as the linear decomposition of the original features and the weights or loads are representative of the feature importance in relation with each component.

Figure 12 – Principal components analysis



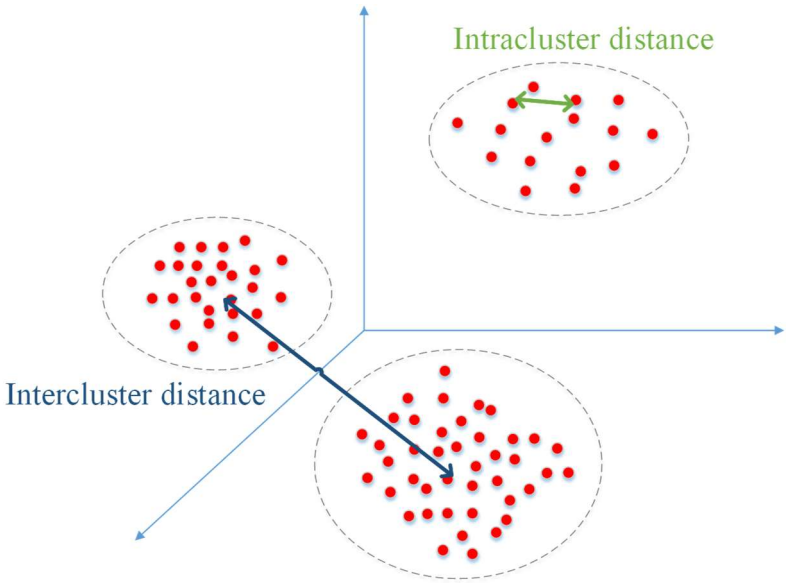
Source: the author

Relevant applications of PCA for dimensionality reduction were found in the social sciences, especially in the social vulnerability index formulation (CUTTER; BORUFF; SHIRLEY, 2003). Moreover, it is broadly used in the composite indicators analysis (BECCARI, 2016) as well as in the urban planning, as a part of the procedure to classify city areas (IBES, 2015). Especially for these studies, the reduction aims at removing the noise and correlated variables from the original dataset.

4.1.2 Clustering

In the clustering techniques, no class should be predicted, and the instances should be divided into similar groups. Therefore, cluster is defined as a group of similar instances. To determine the similarity between groups, those techniques have an input based on general distances or proximity matrix. There are two types of distance measures, intra and inter clusters (Figure 13). The goal of the clustering procedure is to minimize the within-group similarity (intra cluster distance) and to maximize the distance between distinct clusters (inter cluster distance) (JOSEPH et al., 2016). They are useful to quantify how different the groups are and to compare the similarity or distance between each attribute. The dissimilarity between clusters is intrinsically related to the accuracy of the algorithm. The more distinguished the groups are, the better the cluster assignment is.

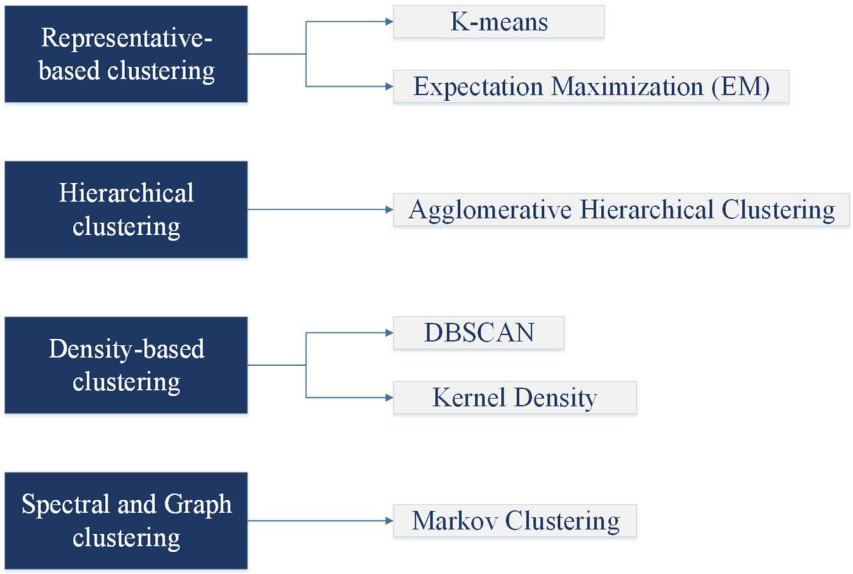
Figure 13 – Distances measures between clusters



Source: the author

According to Zaki & Meira (2013), the clustering algorithms can be: (i) representative, which divide the dataset into groups and desired number of clusters and for each group, there is a centroid that summarizes the cluster content and generally represents the mean of points assignment; (ii) hierarchical clustering, which aim at creating a sequence of divisions, that can be visualized in a dendrogram, a tree that represents the clusters arrangements, creating a hierarchy of group importance and using it to iteratively create similar groups; (iii) density-based, which are based on the local density of points instead of the distance; (iv) graph clustering, which can use graph as a reference aiming to cluster the nodes by using the edges and their weights, which represents the similarity between nodes (Figure 14).

Figure 14 - Main clustering algorithm classes

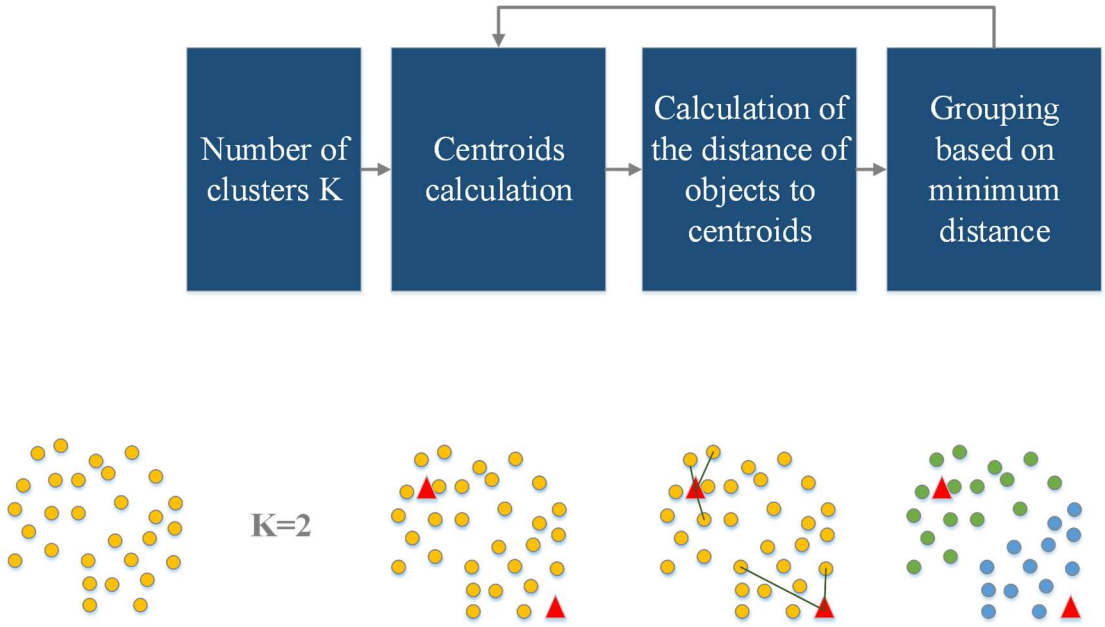


Source: adapted from Zaki & Meira (2013)

K-Means is a popular algorithm due to its intuitive character and low computational cost (ZAKI; MEIRA, 2013; JOSEPH et al., 2016). It involves two main steps: the cluster assignment and the centroid update. Firstly, the K number of clusters is set, and each point should be assigned to the closest mean. Therefore, the groups of points close to the mean value constitute one cluster. In the next step, the centroids of each cluster are updated (Figure 15). The convergence occurs if the cluster centroid does not change of a given interaction to other. The algorithm has been enhanced since its publication in Macqueen (1967). Some application of K-Means can be verified in the remote sensing area for image segmentation (DHANACHANDRA; MANGLEM; CHANU, 2015) and precipitation

estimation (MOKDAD; HADDAD, 2017) as well as the improvements in the algorithm for time series analysis (HUANG et al., 2016).

Figure 15 – K-means algorithm



Source: the author

Hierarchical clustering is a hard-clustering categorization such as the K-Means algorithm. It is based on a sequence of nested partitions, being agglomerative or divisive. The first approach considers a bottom-up approach, which starts with each point in a cluster and merges them into similar groups. The second involves a top-down approach and, at the beginning, starts with all points in the same cluster and splits them it until all points are in separate clusters (ZAKI; MEIRA, 2013). As most clustering techniques, the input is a proximity matrix. The similarity can be measures as the minimum, maximum or the average distance between two points in the clusters, distance of mean centroids or the increase in the sum of squared errors. The latter is known as Ward's method.

The Expectation Maximization (EM) algorithm is useful for incomplete data problems once it generates cluster based on the Gaussian Mixture Model. In the first step, the parameters of the probability distribution, median and covariance matrix are estimated. Next, the log likelihood expected value, i.e. the conditional probability, is maximized.

The abovementioned method of K-means and EM can be also called representative-based clustering due to the ability to identify similar groups with a single point in the centrality. They are able also to find ellipsoid clusters, in other words, the data

agglomeration presents a shape of an ellipsoid. Nevertheless, some datasets present nonconvex curves as clusters shapes. To fill this gap and to detect clusters with a nonconvex curve, the density-based algorithms use not only the distance between groups but also the local density of points to determine clusters. The general approach of the DBSCAN, a well-known density-based algorithm, is to define a sphere with a given radius around one point and the minimum number of points that features a cluster. If one point is categorized into the distance limit and the group has more than the minimum number of points, it is considered as a core point and the two or more form a similar group. This density search, nevertheless, is strongly related to the value of the radius and in some cases, if the dataset is not known or no prior information is given, the application can be limited. The method sensitivity is related to the input parameters. If the radius value is too small, the algorithm can categorize sparse clusters as a noise whereas if it is too large, some denser clusters may be merged (ZAKI; MEIRA, 2013).

This section briefly described the main classes of clustering methods. Other relevant approaches are mentioned in the literature, such as graph, fuzzy, grid-based clustering (GAN; MA; WU, 2007; HAN; KAMBER; PEI, 2012; ZAKI; MEIRA, 2013). The Table summarizes the algorithms presented according to the computational cost, interpretability of results, implementation difficulty and prior knowledge of data. The evaluation was based on ¨the literature and on an exploratory experience (by the author).

Table 7 – Summary of methods according to practical criteria

Algorithm	**Computational cost (training and testing) ***	**Prior knowledge of data***	**Interpretability of results****	**Implementation difficulty****
K-Means	++[2]	+	+	+
Hierarchical clustering	++++	++	++	++
EM	++	+	++	+++
DBSCAN	+++	++++	++	+

*+ Extremely low; ++ Low; +++ High; ++++ Extremely high

**+Easy; ++ Intermediate; +++ Hard

[2] According (HARRINGTON, 2016)

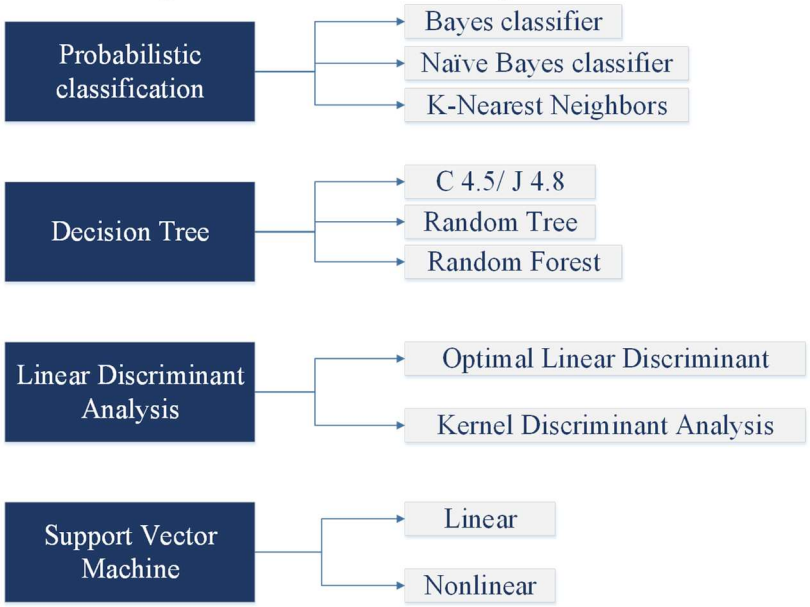
4.2 Supervised learning

Supervised learning generally aims at acquiring knowledge from dataset based on a target variable. For that, it uses association rules to estimates instances based on categorical or numerical values. The two cases task is to map the output based on the input, i.e. estimates a result with a set of explanatory variables (ALPAYDIN, 2010). The procedure trains the model with an initial dataset and applies it to classify the instances to a determined class or continuous value. It is evaluated considering the correct and incorrect instances classified and therefore, the accuracy could be associated to the model. For this, it is necessary to determine a rule or a function of prediction, which aims at relating the explanatory variables to the dependent one.

The model parameters should fit the nonlinearities and interactions existing between the variables in dataset. In the practical aspects, to overcome the potential model bias, the cross validation split the training data into k parts, the model is trained except in one of the folds and the performance is evaluated. The error is computed and the parameters which minimize the error are chosen as the final model (JOSEPH et al., 2016).

Especially, for categorical values, the classifier is the model to categorize the dataset into predefined classes (HAN; KAMBER; PEI, 2012). The algorithms can be: (i) Probabilistic, which deal with the probability of data occurrence to predict the class; (ii) Based on decision trees to split the instances; (iii) Linear Discriminant Analysis, which aims at finding a vector that maximizes the separation between the classes; Support Vector Machines (SVMs) find the optimal hyperplane that maximizes the gap or the margin between the classes (Figure 16).

Figure 16 – Main classification algorithms classes



Source: adapted from Zaki & Meira (2013)

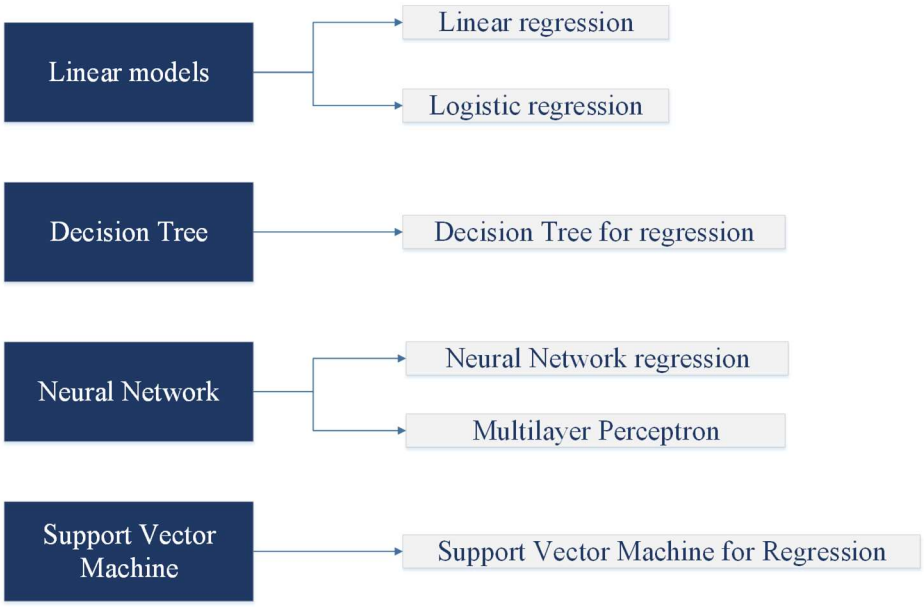
The probabilistic approach is based on a function to estimate the prediction (ZAKI; MEIRA, 2013). One of the most well-known algorithms is the Bayes classifier. It uses the Bayes theorem to classify the instance and maximizes the posterior probability for each class. There is also the Naïve Bayes, which has the same idea as the latter algorithm, but it considers that the attributes are independent. In contrast with the previous abovementioned classifier, the K-Nearest Neighborhoods uses a non-parametric approach to estimate the probability density function. It uses a density estimation method directly considering the data sample and does not make any assumptions about the joint probability. The algorithm classifies one instance "x" according to the K nearest neighborhood class. The class presenting the majority of nearest neighbors, is the class of instance "x".

The literature on classification is considerable extensive and other models are also relevant, such as linear discriminant analysis and support vector machine, but decision tree is the most widely used method (JOSEPH et al., 2016). Decision trees present an intuitive approach for instances classification and is considered as computationally inexpensive (HARRINGTON, 2016). The general idea is to construct a structure that displays, in an organized way, the dataset information. Decision Tree is a flowchart that shows the attributes to be divided. The nodes, with links and arrows, splits the data in branches. The general idea firstly involves, choosing the best attribute to split, then, separating and recalculating the best split until getting the final class. For constructing the tree, the split criterion is an important stage of the algorithm. The attributes to be divided are compared

according to their information gain, or entropy, the measurement of the degree of disorder in a dataset. When the division represents the highest information gain, it is considered the best (HARRINGTON, 2016).

In general words, regression fits a general function to a given number points (HARRINGTON, 2016). Generally, it aims at relating a continuous variable, the target, based on other dependent ones. The latter variables are generally called explanatory once they are correlated and explain the target variable. The training focus on estimating the regression weights, in the case of regression, or the parameters of the predictor, in case of classification. Then, test applies the model in a different dataset that considered training for performing the prediction (HAN; KAMBER; PEI, 2012). The classification algorithms are, for example, Decision Tree and Support Vector Machine models. The regression algorithms are, such as linear, logistic, Multilayer Perceptron and Neural Network (Figure 17). It is important to highlight that some software, such as Weka, implement some regression function as logistic and multilayer perceptron for the classes prediction and some classifiers as decision tree for regression.

Figure 17 – Main regression algorithms classes



Source: adapted from Witten et al. (2011)

In this work, the focus is discovering hidden information about income and accessibility data. Given this objective, supervised learning is applied mainly to explore relations between the income, as the target variable, and the accessibility indicators, the explanatory variables. With the loads obtained as coefficient of the regression functions to

estimates income, it is possible to determine the degree of influence of each variable increasing or decreasing income values. The function can be linear or non-linear, which aims at informing the degree of dependence between the variables. The hidden relation can be identified, such as, testing the behavior of the target feature adding one point in the explanatory variable. Additionality to this, the parameters of the model inform the performance of the chosen function.

4.3 Machine learning techniques for spatial data

Christofoletti (1999) defines the statistics of spatial data those techniques related to model geographical data without necessarily considering spatial information. Anselin (2006) presents the concept of spatial statistics as those that consider geographical features of the data in the model, as spatial correlation and neighborhood information. The ML techniques model different type of data, with or without spatial information and the algorithms can be adapted to consider geographical information in its formulation.

To illustrate the latter case, some studies present adapting clustering techniques. Rus; Kruse (2011) propose an adaptation of agglomerative hierarchical clustering for spatial data of precision agriculture. The approach demonstrates practical advantages to visualize and to deal with fertilizing regions, however, it requires specificities of precise agricultural dataset, such as hexagonal grid and spatial autocorrelation. The K-Means algorithm is also adapted for spatial statistics, being called as K-Spatial Medians. This algorithm uses the spatial median as the reference point for the cluster nevertheless, it requires high computational costs (JIN; JUNG, 2010).

Regression algorithms can also be adapted or developed for spatial statistics purposes. Geographically Weighted Regression (GWR) (BRUNSDON; FOTHERINGHAM; CHARLTON, 1996) attempts to capture the variation in data structure over the space, calibrating a multiple regression model with the different relationships between the points in the space. The application of this model was verified in the transportation literature relating the effects of transportation infrastructure on the land value (DU; MULLEY, 2007).

In this work, no prior geographical information, as neighborhood and coordinates were used in the models. The spatial information is considered in the accessibility indicators,

which brings the geographically distributed opportunities. Other related work explores accessibility data by using spatial statistics techniques for analyzing distributional effects of new transportation infrastructure (PEREIRA; BANISTER; WESSEL, forthcoming). However, to the best of our knowledge, few studies can be found that use the classical ML techniques for analyzing spatial inequalities, which sometimes represent an easy and user-friendly alternative for spatial data analysis. Thus, the classical approach of the ML algorithms implemented in Weka software were used.

5 Spatial inequalities in São Paulo

This chapter encompasses the analysis of income and accessibility data by using unsupervised and supervised techniques. First, the indicators formulation is presented and in the following subsections, the data analysis.

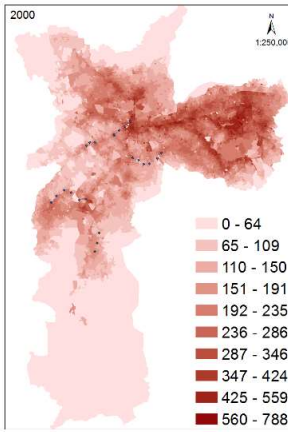
5.1 The indicators formulation

The first procedure involves calculating the accessibility indicators. As abovementioned, the travel time inference changed but the offer of urban facilities is the same on both years of analysis. The main changes, between 2000 and 2010, occur close to metro station areas, especially in the lilac line. Because of the peripheral location, the travel time changed considerably. In some cases, as the accessibility to culture facilities, the city center concentrates the main part of the facilities and there is no relevant difference between the years.

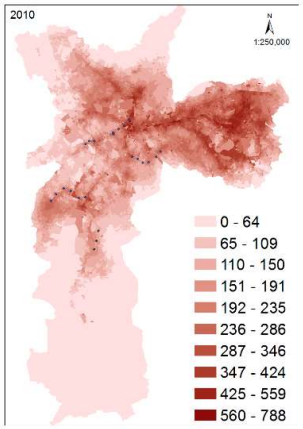
Figure 18 depicts the number of respective urban facilities to be reached given a travel time. The dots are the new metro and train stations implemented between 2000 and 2010. The division scale was natural breaks.

For the pre-processing procedure, some descriptive statistics are calculated for different dataset configuration. The accessibility measures do not present missing values because where there are no facilities, the indicator is zero. In the case of removing the instances with missing income values, it considerably changes the average value of accessibility measures – that represents six other variables of the dataset. Therefore, the missing values presented in the income variable were replaced by the mean to do not affect the other variables with no missing to be replaced.

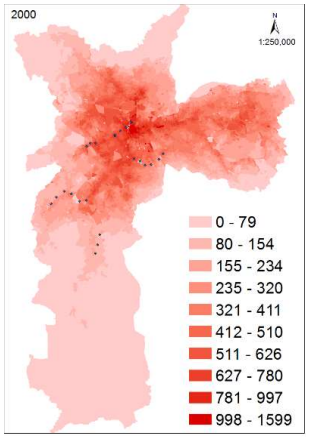
Figure 18 - Accessibility measures of 2000 and 2010



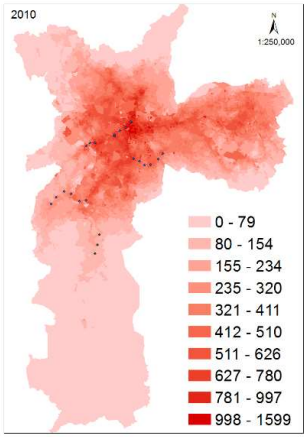
8a. Number of public schools to be reached in 45 minutes (2000)



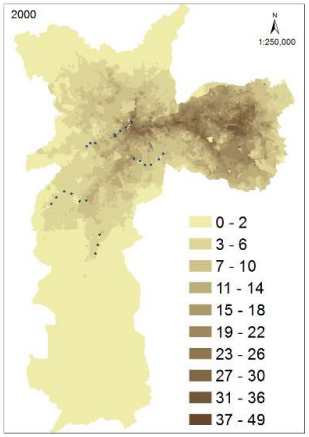
8b. Number of public schools to be reached in 45 minutes (2010)



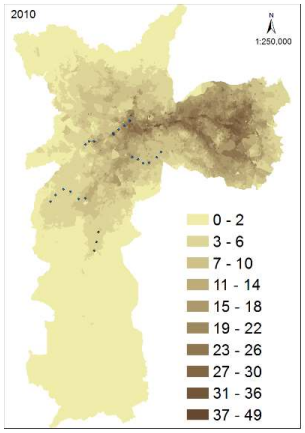
8c. Number of private schools to be reached in 45 minutes (2000)



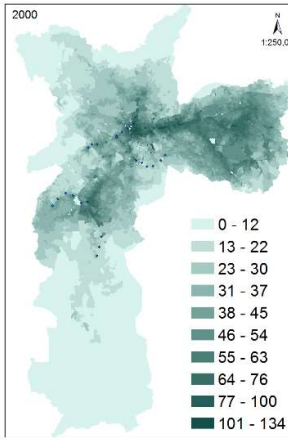
8d. Number of private schools to be reached in 45 minutes (2010)



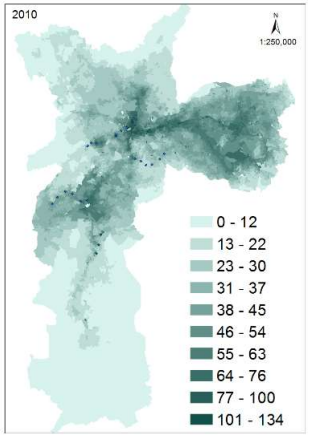
8i. Number of sports facilities to be reached in 50 minutes (2000)



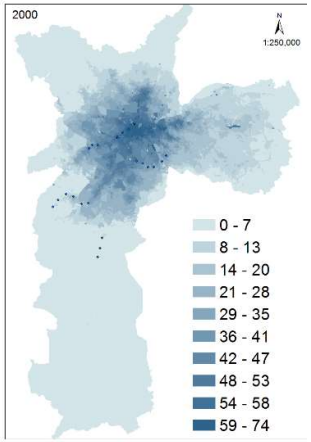
8j. Number of sports facilities to be reached in 50 minutes (2010)



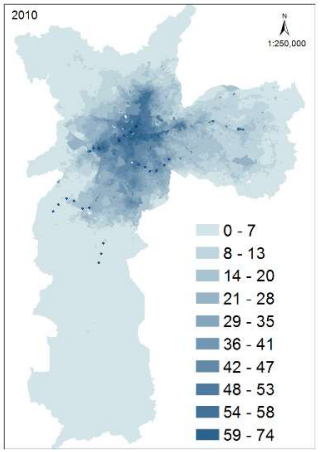
8e. Number of health centers to be reached in 60 minutes (2000)



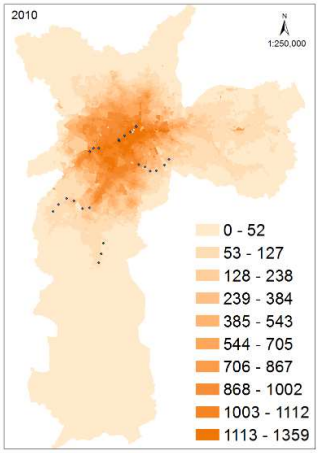
8f. Number of health centers to be reached in 60 minutes (2010)



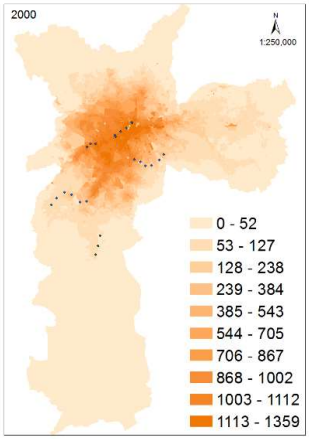
8g. Number of hospitals to be reached in 60 minutes (2000)



8h. Number of hospitals to be reached in 60 minutes (2010)



8k. Number of culture facilities to be reached in 50 minutes (2000)



8l. Number of culture facilities to be reached in 50 minutes (2010)

Source: from Siqueira-Gay; Giannotti; Sester (2017)

The descriptive statistics of the database of 2000 are depicted in Table 8 and 2010 in Table 9.

Table 8 – Descriptive statistic of 2000 dataset

	Access Hospitals	Access Sports Centers	Access Private Schools	Access Health Centers	Access Public Schools	Access Culture	Income
Average	18.20	9.12	296.27	33.70	178.15	253.13	9.76
Median	12.00	7.00	261.00	32.00	167.00	74.00	6.05
Standard deviation	16.57	7.03	192.82	16.19	89.05	358.31	9.99
Interval	72.00	49.00	1581.00	139.00	789.00	1355.00	165.06
Minimum	0.00	0.00	0.00	0.00	1.00	0.00	1.62
Maximum	72.00	49.00	1581.00	139.00	790.00	1355.00	166.68

Table 9 – Descriptive statistic of 2010 dataset

	Access Sports Centers	Access Private Schools	Access Hospitals	Access Health Centers	Access Public Schools	Access Culture	Income
Average	8.28	277.64	17.26	31.87	167.30	243.85	4.24
Median	7.00	234.00	10.00	30.00	156.00	70.00	2.46
Standard deviation	6.89	196.27	16.66	16.50	90.17	354.13	5.02
Interval	49.00	1599.00	72.00	141.00	788.00	1359.00	143.75
Minimum	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Maximum	49.00	1599.00	72.00	141.00	788.00	1359.00	143.75

5.2 Unsupervised analysis

The unsupervised analysis aims at discovering hidden patters in the dataset, revealing the heterogeneity in the data, especially, in the context of transportation planning, the groups composition. To form the groups, first the dimensionality reduction by using PCA is performed and then the clustering algorithm K-Means is applied. The proximity between the instances and how they are grouped are helpful to identify, describe and evaluate the transportation inequalities across the São Paulo city and in the surroundings of new metro stations.

The data analysis followed the question:

- Has the low-income population low accessibility to opportunities?

5.2.1 Dimensionality reduction

The PCA algorithm aims at reducing the complexity of the dataset, transforming the original data into a new dataset, with less complex variables. For visualizing the relations between the variables, the correlation matrix of 2000 dataset is depicted in Table 10. It shows the relation between two variables in the dataset. For instance, the accessibility to culture is highly correlated with accessibility to hospitals. It is confirmed by the maps shown in Figure 18 with high levels of access to such facilities in the city center. The spatial pattern of health centers, sports facilities, and public schools are also similar, therefore, these variables represent high correlation values.

Table 10 – Correlation matrix of the 2000 dataset

	Access Hospitals	**Access Sport Centers**	**Access Priv Schools**	**Access Health Centers**	**Access Public Schools**	**Access Culture**	**Income**
AccessHospitals	1	0.41	0.86	0.5	0.32	0.93	0.53
AccessSportCenters	0.41	1	0.67	0.86	0.84	0.29	0.06
AccessPrivSchools	0.86	0.67	1	0.75	0.68	0.8	0.39
AccessHealthCenters	0.5	0.86	0.75	1	0.87	0.4	0.1
AccessPublicSchools	0.32	0.84	0.68	0.87	1	0.2	-0.04
AccessCulture	0.93	0.29	0.8	0.4	0.2	1	0.6
Income	0.53	0.06	0.39	0.1	-0.04	0.6	1

Source: the author

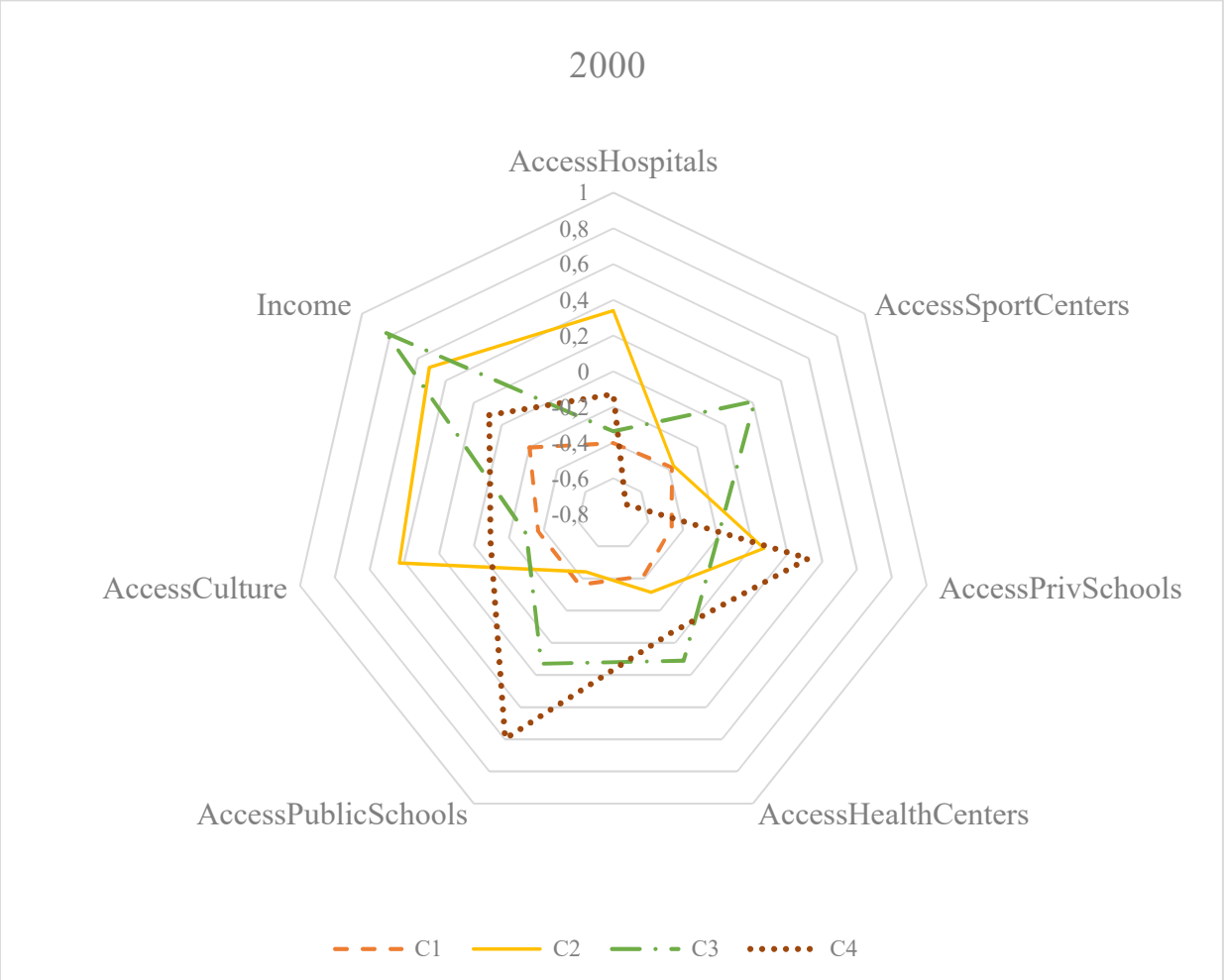
The original dataset presented seven variables: six accessibility measures and the monthly income in m.w. The PCA performance resulted in 97% of the dataset being explained by 4 main components (Table 11). The first main component (C1) explains about 61% of the variance of the original dataset and it presents low loads values in relation to all variables (Figure 19). The map (Figure 20) shows this component highly correlated to the urban fringe. The negative values are related to center-east zone centrality, that reveals the negative dependence with the main metro stations in the São Paulo municipality (Figure 5). The second component (C2) explains 26% of the original variance, accumulating 87% of the total. This component is positively related to the centrality (Figure 20) and it is highly correlated to high accessibility to hospitals, culture and high income (Figure 19). The other two components (C3 and C4) presents, respectively, correlation to income, accessibility sports and health centers and public and private schools. The C3 map shows a distinguished pattern to the city center and C4 do not present a clear difference related to the city center and peripherical region.

Table 11 – Cumulative percentage of variance of the 2000 dataset

Component	Cumulative percentage of variance explained by each component
C1	0.61
C2	0.87
C3	0.94
C4	0.97

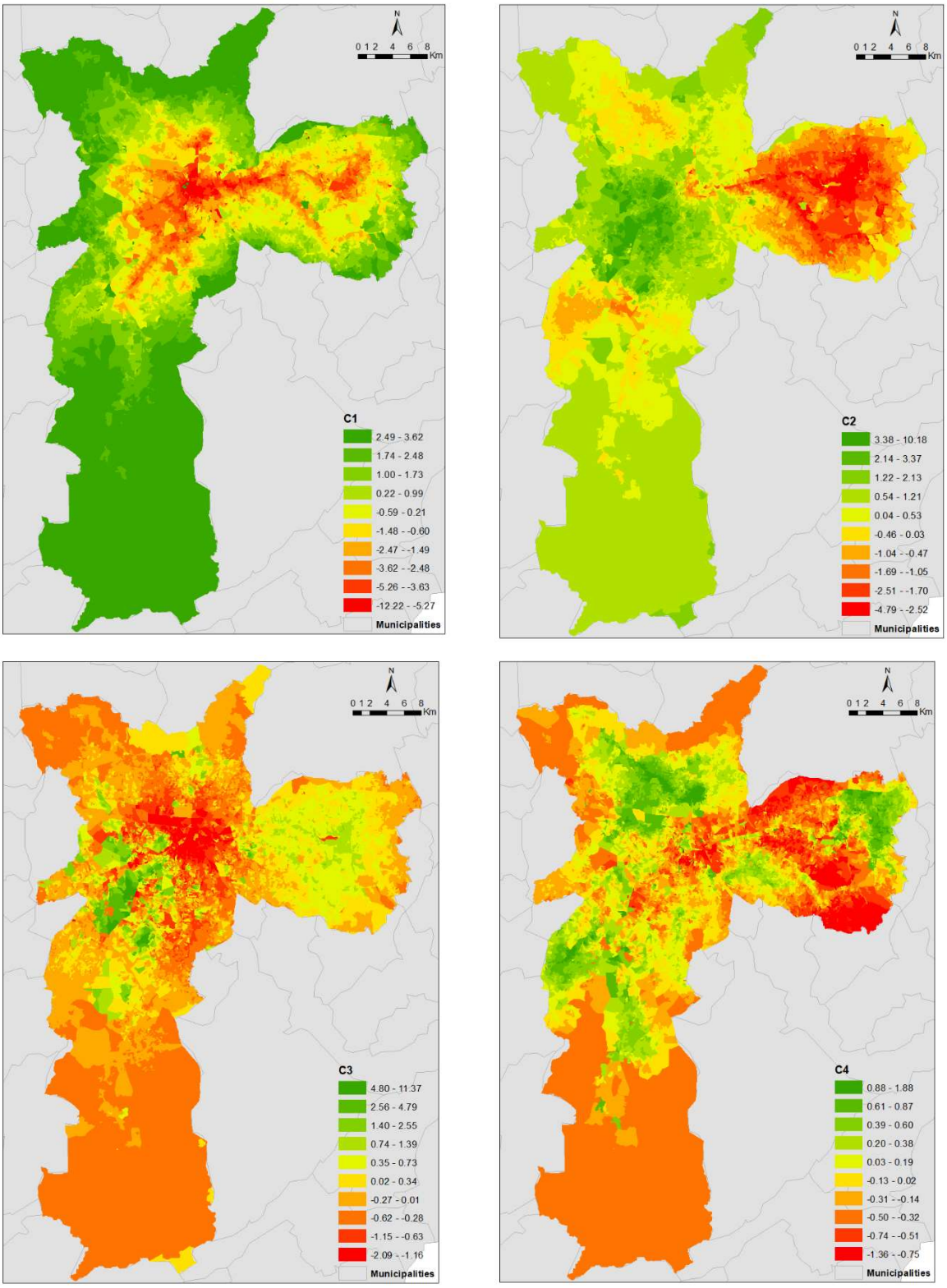
Source: the author

Figure 19 – The proportion of the original variables in each eigenvector for 2000 dataset



Source: the author

Figure 20 – Principal components of 2000 dataset



Source: the author

The analysis of the correlation matrix of 2010 dataset (Table 12) reveals a high correlation between accessibility to sports facilities, health centers and public schools, as also can be seen in 2000 (Table 10). The accessibility to cultural facilities, hospitals and private schools remain highly correlated as in 2000.

Table 12 – Correlation matrix of 2010 dataset

	Access Sport Centers	**Access Priv Schools**	**Access Hospitals**	**Access Health Centers**	**Access Public Schools**	**Access Culture**	**Income**
AccessSportCenters	1	0.7	0.43	0.81	0.84	0.34	0.07
AccessPrivSchools	0.7	1	0.83	0.72	0.73	0.81	0.34
AccessHospitals	0.43	0.83	1	0.54	0.36	0.89	0.45
AccessHealthCenters	0.81	0.72	0.54	1	0.81	0.42	0.12
AccessPublicSchools	0.84	0.73	0.36	0.81	1	0.27	-0.01
AccessCulture	0.34	0.81	0.89	0.42	0.27	1	0.53
Income	0.07	0.34	0.45	0.12	-0.01	0.53	1

Source: the author

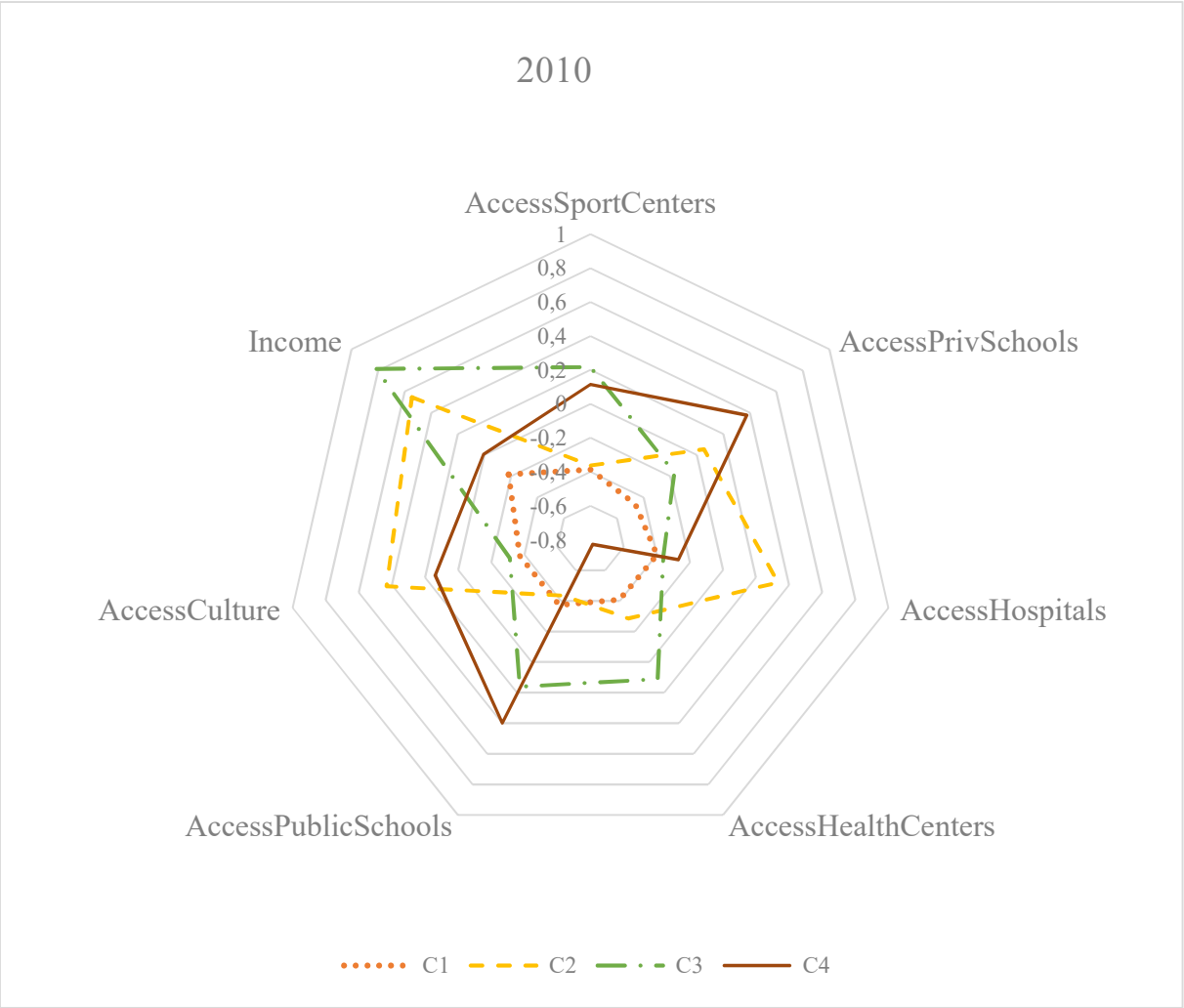
The transformation applied to 2010 dataset reveals also four main components explaining 96% of the original variance (Table 13). The first component (C1) also explain 61% and present similar relation with negative values to all variables (Figure 21) and spatial pattern (Figure 22) of 2000 dataset. The same occurs to the second principal component (C2), which explains about 85% of the original variance. The main differences are in the fourth component (C4), which is also related to public and private schools. However, the map (Figure 22) do not show the same spatial patterns of 2000 (Figure 20).

Table 13 - Cumulative percentage of variance of the 2010 dataset

Component	**Cumulative percentage of variance explained by each component**
C1	0.61
C2	0.85
C3	0.93
C4	0.96

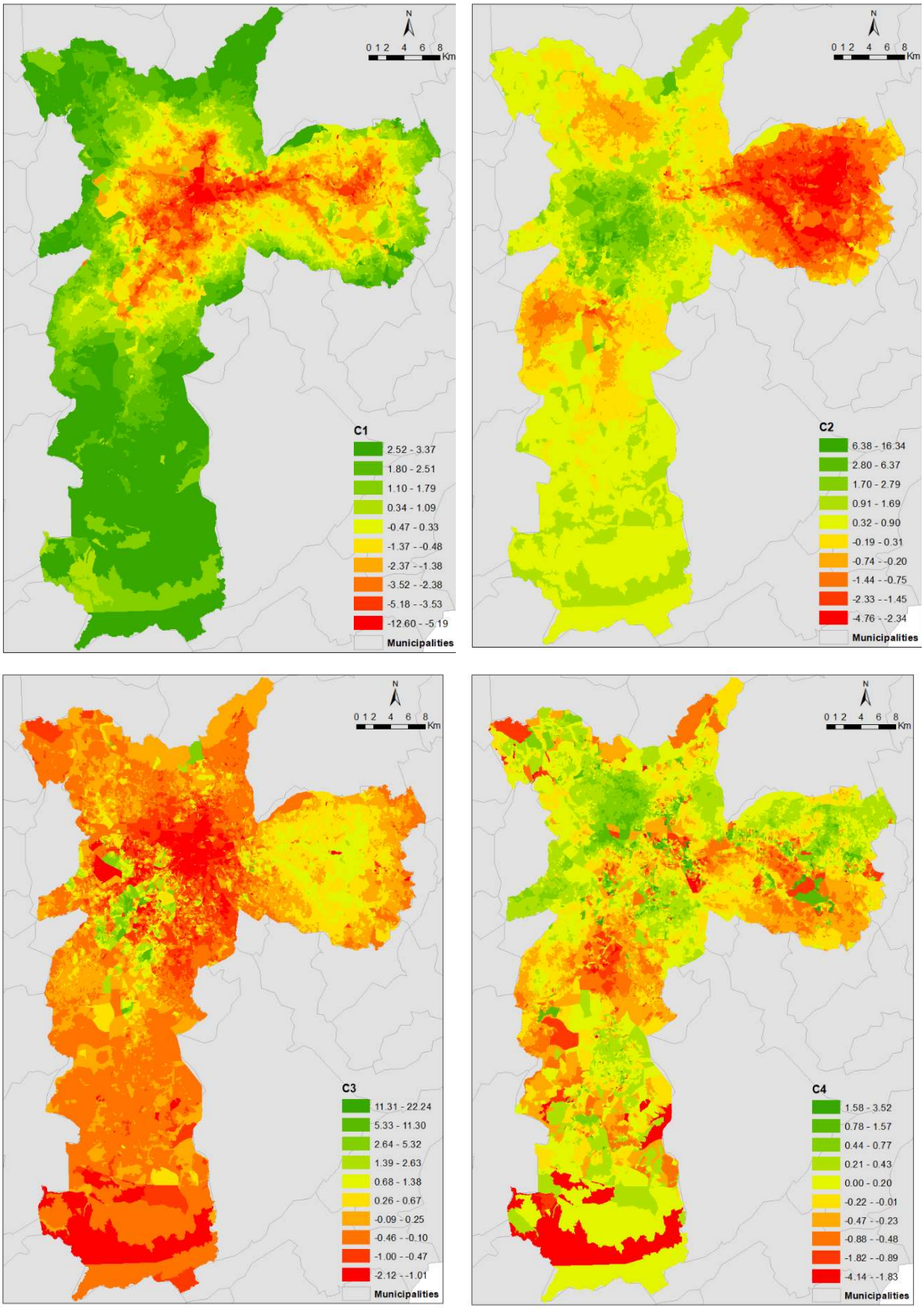
Source: the author

Figure 21 - The proportion of the original variables in each eigenvector for 2010 dataset



The two first main components were chosen to be the new dataset for clustering analysis. The similarity between the two components of both years of analysis resulted in a similar number of clusters and groups patterns of 2000 and 2010. However, in the cluster composition comparing the two years, it is possible to note considerable differences. The next section presents the results of the clustering, the maps, graphs and differences obtained in each year.

Figure 22 – Principal Components of 2010 dataset

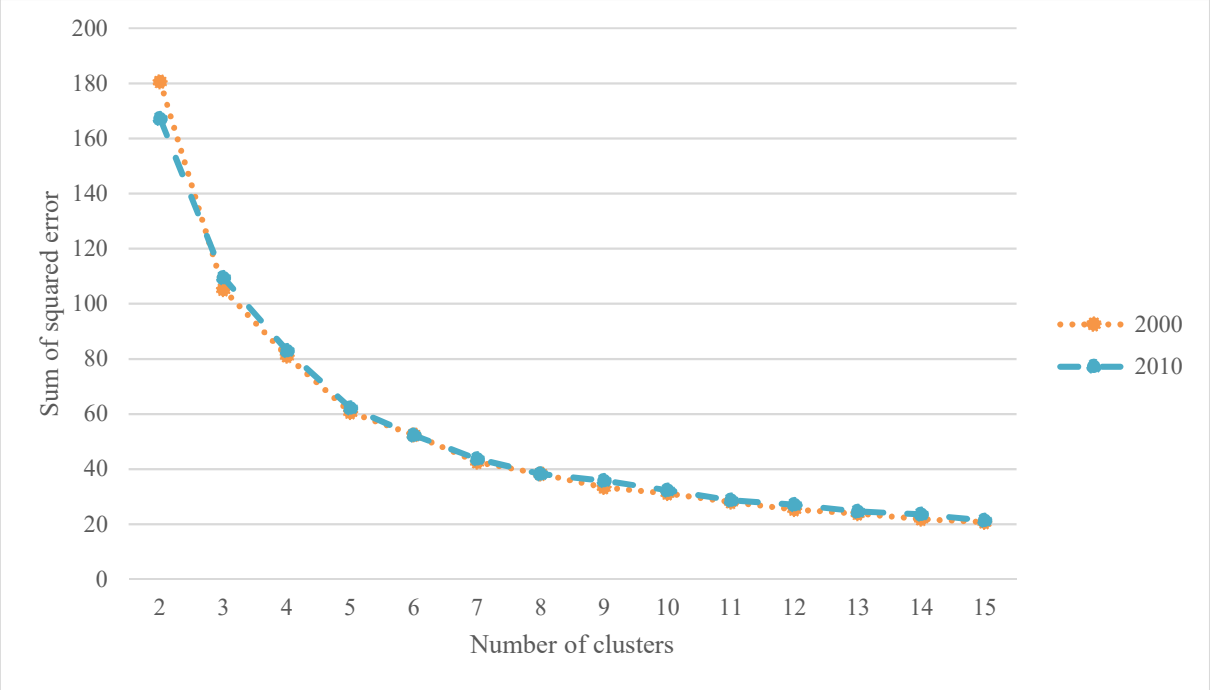


Source: the author

5.2.2 Clustering

For the clustering analysis using K-Means, the first step is to determine the number of clusters K. For that, the elbow curve (Figure 23) displays the sum of squared error, that represents the distance between the clusters. The K-Means algorithm aims at minimizing such distance. The "best" number of clusters is considered the inflection of the curve, once after the break, the error between the clusters do not change considerably. Hence, after the break the error is not significantly influenced by the number of clusters. Figure 23 depicts the number of seven clusters for both 2000 and 2010 dataset.

Figure 23 – Elbow curve



Source: the author

Some model's parameters were tested. The initialization method "random" present similar error as K-Means++ but a higher number of iterations. Considering the presented dataset, the execution time is 0.24 seconds, therefore, there is no considerable difference between the two methods. In relation to the type of distance considered between the clusters, the Euclidean distance present the smallest sum of squared errors in relation to the Manhattan distance implemented in Weka. The consideration of only two first main components also considerably decrease the sum of squared error (from around 170 to 2000 and 121 to 2010 to around 40 on both years) and the time to build the model (0.5 to 0.24).

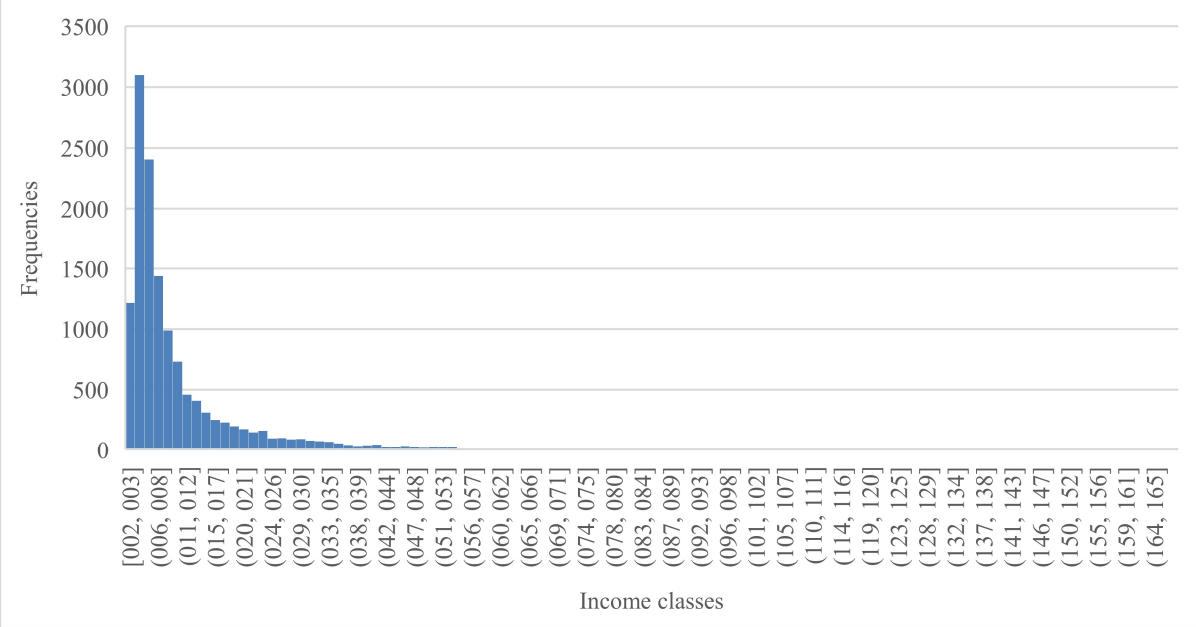
Table 14 - Models parameters

K-Means	2000	2010
Distance Function	Euclidean	Euclidean
Number of seeds	10	10
Initialization method	Random	Random
Number of clusters	7	7
Number of iterations	47	55
Within cluster sum of squared errors	42.57	43.67

Source: the author

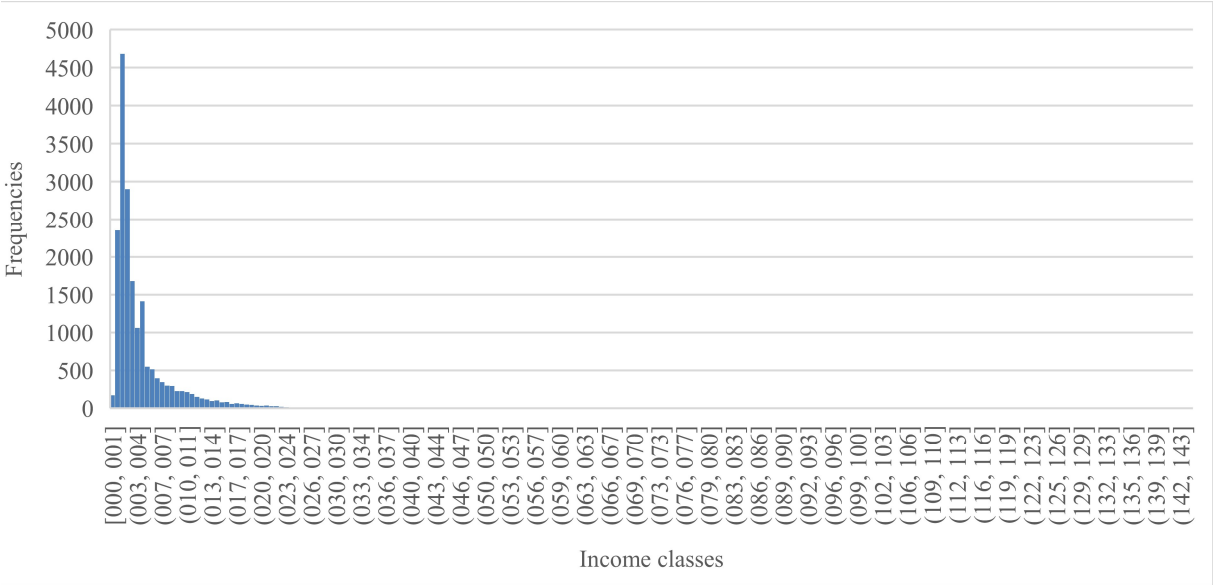
For the analysis, the clusters were classified according to the income distribution of 2000 (Figure 24) and 2010 (Figure 25). The histogram of 2000 is more spread than 2010, thus, the most part of the population in 2000 is earning up to ten m.w. and in 2010 the most part is represented by six m.w., with the higher frequency in the class of three m.w. Because of this, the division of what is low and high income in common sense for both years can bring some restrictions in the analysis. To determine the income level of each cluster, low or high, the information about the quartiles of the distribution was considered Table 15.

Figure 24 - Histogram of 2000 income



Source: the author

Figure 25 – Histogram of 2010 income



Source: the author

Table 15 – Quartile limits of income in minimum wages distribution – all data

	2000	2010
1º Quartile	1.25	0.55
2º Quartile (Median)	1.83	0.84
3º Quartile	3.29	1.55

Source: the author

To determine the limit of high and low-income, the median value of the income in each cluster (Table 16 and Table 17) is compared with the general median value of the whole income values (Table 15). The values of the median of income of each cluster that is below the general median, 1.83 in 2000 and 0.84 in 2010, are considered as low. The others are set as high. The result is four clusters of low income and three of high income in 2000 and 2010.

Table 16 – Descriptive values of income in minimum wages of each cluster of 2000

Cluster	Median	Standard deviation
6 (Low income)	1.59	0.69
3 (Low income)	1.46	0.91
7 (Low income)	1.34	1.54
2 (Low income)	1.74	0.81
5 (High income)	3.03	1.77
1 (High income)	3.99	2.84
4 (High income)	8.11	4.60

Source: the author

Table 17 Descriptive values of income in minimum wages of each cluster of 2010

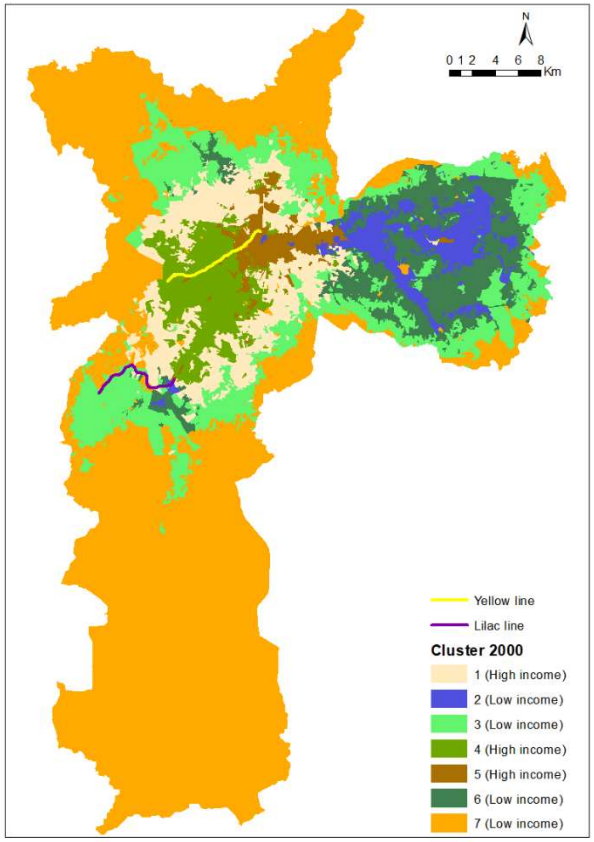
Cluster	Median	Standard deviation
3 (Low income)	0.78	0.41
7 (Low income)	0.78	0.82
6 (Low income)	0.83	0.46
5 (Low income)	0.83	0.97
4 (High income)	0.84	1.23
1 (High income)	0.84	2.03
2 (High income)	0.85	2.97

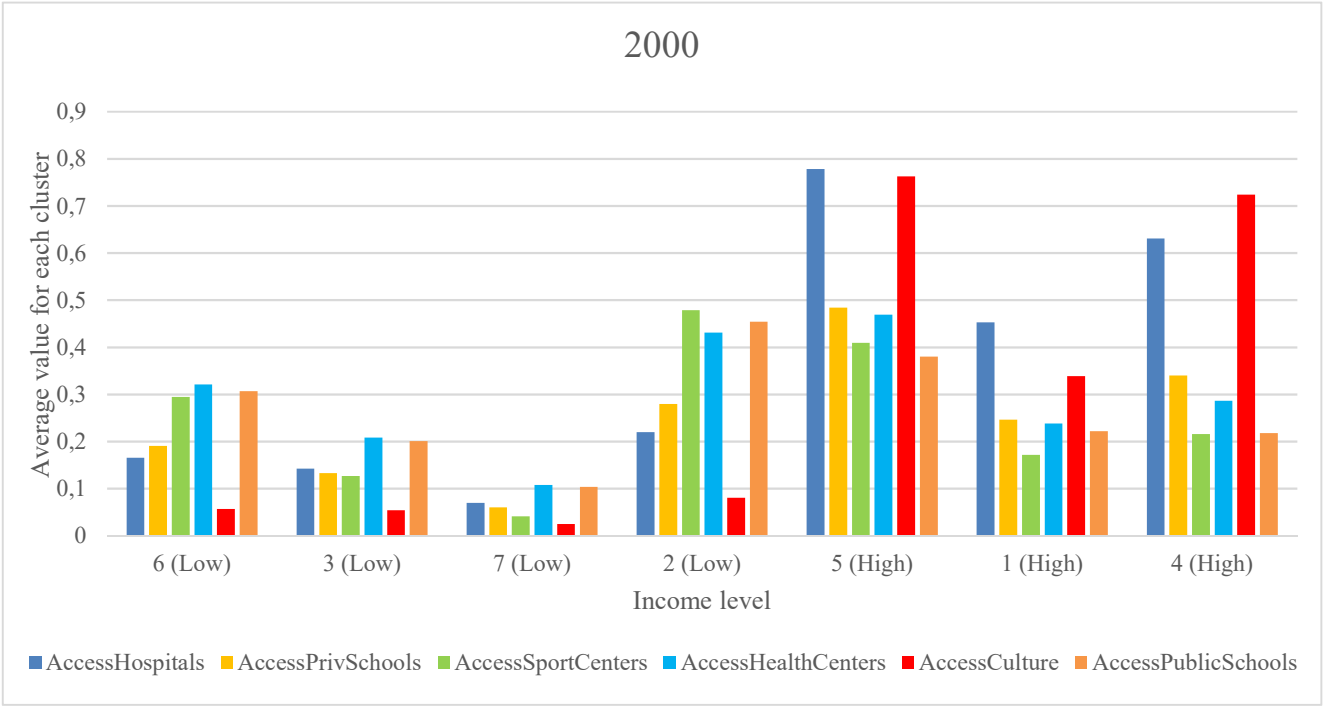
Source: the author

The standard deviation reveals that the clusters are not homogeneous in relation to income distribution. As can be seen in Figure 26, different regions of the city, with the population in distinguishing socioeconomic condition, are grouped. In 2000 dataset, the cluster number four presents high standard deviation because it involves the central region of the city as well as the south. Thus, it encompasses population with low and high levels of income (Figure 4). The cluster number seven also present a relevant standard deviation in relation to the median value. It groups also the population in the west and north region (intermediate level of income) with those located in the urban fringe (Figure 4).

For the 2010 dataset, the median values are lower and more homogenous than 2000 (Table 17). The cluster number five in 2010 show a similar spatial pattern with cluster number seven in 2000. It also presents standard deviation larger than the median value, revealing the larger dispersion of income values within the cluster composition. The cluster number two in 2010, as the cluster number four in 2000, aggregates different income level in its composition.

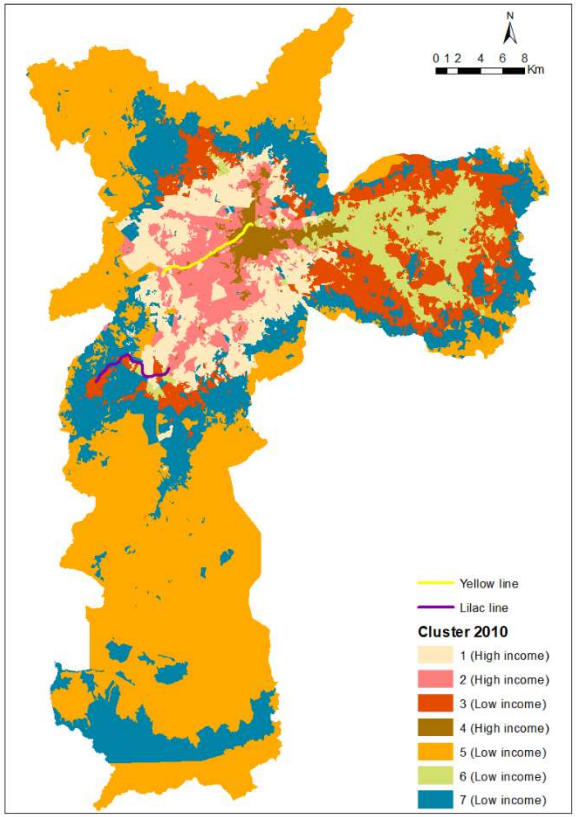
Figure 26 – Map and cluster's composition of 2000 dataset

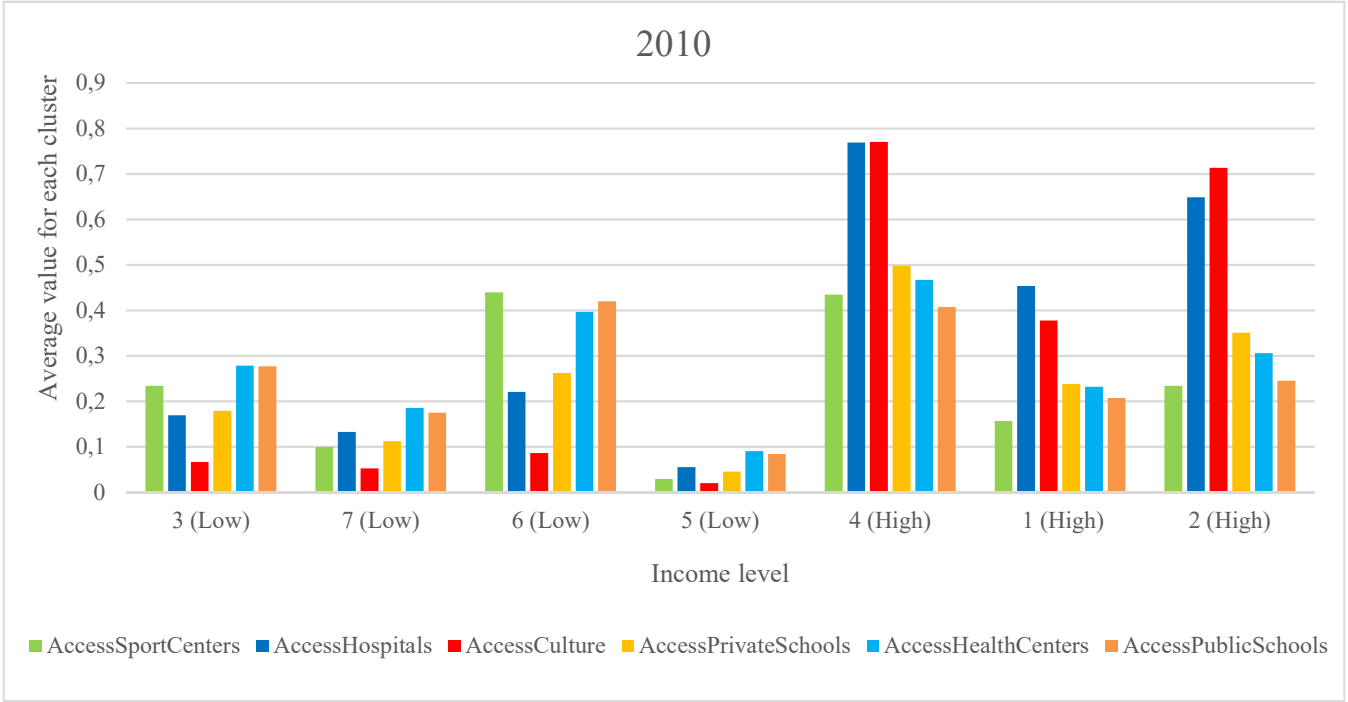




Source: the author

Figure 27 - Map and clusters composition of 2010 dataset





Source: the author

In 2000, the low-income clusters, in general, present a low value of accessibility to hospitals and cultural facilities (Figure 26). Some of these (number three and seven) present low level to sport and health centers but it is not a rule to all low-income clusters. The low level of accessibility to public schools can be seen mainly in cluster seven, located in the urban fringe. The cluster five, which connects the city center and the east zone, present high level of income and is very good located to access hospitals and cultural facilities.

The three high-income clusters represent about 20% of the population (Table 18). The low-income cluster (number seven) in the urban fringe has 27% of the total population and presents the lowest level of accessibility to all facilities. The total amount of low income is 80% of the population, with only 15% with a better accessibility condition.

Table 18 – The summary of clusters composition - 2000 dataset

Clusters	Percentage of population	Description	
		Income and location	Accessibility
1	11%	High-income cluster located in the border of city centrality[3]	It presents high access to hospitals and culture. Intermediate level of access to the other facilities
2	7%	Low-income cluster located in the east zone.	It presents the highest level of accessibility to all facilities (excluding sport centers and public schools)
3	30%	Low-income cluster located in the region between the inner city and the urban fringe	It presents low accessibility to all facilities. The higher access levels are for health centers and public schools
4	7%	High income cluster located in the city centrality	It presents high accessibility level to all facilities, especially to culture and hospitals
5	3%	High income cluster located in the city center, connecting the centrality to the north-east zone, located close to metro stations	It presents the highest level of accessibility to all facilities.
6	15%	Low-income cluster located in the center area of east region	It presents the lowest income level but with high access to sport centers, health centers and public schools. It present low access to culture facilities
7	27%	Low-income cluster located in the urban fringe	It presents the lowest level of accessibility to all facilities

Source: the author

In 2010, the cluster number five with the lowest level of accessibility to all facilities is in the urban fringe and represent 20% of all population of São Paulo (Table 19). The general

[3] City centrality in São Paulo is known as “centro expandido”. It is not coincident with the geographical center of the city but a result of urbanization process.

tendency in the cluster of low income is the low level of accessibility to culture facilities and hospitals. The sports centers appear better distributed to the low-income clusters than to the high-income. In 2010, the high-income cluster involve 21% of the São Paulo population (Table 19).

Table 19- The summary of clusters composition – 2010 dataset

Clusters	Percentage of population	Description	
		Income and location	Accessibility
1	10%	High-income located in the in the border of city centrality	It presents high accessibility to hospitals and cultural facilities and the lowest level of accessibility to health centers (on average considered as intermediate considering the other groups)
2	8%	The cluster presents the highest level of income located in the city centrality	It presents high accessibility level to all facilities, especially to culture and hospitals
3	19%	Low-income cluster located in the center area of east region	It presents low access to culture facilities and intermediate level of the other facilities
4	3%	High income cluster located in the city center, connecting the centrality to the north-east zone, located close to metro stations	It presents the highest level of accessibility to all facilities
5	20%	Low-income cluster located in the urban fringe	It presents the lowest level of accessibility to all facilities
6	9%	Low-income cluster located in the east region	It presents low level of access to culture facilities, intermediate access to private schools and hospitals and high to sport facilities, health centers and public schools
7	31%	Low-income cluster located in the region between the inner city and the urban fringe	It presents low level of access to culture facilities and intermediate level of access to hospitals, sport facilities, health centers, public and private schools

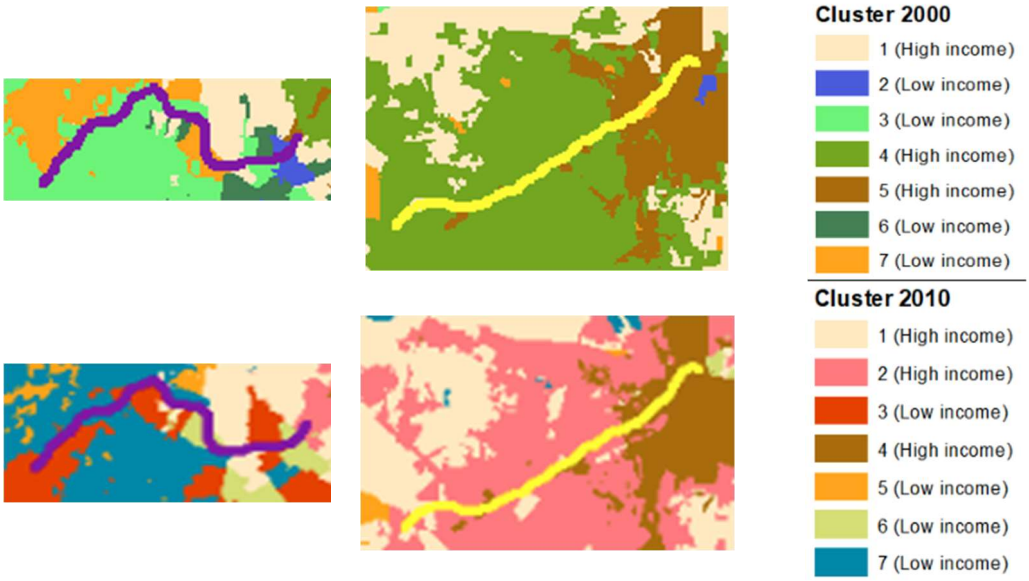
Source: the author

Some clusters of 2000 and 2010 presents similar spatial pattern (in the same color in Figure 26 and Figure 27). The cluster number one of both years represent the border of city centrality. The accessibility to all facilities follows the same trend in the clusters of high income in 2000 than 2010. In 2000, the cluster number seven present a similar spatial pattern as the number five of 2010 and both present the lowest level of accessibility to all facilities.

Especially in the surroundings of metro stations, the lilac line, in the left side in Figure 28, present the predominance of cluster number seven and three in 2000, both have low-income population and low level of accessibility to all facilities (Table 18). In 2010, the predominance is of cluster number seven, with the low level of accessibility mainly to culture facilities (Table 19). In the yellow line, on right side in Figure 28, it is possible to note similar spatial pattern in the cluster dispersion, on both years, with clusters of high income and high accessibility to all

facilities. The differences are larger in the clusters composition of surroundings of the lilac line in relation to the yellow line.

Figure 28 – Surroundings of metro stations of 2000 (left) and 2010 (right)



Source: the author

5.3 Supervised analysis

In the supervised analysis, a target variable can be investigated. It is possible to identify a relation between the variables in the dataset based on the weights of the regression function. Given this potentiality, the following questions are investigated:

- What is the relation between income and other accessibility?

5.3.1 Linear regression

For estimating the change in the accessibility value, the influence of other variables can be explored using the regression. Although the functions are useful to estimate, also the model aims at informing about better the relations between the variables (BATTY, TORRENS, 2001).

For the 2000 dataset, the linear regression function implemented in Weka was applied. The income values were in m.w. and it was used as a target feature and model parameters are presented in Table 20. The option to eliminate colinear attributes was chosen. The attribute selection tested the M5's method, which the attributes removal followed the smallest

standardized coefficient until no improvement is observed in the estimated error given by the Akaike information criterion, and a greedy selection using the Akaike information metric. Both presented similar performance in the attributes selection. The variables were normalized into zero to one scale for providing comparable coefficients.

Table 20 – Model's parameter of the regression of income considering 2000

Summary of model's parameters	
Correlation coefficient	0.6196
Mean absolute error	4.5715
Root mean squared error	7.8395
Relative absolute error	70%
Root relative squared error	78%
Total Number of Instances	13278

Source: the author

The function is shown in (5.1). All accessibility indicators present relation with income. The accessibility to hospitals, sport centers and culture have a positive relation with income. Therefore, the higher the value of these accessibilities, the higher is the income. In the other hand, the accessibility to health centers, private and public schools have a negative relation. Besides that, considering this formula, the highest load is for accessibility to culture facilities, then it is possible to state that accessibility influence more the income value than the other accessibilities.

$$\begin{aligned} Income = {} & 1.72 \cdot AccessHospitals + 5.21 \cdot AccessSportCenters \\ & -8.29 \cdot AccessPrivSchools - 5.78 \cdot AccessHealth \\ & -11.06 \cdot AccessPublicSchools + 25.33 \cdot AccessCulture + 9.08 \end{aligned} \quad (5.1)$$

For 2010, the model parameters are shown in Table 21. The function (Equation 6.2) resulted of the linear regression analysis show the sport centers and accessibility to culture facilities positively related to income and accessibility to public and private schools negatively related. The distribution of income is considerably different on both years, revealing also lower values in 2010. For example, both accessibility to culture facilities demonstrate to be stronger but 2010 present lower value (12.07).

Table 21 - Model's parameter of the regression of income considering 2010

Summary of model's parameters	
Correlation coefficient	0.55

Mean absolute error	2.1913
Root mean squared error	4.1907
Relative absolute error	72%
Root relative squared error	84%
Total Number of Instances	18953

Source: the author

$$Income = 2.09 \cdot AccessSportCenters - 3.98 \cdot AccessPrivSchools \\ -6.84 \cdot AccessPublicSchools + 12.07 \cdot AccessCulture + 3.86 \quad (5.2)$$

In both years, the correlation coefficient is about 0.6. It occurs probably because the relation between the accessibilities and income is not linear. For further developments, other variables can be tested using the potentialities of regression. Besides that, another algorithm, such as decision trees and neural networks, can support instances classification. In this sense, variables can be assigned considering a target indicator, that can be accessibility or income.

5.4 Discussion

We presented in this section the quantitative analysis of spatial inequalities in a Brazilian metropolis. By using ML techniques, the accessibility indicators were explored to understand the heterogeneity of opportunities inequalities in the territory considering the socioeconomic level of the population. The dependence between the income and accessibilities values was also explored by the regression analysis. Other works also explore the spatial inequalities throughout the São Paulo territory. Marques (2005) explores clusters in the São Paulo Metropolitan Region considering some Census variables of 2000 of education, income, infrastructure, population density among others. The results show a cluster in the city centrality ranked as high income, high scholarly rate, best infrastructure, with elderly and with a decrease growth rate. In the present analysis such cluster is spatially coincident with the cluster number four and present high accessibility to all facilities, especially to hospital and culture.

Considering other works about São Paulo and segregation, Feitosa; Câmara; Monteiro (2007) show, considering the 2000 dataset, the centrality with isolation of high-income families, therefore, the high income is surrounded by population with other income level. Sposati;

Monteiro (2017) discuss the degree of social inclusion and exclusion[4] in the year of 1991, 2000 and 2010 using also the Census data. Sposati; Monteiro (2017) demonstrate also, in the central region, the existence of a group with the highest level of social inclusion, i.e., with a better life condition. It is important to highlight that this group present the lowest number of inhabitants (Table 18 and Table 19), therefore, the good condition of accessibility and quality of life is accessed by a few proportion of population. Such analyses demonstrate that this region, which received the new stations of yellow line, do not present the most unequal group of population. The cluster number seven in 2000 (Figure 26) and five in 2010 (Figure 27) present the lowest level of accessibility and, in the analysis of degree of exclusion presented by Sposati; Monteiro (2017), they comprise the districts with the most intense degree of social exclusion.

Some interesting comparison between the result of Sposati; Monteiro (2017) and the obtained by the present work is the clusters in the border of city centrality - number one in 2000 (Figure 26) and 2010 (Figure 27). They present moderate social exclusion but is classified with high income (with high standard deviation, thus it incorporates both high and, at least, intermediate income), high access to hospitals, culture and intermediate level of access to other facilities. The 10% of population lives in that region and are considered as a homogenous groups by Sposati; Monteiro (2017).

The group in the east zone present severe degree of social exclusion but considering the accessibility level, they do not present the worst condition (Figure 26 and Figure 27). They are classified as low-income and represent about 20% of city's population (Table 18 and Table 19) and lives close to metro station and bus corridors. Besides that, the offer of some facilities as schools and health centers are not critical. The quality of the service is not considered in the accessibility measures, hence, only the presence or absence of infrastructure is assumed.

Especially considering the reality of east zone of the city, Érnica; Batista (2012) investigate the relation between the neighborhood and the education quality. They declare that as more vulnerable the surroundings of the school, worse is the quality of educational opportunities. Because of that, even the east zone providing a considerable number of public

[4] "Social Exclusion" is used hereafter in accordance with the concept presented by Sposati; Monteiro (2017)

and private schools (Figure 18), the population living in such region do not necessary access a good educational service.

In São Paulo, it is possible to note the educational inequalities throughout the city's territory. For instance, the population with the higher income level have graduate degree and attend mostly private schools (PEROSA; LEBARON; LEITE, 2015). For this high-income group, which lives in the centrality, the region is supplied by metro station and present the higher values of private school's accessibility (Figure 18). However, such supply is wasted by a group that do not use public transportation and does not need for reaching different city's regions because they live close to the opportunities. On the other hand, in a distinguished condition is the lilac line, in the south. The population living there need for reaching the opportunities in the centrality and do not present a high level of accessibility to public schools. Further works should explore the differences between the groups considering the age and the daily differences of demand for educational purposes for a better diagnosis of such demand.

In this work, the most critical facility for the low-income population is the cultural centers. The offer of museums and public libraries are mainly concentrated in the city center and positively related to income (see Equation 6.1. and (5.2.). This result evidences the need for providing more cultural opportunities in the peripherical region. A different behavior can be noted by the spatial pattern of accessibility to sport facilities. Higgs; Langford; Norman (2015) highlights the importance of measuring the proximity with sport facilities as a proxy of the quality of life and interaction between the population. In the clusters composition, the cluster of the east zone - number two in 2000 (Figure 26) and number six in 2010 (Figure 27) – even classified with low income and in a vulnerable region of the city, present the highest level of accessibility to sport facilities.

Considering the health opportunities, the accessibility to hospitals is high correlated with accessibility to culture, therefore with high values concentrated in the city centrality (Table 10 and Table 12). In 2000, the linear regression function for income prediction present the accessibility to hospital with a relevant load, demonstrating the dependence between such variables. On the other hand, the accessibility to health center is more spread across the city territory, in the east and south region (Figure 18). Because of this, in the low-income cluster, this type of opportunity do not presents one of the most critical accessibility level. Neutens (2015) present a review of geographical accessibility of health services. He also presents the main challenges of assessing such type of service once it depends of the affordability and costs of the health services, the quality of the service, availability, accessibility and travel impedance between the patients and the service as well as accommodation. In this sense, it is possible to

note that there is a lot to do in the way of exploring more this type of accessibility given the groups condition (age and gender for example, besides the income) in transportation plans.

Given the presented results, it possible to demonstrate the potentiality of exploring the cluster composition, its heterogeneity and dependence between the variables. The techniques, using or not the geographical information in the model, can be explored by planners as a powerful tool to acquire knowledge about the territory and its inequalities. Although such analysis presents potential information, some restrictions also should be considered, such as the Modified Area Unit Problem (MAUP). Recent research points out the need for such evaluation, especially considering the land use-travel interaction (KWAN; WEBER, 2008). Kwan; Weber (2008) find out the invariance of space-time accessibility with the unit of analysis, therefore, the relations do not show substantial variation at different geographic scales. On the other hand, (PEREIRA; BANISTER; WESSEL, forthcoming) demonstrate the dependence of equity assessment to spatial scale and area unit. However, the author highlights that the identified most suitable area can change according to the accessibility measure and opportunity analyzed. Therefore, the is no consensus in the literature and each region or case study should be analyzed in detail. For the planner's practice, such restriction should be considered, and the scales set according to the level of the project details.

6 Outcomes for transportation planning

To conceptually discuss the social aspects of transportation planning, some ethics theories are used to support the discussion of the policies and its outcomes. The analysis of the plans states the following question: to whom is the transportation improvements delivered? The answer is mainly related to the fairest distribution pattern desired to be provided.

The egalitarianism defines as principles that (i) the individual's basic rights and liberty have the same rule the same equality to all, (ii) social and economic inequalities can be understood as fair if they are consequences of situation of fair equality and work to benefit the most deprived group (RAWLS, 2006; PEREIRA; SCHWANEN; BANISTER, 2017). It also declares that the fairness distribution for all groups does not favor one specific group (LUCAS; VAN WEE; MAAT, 2015). Pereira et al. (2017) argue that Rawl's theory that "justice is not about whether some people enjoy greater accessibility than others, but about how institutions and policies deal with such inequalities to minimize inequality of opportunities". According to this, not all population should have the same level of access to opportunities. On the other hand, utilitarianism focus on providing equal human well-being, in other words, utility and welfare, to everybody disregarding his/her social and economic condition (PEREIRA; SCHWANEN; BANISTER, 2017). On the other hand, capabilities approach is based on principles equal respect and according to this, the opportunities, as central and basic capabilities, should be reached by all population, given a specific threshold of minimum attendance level.

The knowledge acquired from the analysis of accessibility measures and socioeconomic data can inform policymakers to discuss some ethics theories paradigms as: (i) what should be delivered in term of opportunities? (ii) to whom and which level of service is expected to be delivered? (PEREIRA; SCHWANEN; BANISTER, 2017).

Firstly, to answer the first question this work analyzed six types of urban facilities: health centers, hospitals, public schools, private schools, sports centers and culture facilities. The metric of cumulative opportunities allows the interpretation of facility offer and travel time in the perspective of place-based measures (NEUTENS et al., 2010). However, it is important to highlight that the change in the number of facilities or even its quality is not addressed in the proposed accessibility measures of 2000 and 2010. The literature evidences the need for improving accessibility measures regarding information about population, for this, more comprehensive transport survey to capture relevant information about the individuals must be undertaken (PEREIRA; SCHWANEN; BANISTER, 2017).

The second question focus on the discussion of the groups and the fairest distribution pattern desired. By using ML techniques, the information about the most critical accessibility to low-income population, thus, the dependence between income and accessibility, revels that the accessibility to culture facilities and hospitals need for receiving more attention. The accessibility levels of the population living in the surroundings of such metro stations are high to almost all facilities. On the other hand, in the lilac line, it is possible to note most heterogeneous groups in the surroundings in 2000 than 2010. The change in the travel impedance in the lilac line present a most significant change in accessibility in comparing with yellow line due to the difference in the transportation supply in both areas.

Based on the already identified lack of knowledge in the practitioner's perception about accessibility metrics and inequalities assessments, the development of frameworks and the user friendly material is encouraged (BOISJOLY; EL-GENEIDY, 2017b). To summarize the potential contribution to transportation planning, two questions are unfolded based on the central research question of "How spatial inequalities could be investigated by using ML techniques to inform the current transportation planning practice?":

- How could spatial inequalities be explored by using ML techniques to inform transportation planning?

- What is the relation of ML contribution with the existing transportation planning structure?

6.1 How could spatial inequalities be explored by using ML techniques to inform transportation planning?

To summarize the contribution of the analyses already presented in this work, Table 22 depicts the potentialities and barriers of using ML techniques. The lack of data is highlighted as one of the major obstacles to the use of accessibility metrics in practice (BOISJOLY & EL-GENEIDY, 2017b). The data required to perform the analysis usually is from the transportation infrastructure, land use and opportunities and socioeconomic variables. In some cases, it can be acquired from different inputs (SESTER, 2000; WENDER et al., 2010).There is another hurdle for implementation related to the practitioners' experience and knowledge (BOISJOLY & EL-

GENEIDY, 2017b), organizational barriers and institutionalization of instruments (SILVA et al., 2017). Especially for ML techniques usage, there is a need for an experienced data analyst who can explore better which techniques are most suitable for the problem at hand.

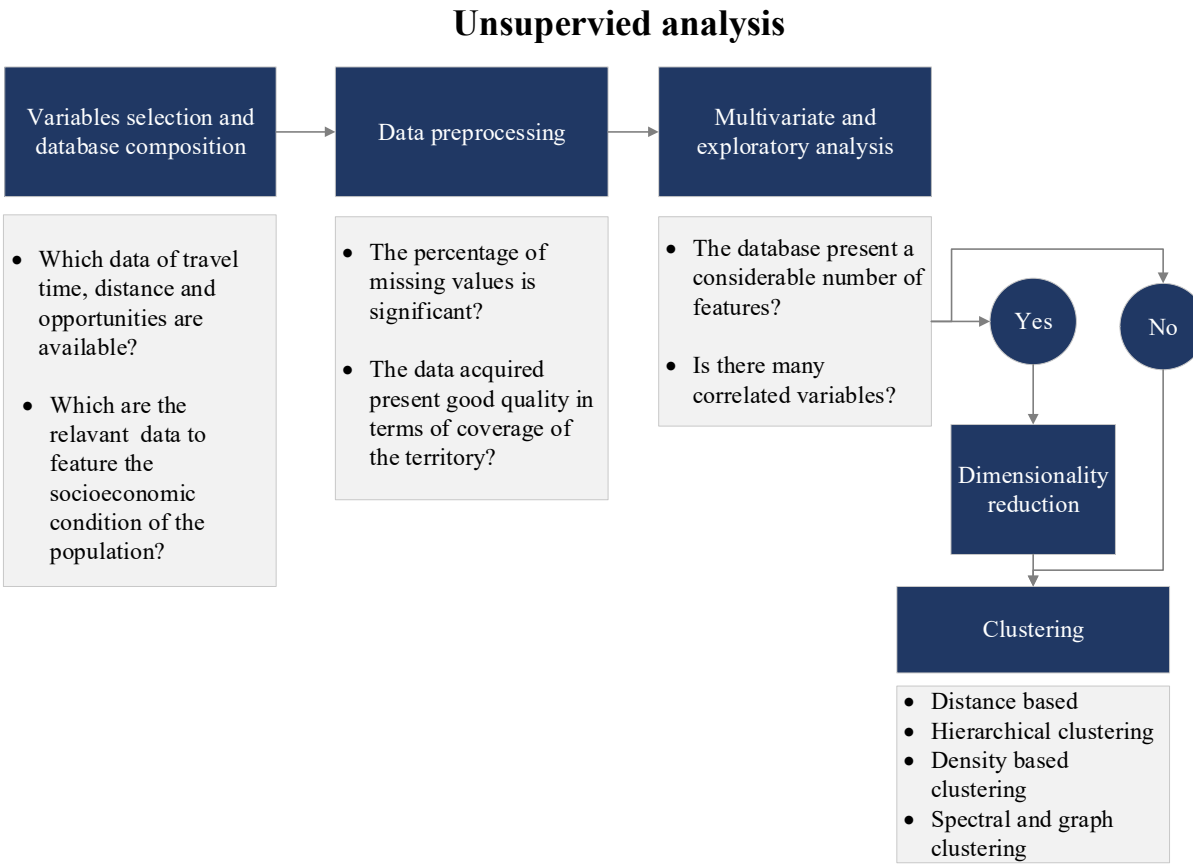
Table 22 – Summary of potentialities and risks of the application of ML techniques to explore inequitable distribution of opportunities

Problem focus on inequitable distribution of opportunities – Exploring differences across social groups		
Technique	**Potentialities**	**Barriers**
Clustering	It discovers hidden patterns of accessibility in the groups composition. It allows to visualize the heterogeneity of the provision of transport and urban facilities	The is a need of practitioner's previous knowledge in data analysis. It is also dependent of the quality and availability of the transportation, urban facilities and socioeconomic data.
Regression	It establishes relation between the one desirable variable and the other of the dataset	

Source: the author

To provide a useful reference for implementation of ML techniques, the main steps of the analysis and respective checklist are provided in Figure 29 and in Figure 30.

Figure 29 – Check list for unsupervised analysis



Source: the author

Step 1. *Variables selection*: This first step is for calculating the indicators based on the available data and analysis purpose.

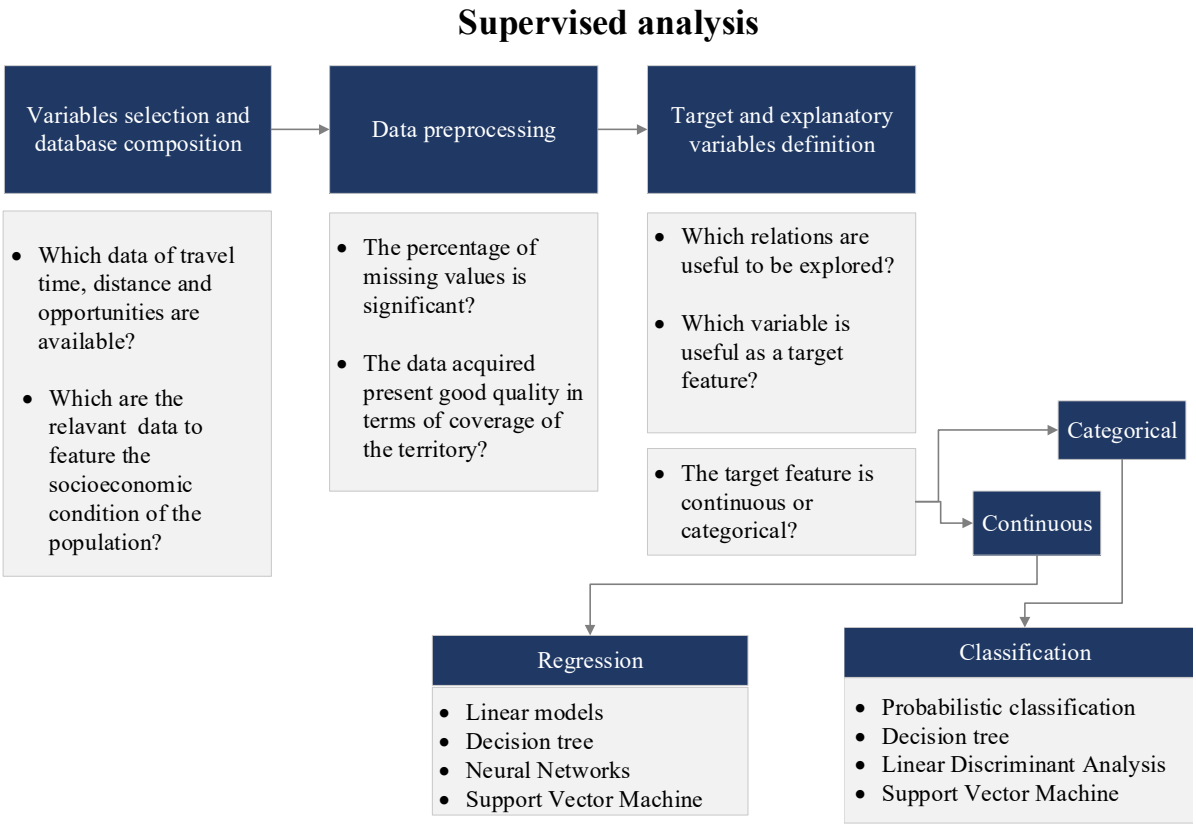
Step 2. *Data pre-processing:* There are three general methods to treat missing values in the database: case deletion; single imputation with the median or mean, or multiples imputation in each instance based on other information of the dataset (OECD, 2008). If a variable present more than 5% of missing values, cases should not be deleted (OECD, 2008). The data should also have consistence, especially in the case of being from user's surveys. For the analysis across the territory, also the coverage should be suitable for the scale of analysis.

Step 3. *Multivariate and exploratory analysis*: Other techniques can be applied, as a composite indicator construction or quantiles analysis to better understand the data. In general, the correlation matrix and descriptive statistic can inform about the complexity of the dataset.

Step 4. *Dimensionality reduction*: This stage is not mandatory for the data analysis. It fits for a complex dataset and improves some ML algorithm performance.

Step 5. *Clustering:* Different algorithms can be applied (for more see section 5.1.2) based on the problem to be analyzed.

Figure 30 – Check list for supervised analysis



Source: the author

Step 1. *Variables selection and database composition*: This stage aims at composing the database based on the availability of data and problem to be analyzed. It is interesting to explore in supervised learning, a variable of interesting (target). In transportation planning, it can be one or more types of accessibility, travel time, opportunities, transportation modes preferences or socioeconomic variables, such as income as analyzed in the present work.

Step 2. *Data preprocessing*: For supervised analysis, the missing values are critical because they interfere with the model (functions or rules) parameters. In this sense, the imputation should be carefully treated. Also, the correlation matrix can already inform about relations in the dataset and depend on the regression algorithm, the function can confirm the information already acquired.

Step 3: *Target and explanatory variables definition*: In this phase, the problems and questions to be answered should be clear. They will guide the selection of what feature should be the aim of the relation inference (target) and explanatory ones.

Step 4: *Supervised method application – Regression or classification:* The algorithm to be applied will depend on the type of already defined target feature: categorical and continuous.

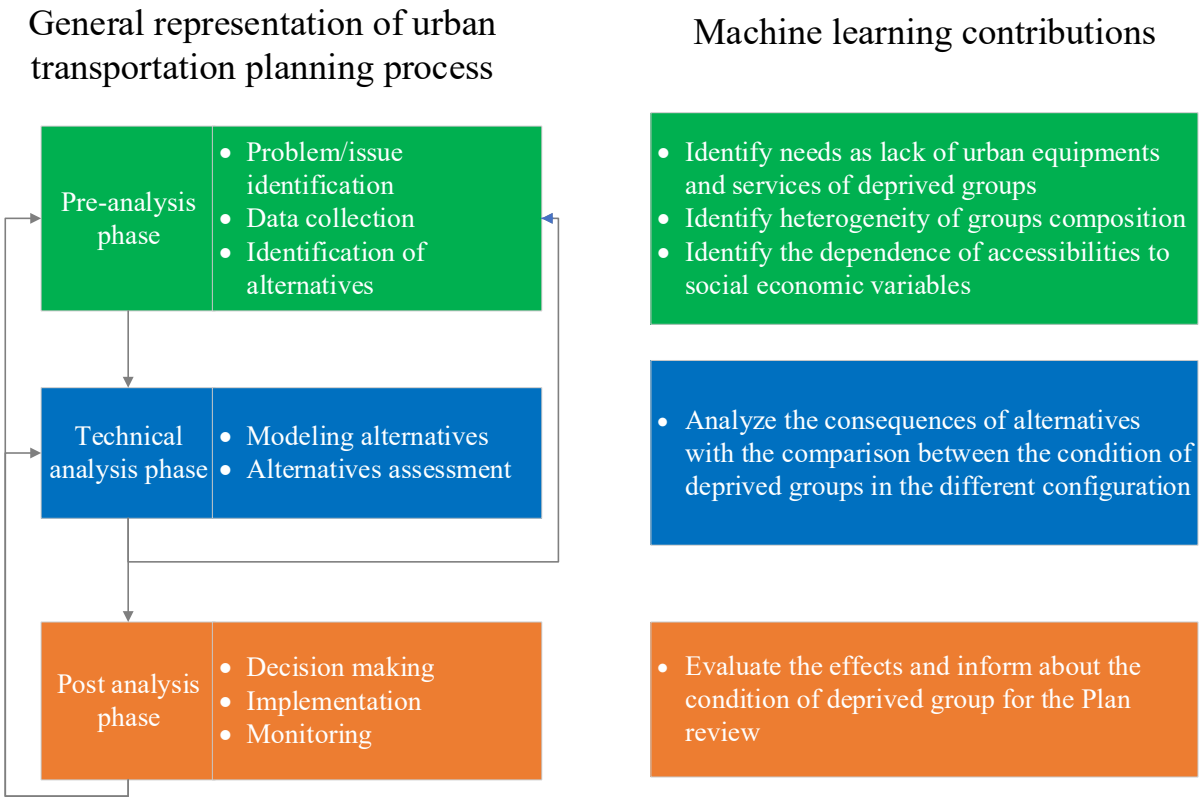
The algorithms are respectively, classification and regression (For more information see section 5.2.).

6.2 What is the relation of ML contribution with the existing transportation planning structure?

Planning as a process can serve mainly for (i) providing information to decision makers about consequences of alternatives in courses of an action; (ii) educating planners, decision makers, and the general public and; (iii) responding to regulations (PAS, 2004). In this sense, transportation planning can be part of learning process about social and spatial issues, related to the fairness distributions of opportunities across the territory.

ML techniques can support transportation planning in different context of analysis such as: (i) using regression for predicting of travel model choice (HAGENAUER; HELBICH, 2017; ZHU et al., 2017) and travel time (GAL et al., 2014); (ii) to quantify interdependence between land use and transport delivery (HU et al., 2016). This work proposes the use of those techniques for identifying patterns and relationships about inequalities. To summarize the findings and provide a useful reference to practitioners, a framework is proposed (Figure 31).

Figure 31 – Framework of the contribution of ML techniques to transportation planning



Source: The first column of "General representation of urban transportation planning process" is adapted from (PAS, 2004) the "Machine learning contributions" are from author

Pre-analysis phase

This stage starts with the problem and issue identification (PAS, 2004). The identification of opportunities and problems should be broadly enough to do not constraints the potential solutions (PAS, 2004). The formulation of goals and objectives defines the desired states toward the planning process should guide the urban area (PAS, 2004). They are derived from the values and paradigms already existing in the community and in the planner's rationality. This part of the planning process that should be more comprehensive and have clear accessibility and inequalities objectives. In this sense, the previous analysis using ML techniques can support a previous diagnosis to identify critical areas and groups. However, to make the analysis feasible, it is necessary to collect, sometimes, a huge amount of data. The data collection is also part of this diagnosis and supports the ML application. Most part of methods aims at gathering travel or traveler information (PAS, 2004). For more comprehensive analysis, also land use information can support the models. Based on this, the alternatives should be identified and described. They should present different situations to cover the initial gaps and drawbacks identified in the diagnosis.

Technical analysis phase

In this phase, some mathematical models are applied (PAS, 2004). Some mathematical models of travel behavior, land use-activity system models and models of impact prediction are examples already applied in practice (ORTUZAR; WILLUMSEN, 2011). The Urban Transportation Model System (UTMS), also namely as four-step model, developed in studies in Detroit and Chicago involves the stages of: (i) generation, which aims at predicting the number of trips produces and attracted to each traffic analysis; (ii) trip distribution, which predicts where the trips go; (iii) mode choice, which addresses how the trips are made by each mode of travel; (iv) trip assignment, which predicts the route used by the trips. All this mathematical background already supports the decision making of transportation plans. However, according to the objectives of promoting equity and reducing inequalities, these methods should focus to predict and deliver travel and opportunities considering the existing critical areas and groups in the urban region. This stage also involves the alternatives assessment and selection of an alternative to being implemented. In this sense, the application of ML techniques can support with information the comparison of different alternative's effects.

Post analysis phase

This phase implements the chosen alternative and monitors the system performance. The already acquired knowledge aims at informing besides the planner, also the community. The information should be updated to the plan review. The programs for monitoring the effects of the implementation should also be implemented.

7 Conclusions

This work aims at investigating the benefits of using ML techniques to identify, describe patterns and investigates the relationships in accessibility of spatial inequalities considering accessibility and socioeconomic dataset. The literature shows that few accessibility and equity assessment are considered in transportation planning practice. The analysis of nine Brazilian plans reveals that no measure is presented to assess spatial inequalities. In this context, works are encouraged to show the value of some techniques and innovative approach to support the development of transportation plans. Some supervised and unsupervised algorithms were applied on accessibility and socioeconomic dataset. The main goal is to explore applications of ML techniques and provide useful insights for further developments of spatial inequalities assessment in transportation practice. This work seeks to start an awareness of the application of ML to assess inequalities, especially in Brazilian reality.

The accessibility measure used was the cumulative opportunities to evaluate the potential number of urban equipment to be reached given a travel time. The dataset used was dated from 2000 and 2010, with no changes in the offer of urban facilities. Because of this, the changes in the accessibility indicators are caused only by the differences in the travel time of new metro lines. The income was taken as a reference to feature the deprived group composition.

The unsupervised learning focus on grouping the instances based on similarities among accessibility level of different urban facilities. The analysis focuses to feature the heterogeneity, especially of the low-income groups. The dimensionality reduction revels two main groups on both years, the urban fringe in all regions of the city and the urban centrality. The last, present negative relations with values in the east region of the city. In the clustering analysis, it was possible to realize the distinguished condition between the peripheries in the city. In the east zone, a cluster of low-income does not present low accessibility to all facilities. The most critical access to urban equipment for low-income population are hospital and culture facilities.

The supervised learning revels relations between income and the accessibility. The income variable is positively related to accessibility to culture facilities, therefore, as higher is that accessibility, higher is the income.

Further works could be developed comparing the patterns acquired using different techniques, such as those using and not using the spatial statistics into the model formulation. Other datasets for capturing other categories of inequalities, such as (i) social inequalities, with gender, elderly among others; (ii) transportation mode inequalities, comparing the public and

private, or bicycle and walk. Regarding the accessibility calculation, the network could consider also the bus changes over the years. Besides that, competition measures, vacancies and other restrictions can enhance the complexity of the indicators.

References

ALPAYDIN, E. **Introduction to Machine Learning**. 2. ed. London: The MIT Press, 2010. 579p.

ANSELIN, L., SYABRI, I., KHO, Y., 2006. GeoDa: An introduction to spatial data analysis. **Geographical Analysis**. 38, 5–22.

BATTY, M; TORRENS, P; 2001. **Modeling complexity: the limits to prediction**. (CASA Working Papers 36). Centre for Advanced Spatial Analysis: London, UK.

BANISTER, D. **Transport planning**. 2. ed. London and New York: Spon Press, 2002. 328p.

BANISTER, D. The sustainable mobility paradigm. **Transport Policy**, v. 15, n. 2, p. 73–80, 2008.

BECCARI, B. A Comparative Analysis of Disaster Risk, Vulnerability and Resilience Composite Indicators. **PLoS Currents Disasters**, v. 14, n. 1, 2016.

BERTOLINI, L.; CLERCQ, F.; KAPOEN, L. Sustainable accessibility : a conceptual framework to integrate transport and land use plan-making. Two test-applications in the Netherlands and a reflection on the way forward. **Transport Policy**, v. 12, p. 207–220, 2005.

BOISJOLY, G.; EL-GENEIDY, A. M. How to get there? A critical assessment of accessibility objectives and indicators in metropolitan transportation plans. **Transportation Research Board 96th Annual Meeting**, v. 55, n. February, p. 38–50, 2017a.

BOISJOLY, G.; EL-GENEIDY, A. M. The insider: A planners' perspective on accessibility. **Journal of Transport Geography**, v. 64, n. August, p. 33–43, 2017b.

BRUNSDON, C.; FOTHERINGHAM, a S.; CHARLTON, M. E. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. **Geographical Analysis**, v. 28, n.

4, p. 281–298, 1996.

CHRISTOFOLETTI, A. **Modelagem de sistemas ambientais**. São Paulo: Blucher, 1999. 236p.

CUTTER, S. L.; BORUFF, B. J.; SHIRLEY, W. L. Social Vulnerability to Environmental Hazards. **Social Science Quarterly**, v. 84, n. 2, p. 242–261, 2003.

DELBOSC, A.; CURRIE, G. Using Lorenz curves to assess public transport equity. **Journal of Transport Geography**, v. 19, n. 6, p. 1252–1259, 2011.

DHANACHANDRA, N.; MANGLEM, K.; CHANU, Y. J. Image Segmentation Using K - means Clustering Algorithm and Subtractive Clustering Algorithm. **Procedia Computer Science**, v. 54, p. 764–771, 2015.

DU, H.; MULLEY, C. Relationship Between Transport Accessibility and Land Value: Local Model Approach with Geographically Weighted Regression. **Transportation Research Record: Journal of the Transportation Research Board**, v. 1977, n. 1, p. 197–205, 2007.

EIBE, F.; HALL, M. A.; W., I. H. The WEKA Workbench. In: **Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"**. 4 ed. Morgan Kaufmann, 2016.

EL-GENEIDY, A.; BULIUNG, R.; DIAB, E.; VAN LIEROP, D.; LANGLOIS, M.; LEGRAIN, A. Non-stop equity: Assessing daily intersections between transit accessibility and social disparity across the Greater Toronto and Hamilton Area (GTHA). **Environment and Planning B: Planning and Design**, v. 43, n. 3, p. 540–560, 2015.

ÉRNICA, M.; BATISTA, A. A. G. A escola, a metrópole e a vizinhança vulnerável. **Cadernos de Pesquisa**, v. 42, n. 146, p. 640–666, 2012.

FEITOSA, F.; CÂMARA, G.; MONTEIRO, A. De Conceitos a Medidas Territoriais: A Construção de Índices Espaciais de Segregação Urbana. **Geoinformação em Urbanismo: Cidade Real vs. Cidade Virtual**, p. 86–105, 2007.

FORKUOR, G.; HOUNKPATIN, O. K. L.; WELP, G.; THIEL, M. High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. **Plos One**, v. 12, n. 1, 2017.

FOTH, N., MANAUGH, K., EL-GENEIDY, A.M., 2013. Towards equitable transit: Examining transit accessibility and social need in Toronto, Canada, 1996-2006. **Journal of Transport Geography**. 29, 1–10.

FUSCO, G.; COLOMBARONI, C.; ISAENKO, N. Short-term speed predictions exploiting big data on large urban road networks. **Transportation Research Part C: Emerging Technologies**, v. 73, p. 183–201, 2016.

GAL, A.; MANDELBAUM, A.; SCHNITZLER, F.; SENDEROVICH, A.; WEIDLICH, M. Traveling time prediction in scheduled transportation with journey segments. **Information Systems**, v. 64, p. 266–280, 2014.

GAN, G.; MA, C.; WU, J. **Data Clustering: Theory, Algorithms, and Applications**. ASA-SIAM Series on Statistics and Applied Probability. 2007. 488p.

GARCIA, C. S. H. F. **Strategic assessment of accessibility on urban mobility networks strategic assessment of accessibility on urban mobility networks**. Doctoral thesis. Universidade de Lisboa, 2016.

GEURS, K. T.; VAN WEE, B. Accessibility evaluation of land-use and transport strategies: review and research directions. **Journal of Transport Geography**, v. 12, n. 2, p. 127–140, 2004.

GIL, A.C. **Métodos e técnicas de pesquisa social.** 6ª Edição. São Paulo: Editora Atlas. 2008. 200p.

GUZMAN, L. A.; OVIEDO, D.; RIVERA, C. Assessing equity in transport accessibility to work and study: The Bogotá region. **Journal of Transport Geography**, v. 58, p. 236–246,

2017.

HAGENAUER, J.; HELBICH, M. A comparative study of machine learning classifiers for modeling travel mode choice. **Expert Systems with Applications**, v. 78, p. 273–282, 2017.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining : Concept and Tecniques**. Elsevier. 2012. 740p.

HARRINGTON, P. **Machine Learning in Action**. New York: Manning. 2016. 382p.

HEUNG, B.; HO, H. C.; ZHANG, J.; KNUDBY, A.; BULMER, C. E.; SCHMIDT, M. G. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. **Geoderma**, v. 265, p. 62–77, 2016.

HIGGS, G.; LANGFORD, M.; NORMAN, P. Accessibility to sport facilities in Wales: A GIS-based analysis of socio-economic variations in provision. **Geoforum**, v. 62, p. 105–120, 2015.

HU, N.; LEGARA, E. F.; LEE, K. K.; HUNG, G. G.; MONTEROLA, C. Impacts of land use and amenities on public transport use, urban planning and design. **Land Use Policy**, v. 57, p. 356–367, 2016.

HUANG, X.; YE, Y.; XIONG, L.; LAU, R. Y. K.; JIANG, N.; WANG, S. Time series k-means: A new k-means type smooth subspace clustering for time series data. **Information Sciences**, v. 367–368, p. 1–13, 2016.

IBES, D. C. A multi-dimensional classification and equity analysis of an urban park system: A novel methodology and case study application. **Landscape and Urban Planning**, v. 137, p. 122–137, 2015.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Censo Demográfico 2000: Agregado por Setores Censitários dos Resultados do Universo**. 2003.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Base de informações do Censo Demográfico 2010 : Resultados do Universo por setor censitário**. 2011.

JIN, S.; JUNG, B. C. Sample Based Algorithm for k-Spatial Medians Clustering. **The Korean Journal of Applied Statistics**, v. 23, n. 2, p. 367–374, 2010.

JOSEPH, J.; TORNEY, C.; KINGS, M.; THORNTON, A.; MADDEN, J. Applications of machine learning in animal behaviour studies. **Animal Behaviour**, v. 124, n. December, p. 203–220, 2016.

KOGA, D. Território entre pobreza e exclusão social. In: **Medidas de cidades: entre territórios de vida e territórios vividos**. 2 ed. São Paulo: Cortez, 2011. 331p.

KWAN, M. P.; WEBER, J. Scale and accessibility: Implications for the analysis of land use-travel interaction. **Applied Geography**, v. 28, n. 2, p. 110–123, 2008.

LOGIODICE, P. C. R. **Avaliação de impacto do aumento da acessibilidade a partir de simulações da malha metroviária**. Trabalho de formatura. Universidade de São Paulo. 2017.

LUCAS, K. Transport and social exclusion: Where are we now? **Transport Policy**, v. 20, p. 105–113, 2012.

LUCAS, K.; VAN WEE, B.; MAAT, K. A method to evaluate equitable accessibility: combining ethical theories and accessibility-based approaches. **Transportation**, 2015.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability**, v. 1, n. 233, p. 281–297, 1967.

MAHRSI, M. K. El; CÔME, E.; OUKHELLOU, L.; VERLEYSEN, M. Clustering Smart Card Data for Urban Mobility Analysis. v. 18, n. 3, p. 712–728, 2017.

MANAUGH, K.; BADAMI, M. G.; EL-GENEIDY, A. M. Integrating social equity into urban transportation planning: A critical evaluation of equity objectives and measures in transportation plans in north america. **Transport Policy**, v. 37, p. 167–176, 2015.

MARQUES, E. C. Espaço e grupos sociais na virada no século XXI. In: MARQUES, E. C.;

TORRES, H. (Orgs). **São Paulo: segregação, pobreza e desigualdades sociais**. São Paulo: Editora Senac, 2005. 329p.

MASCARO, J.; ASNER, G. P.; KNAPP, D. E.; KENNEDY-BOWDOIN, T.; MARTIN, R. E.; ANDERSON, C.; HIGGINS, M.; CHADWICK, K. D. A tale of two "Forests": Random Forest machine learning aids tropical Forest carbon mapping. **PLoS ONE**, v. 9, n. 1, p. 12–16, 2014.

MIGUEL-HURTADO, O.; GUEST, R.; STEVENAGE, S. V.; NEIL, G. J.; BLACK, S. Comparing Machine Learning Classifiers and Linear/Logistic Regression to Explore the Relationship between Hand Dimensions and Demographic Characteristics. **Plos One**, v. 11, n. 11, p. e0165521, 2016.

MILLER-COLEMAN, R. L.; DODSWORTH, J. A.; ROSS, C. A.; SHOCK, E. L.; WILLIAMS, A. J.; HARTNETT, H. E.; MCDONALD, A. I.; HAVIG, J. R.; HEDLUND, B. P. Korarchaeota diversity, biogeography, and abundance in Yellowstone and Great Basin hot springs and ecological niche modeling based on machine learning. **PLoS ONE**, v. 7, n. 5, 2012.

MOKDAD, F.; HADDAD, B. Improved Infrared Precipitation Estimation Approaches based on k-means Clustering: Application to North Algeria using MSG-SEVIRI Satellite Data. **Advances in Space Research**, 2017.

NEUTENS, T. Accessibility, equity and health care: Review and research directions for transport geographers. **Journal of Transport Geography**, v. 43, p. 14–27, 2015.

NEUTENS, T.; SCHWANEN, T.; WITLOX, F.; DE MAEYER, P. Equity of urban service delivery: A comparison of different accessibility measures. **Environment and Planning A**, v. 42, n. 7, p. 1613–1635, 2010.

OECD (Organisation for Economic Co-operation and Development). **Handbook on Constructing Composite Indicators: Methodology and User Guide** (Vol. 3). 2008. 162p.

ORTUZAR, J. de D.; WILLUMSEN, L. G. **Modelling transport**. 4 ed. UK: Wiley. 2001. 608p.

PÁEZ, A.; SCOTT, D. M.; MORENCY, C. Measuring accessibility: Positive and normative

implementations of various accessibility indicators. **Journal of Transport Geography**, v. 25, p. 141–153, 2012.

PAS, E. The urban transportation planning process. In: HANSON, S. (Orgs.). **The Geography of urban transportation**. London and New York: The Guilford Press, 2004. p. 432.

PEREIRA, R. H. M.; BANISTER, D.; WESSEL, N. Distributional effects of transport policies on inequalities in access to opportunities in Rio de Janeiro. Forthcoming, **2017**.

PEREIRA, R. H. M.; SCHWANEN, T.; BANISTER, D. Distributive justice and equity in transportation. **Transport Reviews**, v. 37, n. 2, p. 170–191, 2017.

PEROSA, G. S.; LEBARON, F.; LEITE, C. K. da S. O espaço das desigualdades educativas no município de São Paulo. **Pro-Posições**, v. 26, n. 2, p. 77, 2015.

RAMJERDI, F. Equity measures and their performance in transportation. **In Transportation Research Board 85th Annual Meeting**, n. 1983, p. 67–74, 2006.

RAWLS, J. **A theory of justice: revised edition**. Cambridge: Harvard University Press. 2006. 561p.

REMITA, M. A.; HALIOUI, A.; DIOUARA, A. A. M.; DAIGLE, B.; KIANI, G.; DIALLO, A. B. CASTOR: A machine learning platform for reproducible viral genome classification. **bioRxiv**, p. 1–7, 2016.

RIBEIRO, J. Predicting Risk of Suicide Attempts over Time through Machine Learning. **Clinical Psychological Science**, v.5 , n. 3, 457-469, 2017.

RUS, G.; KRUSE, R. Machine learning methods for spatial clustering on precision agriculture data. **Frontiers in Artificial Intelligence and Applications**, v. 227, p. 40–49, 2011.

SECRETARIA NACIONAL DE TRANSPORTE E DA MOBILIDADE URBANA (SEMOB); MINISTÉRIO DAS CIDADES. **Caderno de referência para elaboração de plano de mobilidade urbana**. 2015. 238p.

SESTER, M., 2000. Knowledge acquisition for the automatic interpretation of spatial data. **International Journal of Geographic Information Science**, vol. 14, p. 1-24.

SILVA, C.; BERTOLINI, L.; TE BRÖMMELSTROET, M.; MILAKIS, D.; PAPA, E. Accessibility instruments in planning practice: Bridging the implementation gap. **Transport Policy**, v. 53, n. July 2015, p. 135–145, 2017.

SIQUEIRA-GAY, J.; GIANNOTTI, M. A.; SESTER, M. Learning spatial inequalities : a clustering approach. In: Proceedings of the XVIII Brazilian Symposium of Geoinformatics, Salvador. 2017.

SPOSATI, A.; MONTEIRO, M. Exclusão/Inclusão social nos distritos. In: **Desigualdades nos territórios da cidade: métricas sociais intraurbanas em São Paulo**. São Paulo: EDUC, 2017. p. 128.

SUBASI, A.; KEVRIC, J.; ABDULLAH CANBAZ, M. Epileptic seizure detection using hybrid machine learning methods. **Neural Computing and Applications**, 2017.

TOMASIELLO, D. B. **Modelos de rede de transporte público e individual para estudos de acessibilidade em são paulo**. 2016. Dissertação de mestrado. Universidade de São Paulo, 2016.

VAN WEE, B. Accessible accessibility research challenges. **Journal of Transport Geography**, v. 51, p. 9–16, 2016.

WEE, B.; GEURS, K. Discussing equity and social exclusion in acessibility evaluations. **European Journal of Transport and Infrastructure Research**, v. 11, n. 4, p. 350–367, 2011.

WERDER, S.; KIELER, B., SESTER, M., 2010. Semi-Automatic Interpretation of Buildings and Settlement Areas in User-Generated Spatial Data. In: Proceedings of the 18th International Conference on Advances in Geographic Information Systems, San Jose, CA.

WEI, Y. D. Spatiality of regional inequality. **Applied Geography**, v. 61, p. 1–10, 2015.

WIELAND, R.; KERKOW, A.; FRÜH, L.; KAMPEN, H.; WALTHER, D. Automated feature selection for a machine learning approach toward modeling a mosquito distribution. **Ecological Modelling**, v. 352, p. 108–112, 2017.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining Practical Machine Learning Tools and Techniques.** 3 Ed**.** Morgan Kaufmann. 2011. 665p.

WONG, K. K. L.; WANG, L.; WANG, D. Recent developments in machine learning for medical imaging applications. **Computerized Medical Imaging and Graphics**, v. 57, p. 1–3, 2017.

YAMAMOTO, Y.; SAITO, A.; TATEISHI, A.; SHIMOJO, H.; KANNO, H.; TSUCHIYA, S.; ITO, K.; COSATTO, E.; GRAF, H. P.; MORALEDA, R. R.; EILS, R.; GRABE, N. Quantitative diagnosis of breast tumors by morphometric classification of microenvironmental myoepithelial cells using a machine learning approach. **Scientific Reports**, v. 7, n. April, 2017.

ZAKI, M. J.; MEIRA, M. J. **Data Mining and Analysis: Fundamental Concepts and Algorithms**. Cambrigde University Press. 2013. 607p.

ZHU, Z.; CHEN, X.; XIONG, C.; ZHANG, L. A mixed Bayesian network for two-dimensional decision modeling of departure time and mode choice. **Transportation**, 2017.