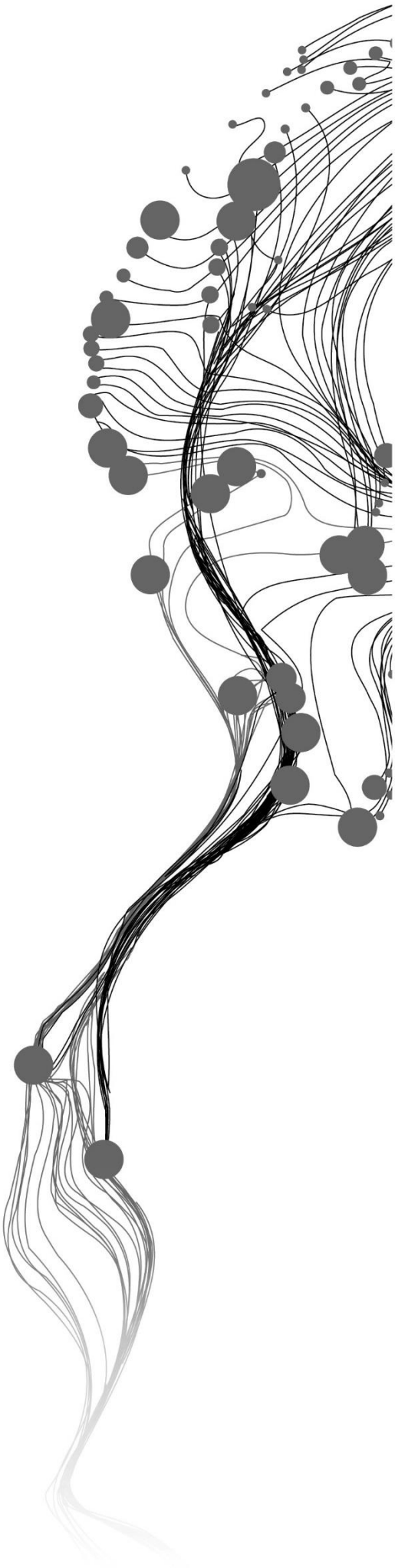
THE DIVERSITY OF DEPRIVED AREAS: APPLICATIONS OF UNSUPERVISED MACHINE LEARNING AND OPEN GEODATA

LORRAINE TRENTO OLIVEIRA
August 2021

SUPERVISORS:
Dr. M. Kuffer
Dr. N. Schwarz



THE DIVERSITY OF DEPRIVED AREAS: APPLICATIONS OF UNSUPERVISED MACHINE LEARNING AND OPEN GEODATA

LORRAINE TRENTO OLIVEIRA
Enschede, The Netherlands, August 2021

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Urban Planning and Management

SUPERVISORS:
Dr. M. Kuffer
Dr. N. Schwarz

THESIS ASSESSMENT BOARD:
Prof. Dr. R.V. Sliuzas (Chair)
Dr. J. C. Pedrassoli (External Examiner)

ABSTRACT

The rapid growth of deprived areas in Low- to Middle-Income Countries (LMICs) is a great urban challenge that requires consistent and updated information specifically about the physical and living conditions in such areas. When available, census data provides detailed socioeconomic information at the household level, but it is resource-intensive and aggregated at area enumeration levels, masking important spatial differences. The increasing availability of very-high-resolution (VHR) imagery has boosted the publication of remote sensing (RS) mapping studies. Yet, only a few studies focus on characterizing the deprived areas because RS studies mostly focus on common physical morphologies of deprived areas and, thus, oversimplify their features. Moreover, even fewer studies address city-wide analysis due to the VHR acquisition and computational costs. To address these gaps, this research explores the intra-urban diversity of deprived areas using solely open RS and geo-data sources for São Paulo, Brazil. More specifically, it makes use of the potential of unsupervised machine learning (ML) models to capture intra-urban differences in deprived areas. First, based on literature, a pool of GIS- and RS-based features is developed to derive morphological and environmental characteristics of the study area. Next, a k-means clustering model is trained while running several optimisation experiments, including feature selection techniques and the inclusion of census-derived features. A feature importance tool is coupled to the k-means model to stress the relevance of specific features for each of the four resulting cluster types. The first cluster, "Infant settlements in open spaces", is characterised by low accessibility to services and infrastructures, very sparse occupation and presence of vegetation. The second, "Unordered and poorly consolidated settlements", is marked by steep terrain, lack of infrastructure and relatively low population densities. The third, "Less deprived settlements connected to non-residential areas", is identified mainly by more regular layout and mixed land uses. And the fourth, "Densely urbanized and mature settlements with irregular layout", is highly influenced by built-up density and complex (slum-like) morphology. The qualitative validation evinces that the unsupervised model successfully captures the intra-urban diversity of deprived settlements in São Paulo, stressesing higher precariousness for the second identified cluster. The assessment demonstrates that the proposed approach can be an alternative to current characterization studies using solely open data, providing a gridded output that supports the scalability of the model and its transferability to different cities. The cluster types are profiled and can be comprehensively used for the decision-making process. Moreover, this study offers an additional and important perspective to the characterization analysis with the census-derived features. For further research, the utmost suggestion is transferring the approach to other Brazilian cities and scaling it to a regional and national scale.

Keywords: deprived areas; open data; GIS; urban remote sensing; census; unsupervised learning; clustering algorithm; Low-to Middle-Income Countries

ACKNOWLEDGEMENTS

I came to ITC as an architect and urban planner, with no knowledge of remote sensing or programming skills. I am used to peer-pressure, work plans, deadlines and other stressful situations in the academic environment, but nothing like this experience in a foreign land, living alone in a hotel room, concerned with what is happening with my loved ones in the pandemic and constantly questioning if I would make it at all. Yet, I made it and I feel so full of gratitude while writing the final words of this incredible journey. Therefore, I start this section by thanking God for strengthening my heart, for His faithfulness, and for surprising me.
I would like to express my deepest gratitude and appreciation to my research supervisors, Dr Monika Kuffer and Dr Nina Schwarz. Your patience, constant encouragement, insightful guidance and valuable critiques had paramount importance not only to my work but my life. You are the full package: excellent technical skills and professionalism combined with your incredibly humble and caring personalities. You built confidence in me when it was shaken; you pushed my boundaries; gave me motivation and taught me so much. Without your support, I would not have reached my goal. I will be forever thankful for you.
My strong thankfulness to Alexandra Peixe for your professional feedback, insights and unique suggestions on the validation procedure. Hopefully we can work together on the further deployments of this MSc thesis results in Brazil. Also, I wish to thank Dr Caroline Gevaert for great technical assistance on the proposal phase, indicating possible uncertainties sources and developing fronts for the work. I would also like to give a special thanks to Dr Flávia Feitosa, UFABC, for providing amazing consultancy during the construction of this thesis. From data, features and modelling aspects, she offered valuable suggestions that were very useful for this study. Also, great thanks to my internship supervisor, Frank van Rijn, for his encouragement and valuable lessons, which I applied here. High appreciation to Dr Raian Maretto for all advice and support regarding machine learning tools and techniques. A final special thanks to Dr Richard Sliuzas for the great detailed feedback and discussion on the proposal defence and mid-term assessments.
I wish to acknowledge the support of my ITC friends, my ultimate sincere supporters. I wish I could mention all of you here, expressing how each of you immensely helped me. You made this journey way more fun; you gave me great thesis advice and true understanding for when I felt frustrated or anxious. I am still amazed on how much of a family we have built here, even with totally different cultures, languages and values. It would take me decades to learn all I learned here with you about kindness, empathy and true support. I also would like to extend it to my Brazilian friends, that even with all the distance, made them feel close to me, constantly checking in and giving words of encouragement during these two years.
Last but not least, it is wholeheartedly appreciated the great support of my loving family for all support. Immense gratitude for my father and his utmost example of commitment and assertiveness, my mother and her total devotion to make me thrive, my sister and her unfailing presence, my Godmother and her constant advice and care, my boyfriend and his unconditional encouragement. Thank you for supporting my decision to pursue this degree and participating throughout the process. I love you all so much.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AGSN	Aglomerados Subnormais *(Subnormal settlements)*
CRAN	Comprehensive R Archive Network
DEM	Digital Elevation Model
DMP	Data Management Plan
DRM	Disaster Risk Management
EDA	Exploratory Data Analysis
EO	Earth Observation
GDP	Gross Domestic product
GIS	Geographic Information System
GLCM	Gray-Level Co-occurrence Matrix
HOT	Humanitarian OpenStreetMap Team
IBGE	Instituto Brasileiro de Geografia e Estatística *(Brazilian institute of Geography and Statistics)*
IDEAMAPS	Integrated Deprived Area Mapping System
LMIC	Low- to Middle-Income Country
LSM	Landscape Metrics
MAUP	Modifiable Areal Unit Problem
ML	Machine Learning
MSD	Mean-square distance
NDVI	Normalized Difference Vegetation Index
NTL	Night Time Light
OSM	Open Street Map
PAN	Panchromatic
PCA	Principal Component Analysis
RS	Remote Sensing
SDG	Sustainable Development Goal
SQL	Structured Query Language
USGS	United States Geological Survey
VHR	Very High Resolution
VIIRS	Visible Infrared Imaging Radiometer Suite
WCSS	Within-Cluster Sum of Square
WDPA	World's Database of Protected Areas
WSF	World Settlement Footprint

1. INTRODUCTION

Many cities of the world face unprecedented growth of urban poor populations and acquiring information about them is a constant demand. Current approaches rely on census or Earth Observation (EO) data, subject to a series of constraints that reflects the complexity of deprived areas. Conversely to the slum concept, deprived areas reflect the living conditions beyond the household level, incorporating multiple social, environmental and ecological influences at the neighbourhood level. Meanwhile, efforts on open-source infrastructures and machine learning (ML) studies indicate new ways of addressing existing data and methodological issues. This research aims to generate useful insights into the diversity of deprived areas by exploring the potential of unsupervised ML models and multiple open geodata sources to account for the various living standards of deprivation in cities.

1.1. Background and justification

Recognizing the emergence of urban areas with poor living conditions is crucial for the sustainability of cities across the globe. According to United Nations-Habitat (2015), 1 out of 8 people in the world lives in slum-like settlements, facing a lack of basic service provision, inadequate housing and high poverty levels. In Low- to Middle-Income (LMIC) cities, the situation is even more aggravated due to unprecedented urban growth rates and weak institution's frameworks (Bastos da Cunha et al., 2015). The local governments do not fully acknowledge the existence of poor settlements, resulting in urban plans that focus mainly on the formal population (Lucci, Bhatkal, & Khan, 2016). This lack of planning capacity, combined with the enormous housing demands of the urban population, boost the development of deprived areas (Kohli, Sliuzas, Kerle, & Stein, 2012). In this context, several institutions for the past decades emphasize poverty and deprivation as main urban challenges to be tackled and propose policies to improve the living conditions of deprived neighbourhoods, e.g., in the SDG 11[1] (United Nations, 2018). Yet, there are still limitations to address inequalities and poverty in urban areas, regarding uncertainties of definitions, availability of data and technology solutions.

According to UN-Habitat (2015), the traditional and widely used concept of 'slums' encompasses one of the most extreme forms of deprivation and relative poverty worldwide. Their definition infers the classification of slum households when there is a lack of access to at least one of the following dimensions: water safety, acceptable sanitation, durable housing, sufficient living areas, and tenure security. It defines a poor and unsafe living environment with no or inadequate basic service provision. Despite the undeniable significance of the term, it does not fully portray how diverse slums are. By limiting the spatial perspective to the household level, policies and plans fail to address the area characteristics of these settlements and the multiple conditions they face (Olthuis, Benni, Eichwede, & Zevenbergen, 2015). In addition, the 'slum' concept is frequently used to express homogeneous appearance in terms of morphology, environment conditions and socioeconomic status (Roy, Lees, Palavalli, Pfeffer, & Sloot, 2014), driving most studies to rely on the existing commonalities among slum settlements. However, on the ground, the manifestation of poverty and deprivation differs locally with heterogeneous characteristics at the intra-city and inter-city levels (Engstrom, Ofiesh, Rain, Jewell, & Weeks, 2013). This plurality hampers the creation of universal definitions and methods for capturing deprived areas, that account for the complexity of urban poverty, which creates a wicked problem (Gevaert, Kohli, & Kuffer, 2019). While most studies develop efficient methods to capture deprivation by oversimplifying it as a dichotomy of slum versus non-slum areas, they overlook its

[1] SDG target: "By 2030, *ensure access for all to adequate, safe and affordable housing and basic services and upgrade slums*".

diverse nature, hampering the creation of more effective and contextualised pro-poor policies (Thomson et al., 2020).

Recent studies provided empirical evidence on the prevailing characteristics of slums about their housing and environment conditions that are easily recognized in the urban landscape (Lilford et al., 2019; H. Taubenböck & Kraff, 2014). Acknowledged by the multiplicity and the dynamicity of deprived areas, there is an urgent need for contextual, up-to-date, and detailed spatial data to characterize the complexity of these neighbourhoods (Kohli, Warwadekar, Kerle, Sliuzas, & Stein, 2013). However, especially in LMICs, several limitations are encountered, considering the low resources flow allocated to address such urban issues (Lucci et al., 2016). Given the advancements in EO data and infrastructure in the past decades, recent research investigates its availability to study the morphology of deprived settlements as an alternative to census survey data, mostly dedicated to identifying and delineating slum boundaries (Taubenböck, Kraff, & Wurm, 2018). Though finding their location is indisputably necessary, only a few studies try to characterize them and capture their diversity, acquiring contextual knowledge to customize plans and place-based interventions (Kuffer, Pfeffer, Sliuzas, & Baud, 2017). Therefore, more research is required to provide detailed information considering the local living conditions of deprived areas. The following section introduces the issues with the available methods and states the need for alternative approaches on urban deprivation for LMICs.

1.2. Research problem

The spatial differences of deprived areas need to be effectively captured and to describe their local living conditions (Kuffer, Orina, & Sliuzas, 2017). The present research acts towards the characterization of deprived areas and the gaps found in the literature. The first gap concerns data acquisition and availability. Commonly, fieldwork produces information at the household level, which creates data-rich repositories that are commonly aggregated at administrative units (Mahabir, Agouris, Stefanidis, Croitoru, & Crooks, 2020). The data aggregation process can hide important characteristics by generating spatial generalization biases because large administrative units are assigned with the same average value, creating false representations of the reality on the ground (Martínez, Pfeffer, & Baud, 2016). Work as Carr-Hill (2013) and Gevaert, Kohli and Kuffer (2019) also plead that census can omit large population categories, undercounting them by design (e.g., excluding homeless or refugees) or by practical hindrances. Besides, field data is temporally inconsistent, time and human-resource consuming (Thomson et al., 2020). Hence, when available in LMICs, they space out over a decade or more (Owen & Wong, 2013). The other option is EO data. The increasing availablity of Very-High-Resolution (VHR) imagery – high temporal and sptial resolution - allows the urban landscape to be analysed at a fine geographical scale, following the dynamic nature of deprived areas (Duque, Patino, & Betancourt, 2017). The gridded data can be aggregated in different spatial units without following the traditional administrative boundaries from census data, preventing neighbourhood stigmatization (Abascal et al., 2021). Nonetheless, even though the costs have been reducing throughout the years, the acquisition of VHR images for a large spatial coverage area, for instance, at city scale, is a cost-prohibitive factor (Aguilar & Kuffer, 2020). Until the present date, the high costs of VHR restricted their use to research advancements by scholars and high-income governments (Gram-Hansen et al., 2019). In this context, the emergency of open geodata poses as an alternative to such issues especially in LMICs (Chakraborty, Wilson, Sarraf, & Jana, 2015). In crowd-sourced platforms such as Open Street Map (OSM), geospatial data production becomes decentralized, easily up-to-dated, and public engagement due to transparency (Grippa et al., 2018). However, only a few urban studies have comprehensively explored them, primarily because of quality, resolution and integration issues that demand extra time on data preparation, discouraging research on open data frameworks (Mahabir et al., 2020).

Besides data resource constraints, the literature also highlights methodological gaps. Given the trending advancements of ML algorithms, with robust methods and efficient performance, most EO studies extract information from VHR images by recognizing physical attributes and performing binary classification of

slum and non-slum areas (Gevaert et al., 2019). Yet, these observable attributes are not universal, not even in the same city (Kohli, Stein, & Sliuzas, 2016). As the models do not consider the area' specificities and establish one label for all settlements in the study area, inaccuracies are generated to the model (Duque et al., 2017; Kit, Lüdeke, & Reckien, 2012). Meanwhile, a few recent studies use EO data to characterize the diversity of deprived areas and quantify their morphological features (Ajami, Kuffer, Persello, & Pfeffer, 2019; Kuffer, Pfeffer, et al., 2017). Under the supervised-learning umbrella, they depend on local expert knowledge (secondary, fieldwork or expert data), relying on costly field surveys or expert consultancy for training and reference data (Leonita, Kuffer, Sliuzas, & Persello, 2018). This methodological gap brings unsupervised ML models into sight. They can capture abstract data patterns and attach their similarities in detailed information clusters without spending time and computational resources in training and reference data (Jochem et al., 2020).

In this context, the present study addresses the first gap regarding resource constraints by using open data sources, the second gap regarding the unavailability of large training sets by using unsupervised approaches and combine them into a scalable methodology to characterize deprived areas from city to national scale.

1.3. Research objectives and questions

The overall objective of this research is to explore the potential of unsupervised ML for characterizing deprived areas based on morphological and environmental features using solely open data. It has specific objectives and questions:

1. **Derive spatial features for characterizing deprived areas.**
   a) What are their physical characteristics in literature?
   b) Which GIS-based and RS-based features can be extracted from open and free data sources?
   c) What is the optimal unit of analysis to integrate data with different granularities?

2. **Employ unsupervised learning models to capture the intra-urban diversity of deprived areas.**
   a) Which clustering method is appropriate in terms of efficiency and performance?
   b) What (spatial) clustering patterns emerge?
   c) Which of the features are most significant to describe the deprived areas?
   d) To what extent the inclusion of census data can improve the results?

3. **Analyse the application of the proposed model to the characterization of deprived areas.**
   a) To what extent the selected spatial features are capable of capturing the different deprived areas?
   b) How can the results profiles help shape better pro-poor policies and interventions?
   c) What are the pros and cons of the approach?

1.4. Conceptual framework

Figure 1 summarizes the concepts regarding the data and methodological shortcomings found in literature and how this study dealt with them. Until today, household surveys and EO are the most broadly used methods for data acquisition on deprivation (Lilford et al., 2019). On the one hand, census-based methods are time- and resource-intensive, temporally inconsistent and aggregated at area enumeration levels (Kohli et al., 2012). On the other hand, EO-based approaches provide gridded outputs that incorporate the fuzzy boundaries of deprived settlements but can only cover small urban areas due to the high costs of VHR data (Leonita et al., 2018; Small, 2014), which is a significant shortcoming for several LMIC cities (Thomson et al., 2020). The emergence of open geodata is a promising alternative source that still lacks explorative studies due to their granularity, reliability and integration issues.

Moreover, most EO studies focus on the delineation purposes, using a binary classification of deprived and non-deprived according to their physical morphology (Kohli et al., 2016), which stigmatised entire

neighbourhoods as homogeneously deprived instead of recognizing the variations among them (Ajami et al., 2019). It increases uncertainties of the mapping models and precludes the understanding of socio-spatial inequalities for local and equitable spatial planning (Lucci et al., 2016). Describing deprived areas is more complex than delineating them, but the few recent studies that attempt to capture their characteristics also encounter challenges. The use of supervised classification and regression algorithms that require a large number of high-quality training data, that are not only resource- and computationally-expensive, but also complex to be generated e.g. disagreements on interpreted labels, thus unsustainable for LMICs (Jochem et al., 2020; Kuffer, Pfeffer, et al., 2017). These supervised models rely on training and testing datasets from VHR sources, hindering the ability to scale the existing methods (Gevaert et al., 2019). The proposed framework addresses the challenges above, experimenting with the potential of unsupervised ML models, using open data and a continuous approach to capture the variations of deprivation at the city level.

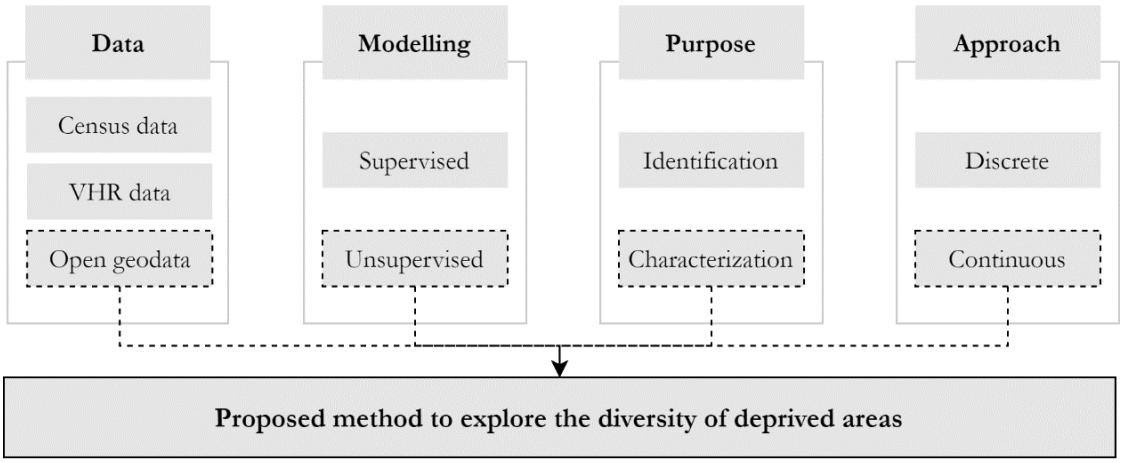


Figure 1. Conceptual framework.

1.5. Research structure

The remainder of this thesis is structured in six chapters. Chapter 2 presents the literature related to the diversity of deprived areas, methodological approaches and unsupervised ML techniques. Chapter 3 describes the research methods, and it is divided into sections. It starts with the description of the study area, the methodology overview and data collection. Next, it follows two main sub-sections encompassing the data processing steps to prepare the model input and the conducted unsupervised machine learning model with validation procedures. Chapter 4 reports the results of the analysis followed by the validation subsection. In Chapter 5, the main findings are discussed, stressing the limitations of the proposed method. Lastly, Chapter 6 provides concluding remarks of the research and directing suggestions for future studies.

2. LITERATURE REVIEW

This chapter reviews the scientific literature and describes the concepts related to the proposed conceptual framework. The first section presents a literature overview on poverty and deprivation and the definition utilized in this research. Then, it provides an overview of the characteristics of deprived areas through spatial lenses. Next, it covers the studies using open and disaggregated data approaches, followed by the review of existing methodology and purposes for capturing deprived areas. Lastly, it presents unsupervised machine learning models, algorithms, and processing techniques.

2.1. The framing of deprived areas

The well-known term 'slum' is commonly used to group a range of neighbourhoods with poor housing structure and environmental conditions and little or no access to basic services (UN-Habitat, 2015a). The definition of the term varies from country to country – even within a single country - along with its physical manifestations (Kuffer et al., 2016). Consequently, the ambiguous concepts result in multiple representations of slums and hamper the existence of a global standard definition for these slum-like settlements (Martínez, Pfeffer, & Baud, 2016). While their dynamic and contextual nature is challenging to develop a global conceptualization for efficient monitoring methods, the standard definitions often ignore their local variations in socioeconomic needs and physical morphology (Gevaert et al., 2019). Thus, when official maps are available in LMICs, they only delineate slum boundaries without accounting for their diversity, which obstructs the efficiency of pro-poor policies such as slum upgrading projects (Kuffer, Pfeffer, et al., 2017).
Acknowledged by this conceptualization dilemma, different institutions attempt to describe groups facing sub-standard living conditions and express urban poverty (Wurm & Taubenböck, 2018). Beyond the primary concept of 'slums' from UN-Habitat (2003), the Sustainable Development Goal (SDG) 11, responsible for fostering policies to identify and quantify spatial inequalities in human settlements, use three different terms - slums, informal settlements and inadequate housing - to contemplate the urban poor in indicator 11.1 (UN-Habitat, 2015b). These definitions seek to facilitate methods to monitor urban poverty and deprivation, but they often identify the groups at the household rather than at the neighbourhood level (Lucci et al., 2016). They ignore the existence of better-off households within precarious neighbourhoods, which leads to entire areas being classified as slums without considering their local environment and the different area-level nuances compared to other neighbourhood communities (Patel, Koizumi, & Crooks, 2014).
A complete review produced by Abascal et al. (2021) identifies the limitations of terminologies at the household level, suggesting the term 'deprived areas', which explicitly express a broader and area-based definition. It refers to neighbourhoods facing deprivation, encompassing social, environmental, and ecological factors that affect the living conditions of an area, beyond the poverty of the single dwelling unit expressed by the 'slum' term. Their definition identifies how these areas are strongly dependent on their local contexts and how heterogenous the intra-city inequality can be, pointing out the demand for tailored methods and better locally target policies (Thomson et al., 2020).
This study acknowledges the terminology dilemmas by stressing the above explanations and employs the term 'deprived area' to refer to settlements in slum-like conditions focusing on a citywide area-based analysis. It is important to state that, regardless of the fuzzy definitions, literature is always trying to detect neighbourhoods with poorer living conditions relative to their surrounding context. They all attempt to understand the dimensions of urban poverty, defining these areas spatially to support pro-poor programmes and interventions, particularly in LMIC (Olthuis et al., 2015). And to do so, extract consistent information is vital to map and monitor these settlements (Kohli et al., 2012).

2.2. The physical features of deprived areas through spatial lenses

Literature has been exploring approaches to acquire information about deprived areas. By EO means, their physical features compose the urban landscape and can be derived from satellite imagery (Sliuzas, Kuffer, & Masser, 2010). Previous studies showed that the physical patterns of deprived areas can be direct proxies of socioeconomic features (Ghaffarian, Kerle, & Filatova, 2018; H. Taubenböck & Kraff, 2014). From RS imagery, spatial, spectral and textural features are interpreted and translated into socioeconomic information, providing insights into the manifestations of deprivation across the globe (Leonita et al., 2018). Meanwhile, studies testify to the need for local adjustments on the choice of these image-based proxies, as deprived areas are complex phenomena (Kohli et al., 2013; Olthuis et al., 2015). Therefore, the mapping task needs to be tailored according to their context, but the adaptation process is not yet clear (Wurm, Taubenböck, Weigand, & Schmitt, 2017).

The review from Kuffer, Pfeffer, & Sliuzas (2016) indicates that deprived settlements are often characterized by their building geometry, density, arrangement, roofing material and site characteristics. They acknowledged the built-up morphology differences among countries, cities and even across the different areas. In turn, Kohli et al. (2012) created an ontology to classify deprived areas – the publication specifically addressed the broader 'slum' term - based on their observable properties, which were conceptualized on three levels: object, settlement and environs. The classification system considers buildings and network accessibility as object level, shape and density of the urban layout as settlement level, and the location and surrounding as environs. With these levels, the local context is in vogue, but as most remote sensing (RS) studies, they did not account for variations across and within these neighbourhoods, which generate high uncertainties to the resulting classification (H. Taubenböck et al., 2018). Considering this limitation, Kuffer et al. (2017) conceptualized that deprived areas differ according to their object types, site characteristics and temporal dynamic. The first two determinants entail aspects that EO data can capture. Object types refer to the differences in size, shape, height and material of building structures and site characteristics refer to the settlement size, shape and density, accessibility to services, land cover, land use, topography, hazardous location. On the ground, these morphological characteristics combine themselves and stress the diversity of slums across and within cities.

In the same direction, the study of Lilford et al. (2019) reinforced the need for research on physical features of slum areas while identifying the most recurrent features in literature to define and categorize them in four domains: built environment, services, ecology and socioeconomic. On a more area deprivation direction, Thomson et al. (2020) reflect on the physical and social characteristics of deprived areas. They suggest physical indicators follow the research of Kohli et al. (2012) and stressed the need for social features that are not easily captured by RS, e.g., water quality, sanitation provision and accessibility to services. In line with this view, the work of Olthuis et al. (2015) suggests that the characterization of deprived areas should be locational-based driven, i.e. tailored upon local man-made and natural environment features. Their research highlights the importance of the surrounding environment for urban interventions that currently focus on housing structure improvements. The land morphology of deprived areas, e.g., hydrography, topography and land cover, should be considered together with the manmade features (Fernandez, 2012).

2.3. An openly available and disaggregated approach

The world stands in the data revolution era. Many resources are available to support decision-making processes, demanding advancements in terms of data science tools and techniques (Batty, 2019). The ultimate demand is to find ways of transforming data into useful, interpretable and up-to-date information to stakeholders (Chakraborty et al., 2015). In the field of urban data, the great challenge is related to data availability. While some LMICs face data scarcity to address and manifest urban poverty, others, as Brazil, India, Kenya, already provide detailed information on poor settlements, subsidizing national tools to make urban data freely available (The World Wide Web Foundation, 2016). However, the urban data is acquired at the household level with census surveys taken in large time intervals and traditionally aggregated to large

political spatial units (Kuffer, Pfeffer, et al., 2017). Setting administrative boundaries as analytical units do not account for the morphological heterogeneities on the ground and creates ecological fallacies by using average values for a whole enumeration unit, which hinder the usefulness and accuracy of the extracted information (Wurm & Taubenböck, 2018). It indicates that even as open-source, census data offers biased results and insufficient capacity to capture all the local nuances of deprived areas. Recently, open geodata platforms, e.g., the United States Geological Survey (USGS), offer a possibility to overcome this issue with the current EO data and infrastructure advancements. They allow free access to satellite imagery that can be used to extract information from explicit spatial features, using spectral, texture and spatial metrics (Kuffer, Pfeffer, & Sliuzas, 2016). These open datasets are precious because they are globally available, up to date, disaggregated and allow scalability (Chakraborty et al., 2015). However, the open data revolution with VHR imagery is still not yet achieved – costs grow exponentially to the study area extent - making it hard to scale the studies (Thomson et al., 2020). As most slum mapping and classification studies aim to prove VHR data's capacity, only a few focus on demonstrating the usability and potential of the openly available gridded datasets (Mahabir et al., 2020). In addition, emerging crowdsourcing technologies, e.g., Open Street Maps (OSM), enable access to other aspects of deprivation that are not easily depicted from HR/VHR imagery (Ajami et al., 2019). Vector layers of services, facilities and infrastructure provide detailed information on deprived areas that could be only available by ground survey, which may not be available.

The main challenge of incorporating these open datasets is the unit of analysis used to integrate them and analyse the study area (Gorelick et al., 2017). It refers to the aggregation level, how it affects the operationalization and resulting information. Features from different open data sources can reflect multiple local characteristics of deprived settlements, but combining these layers at diverse spatial scales requires conversion of the original level of detail, affecting the quality of the information to be extracted (Leyk et al., 2019). For this reason, defining the unit of analysis and, consequently, the aggregation level is a difficult task, and little is documented in the literature about this decision process for pixel-based approaches. Deprivation studies using EO data often vary between pixel or segment-based approaches to define the level of aggregation (Kuffer, Pfeffer, & Sliuzas, 2016). This research understands gridded outputs can be more beneficial, because they can better depict the gradual boundaries of deprived settlements and consider the heterogeneity within each settlement (Thomson et al., 2020). According to literature, defining a sensibly sized grid should consider the spatial extent of the settlements, the input spatial resolution and the study's purpose (Duque et al., 2017; Kit et al., 2012). Moreover, Huang and Zhang (2013) state that finer granularity does not guarantee accurate results, which is reassuring to this research, considering the resolution of the available open-source datasets.

2.4. The purpose: characterising deprived areas

There are several methods available in the literature to identify deprived areas. In general, field surveys provide important data at the household level about deprived settlements (Martínez et al., 2016). However, resulting maps aggregate the information into administrative units that often reflect political governance rather than the urban morphology of the settlements on the ground (Kuffer, Pfeffer, Sliuzas, & Baud, 2016). Given the advancements of RS imagery and the rise of international interest in slum improvement, several studies are relying on image analysis and classification methods (Goldblatt et al., 2018). In this sense, EO data comes in handy and holds a strong position towards creating knowledge repositories in a more inclusive way (H. Taubenböck & Kraff, 2014). The spatial features can be extracted from satellite imagery using qualitative and quantitative methods. Qualitatively, data can be extracted via visual image interpretation, which highly depends on local knowledge (Kohli et al., 2016). Experts observe and manually classify slum areas according to their morphological features, such as building size and layout pattern. This is a straightforward, low-cost method compared to field survey methods, but very time consuming and more reliant on the expert's skills for low error margin in classification (Kohli et al., 2012; Kraff, Wurm, &

Taubenbock, 2020). Quantitatively, (semi-)automatic image classification techniques are performed by algorithms that extract the features as pixel values and quantify them.
Currently, traditional ML algorithms are the most common EO-based methods used (Kraff et al., 2020). Most of them use supervised learning techniques to classify and predict the settlements with systematic and iterative processes (Gram-Hansen et al., 2019; Grippa et al., 2018). They are versatile and combine textural, spectral, and spatial information extracted from RS data (Kuffer, Pfeffer, & Sliuzas, 2016). They are also flexible to the aggregation unit, allowing the choice between pixel or object-based approaches (Wang, Kuffer, & Pfeffer (2019). Some studies focus on ML-OBIA (Duque et al., 2017); others choose pixel-based approaches which require contextual image features to feed the model and provide accurate results (Mudau & Mhangara, 2021). The purpose of the analysis defines the choice of the algorithm. For instance, Decision Trees performed very well in this slum delineation study in Kenya (Mahabir et al., 2020), while a range of image-based features are derived with Random Forest classifier and used as input to generate typologies for an urban area in Mumbai (Kuffer, Pfeffer, et al., 2017).
Even though these studies provided good performance, they still have drawbacks. The major issues relate to access to VHR data and training labels/ground truth reference. Under the supervised learning umbrella, classification and regression algorithms require large training samples as prior knowledge for the learning task, which requires expensive imagery and reference data (Leonita et al., 2018). As an alternative, there are unsupervised learning models. The algorithm learns spatial patterns from unlabelled data and with minimum supervision required. They have been used to classify land cover and land use as Senthilnath et al. (2013) did or to identify urban settlements as performed by Jochem et al. (2020). Within deprivation and slum studies, an unsupervised ML gridded approach was not yet exploited in literature. It can provide an alternative to profile local deprivation nuances, capturing detailed information on their specific characteristics and distinguishing them into spatial clusters.

2.5. Modelling deprived areas with unsupervised learning techniques

Unsupervised learning tasks have been used for decades in science. Scientists are always trying to organize data into classified groups and, unlike supervised techniques, unsupervised ones allow them to organize data in the absence of category labels (Jain, 2008). In the ML field, the computer algorithm learns from data without receiving a class label. They depend on their own to find the structure to the input and discover hidden patterns in data (Han, Kamber, & Pei, 2012). The model studies the underlying structure in a dataset and group objects according to the similarity of their measured characteristics (Soni Madhulatha, 2012).
Clustering is the most common category of unsupervised learning models. As a branch of data mining, it is used to observe the characteristics of data clusters and operate characterization and classification tasks. As a field in ML, it involves segmenting datasets based on shared attributes and aggregating their variables according to these attributes (Scrucca, 2016). Essentially, it is the process of portioning datasets into subsets (clusters), where two properties are met: objects in a cluster are similar to one another and yet different from the objects in other clusters (Han et al., 2012). Compared to supervised tasks, clustering models face the same issues as other unsupervised tasks. Since there are no labels to compare results, validating results is not a straightforward process. It requires internal and external evaluation with a subjective component.
There are four main clustering types: partitional-, hierarchical-, density- and grid-based. The choice of the clustering type follows specific prerequisites: scalability, input data type, model sensitivity, noise handling, dimensionality and interpretability (Han et al., 2012). Scalability is an essential aspect to consider, especially when dealing with large datasets because clustering models can be computationally burdensome (Scrucca, 2016). Some algorithms can only handle categorical or only numerical data types, and this must be considered. Some models are very sensitive to the input, with less robustness to handle dimensionality and noise data. Solorio-Fernández, Carrasco-Ochoa, & Martínez-Trinidad (2019) indicate that input data can increase the computational time and overfitting due to data dependency and redundancy. Hence, several

studies discuss the advantages to apply Exploratory Data Analysis (EDA) techniques before the modelling task (Ibes, 2015; Roy, Bernal, & Lees, 2020).

In ML problems, EDA is an advisable preliminary step that investigates the data, avoiding biased results by analysing the relationships of the input features, assessing their collinearity and helping the selection of the most important ones (Cai, Zhang, & He, 2010). Moreover, literature also indicates the use of dimensionality reduction techniques to facilitate the interpretation of the clustering products (Martino, Rizzi, & Mascioli, 2017). The two main existing approaches are feature extraction and feature selection. For the first approach, Principal Component Analysis (PCA) is the most famous. The unsupervised technique simplifies the number of variables retaining the maximum variation amount by deriving new features from original variables (Hyvärinen, 2015). In the second approach, feature selection is readily readable because it maintains the original features values, in a smaller subset without feature transformation (Li et al., 2017).

Literature shows great advancements and newly released feature selection algorithms, but the great majority stays under the supervised umbrella (Alelyani, Tang, & Liu, 2014). Table 1 lists the few unsupervised techniques found. They face different limitations regarding the model input, the availability of the implemented software and the level of expertise required to run the technique (Li et al., 2017). Last year, Professor Oliver Pfaffel released a novel tool for measuring feature importance for a clustering algorithm as a new R package named "FeatureImpCluster"2 (Pfaffel, 2021). The tool is freely available and well documented in the Comprehensive R Archive Network (CRAN) and has been applied in a few studies (Fisher, Rudin, & Dominici, 2019; Nugrahita & Surjandari, 2020), none of them slum or deprivation-related.

Table 1. Unsupervised feature selection techniques found in the literature. Built on the review of Li et al. (2017).

Feature selection algorithm	**Constraints**
Projective Adaptive Resonance Theory	Binary categorical input
Spectral Feature Selection (SPEC)	Solely implement in MATLAB (not open source)
Laplacian Score	Only runs as supervised learning in WEKA software
Kohonen Self-Organizing Maps	Too computational and knowledge intensive (specific ANN algorithm)

[2] Available at: https://cran.r-project.org/web/packages/FeatureImpCluster/readme/README.html

3. METHODOLOGY

This chapter describes the research methods. It explains the motivation background of the study area, followed by the methodology overview. Then, the sections describe the datasets collected, all the processing steps, data analysis and modelling tasks, including the validation procedures.

3.1. Study area

3.1.1. Motivation for São Paulo as a case study

The three main reasons to choose the city of São Paulo for this study are: *(1) data availability*. The present approach requires the prior delineation of the deprived areas, and the Brazilian Institute of Geography and Statistics (IBGE) has just released a new national layer produced in 2019 (Annex 1). Their intention to publish the content is to help fighting COVID-19 with specific health information on the precarious settlements (IBGE, 2019a). Previous studies argued about the reliability of the same national layer of 2010 compared to the slum layer created by the municipality of São Paulo. Conceptual and cartographic inconsistencies exist, mainly because the layers use different conceptualizations of deprived areas (Pedro & Queiroz, 2019). The updated layer for 2019 has been investigated in N. D. J. Ferreira & Feitosa (2020) and also carefully inspected in the present research to test its reliability and current inconsistencies. Digitation errors were spotted and fixed, ensuring the reliability of the layer (Figure 2). Section 3.4 describes all the processing steps in detail. Besides, compared to the existing municipal layer, the IBGE national layer stands out for the complete spatial coverage and consistency across the country – only a few Brazilian cities have municipal layers of their deprived settlements, and this research is intended to be transferable.

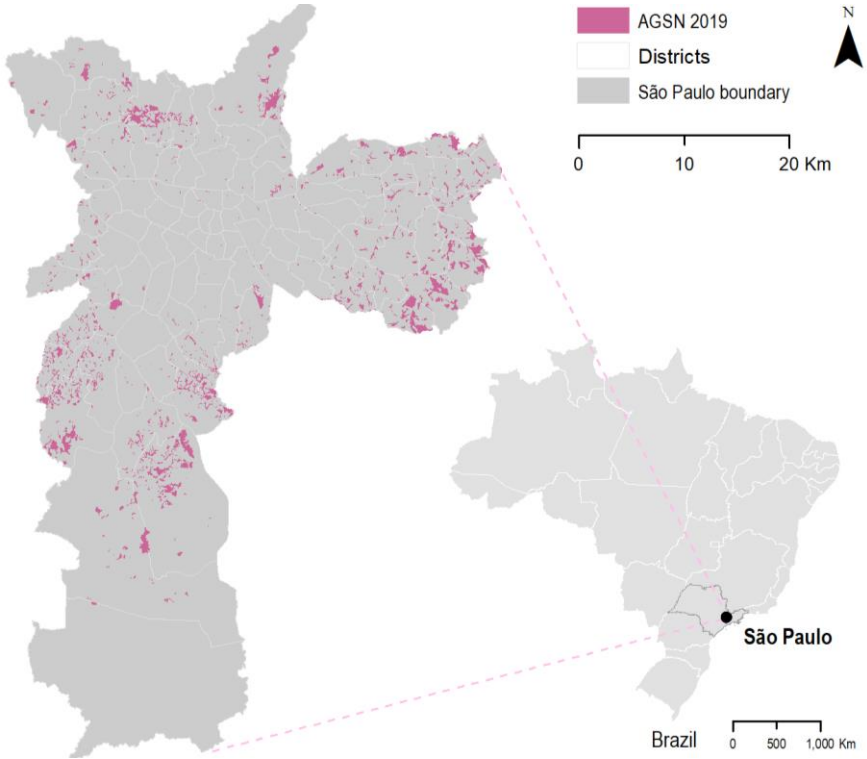


Figure 2. Location of São Paulo and deprived areas from AGSN layer. Source: (IBGE, 2019a).

(2) the release of the MAPPA project for the State of São Paulo, created in partnership with the Agency of Urban Development and the Federal University of ABC region (São Paulo, 2019). They presented different typologies of deprivation (Annex 2) for cities in the Metropolitan Region of São Paulo using a Logistic

Regression algorithm. Still, there is no similar quantitative characterization study for the capital city until the present date. *(3) the city of São Paulo is considerably more organic, dense and complex than other metropolitan cities in Brazil,* especially regarding deprived areas. In the last two decades, the growth of poor settlements happened towards the periphery (Pasternak & D'Ottaviano, 2016), and this growth rate is four times the formal population growth rate. This increasing population adds more urgency and complexity to the case study.

3.1.2. Mapping deprivation in Brazil and in São Paulo

Figure 3. Spatial inequality in São Paulo. Source: The Guardian (2017).

São Paulo is the 4th largest city in population in the world and the most populated one in Brazil, with more than 12.3 million people (IBGE, 2020). It is the state's capital with the same name, located in the southeast of the country. Economically, the city is responsible for 12 percent of the national Brazilian GDP. Wealth and wellbeing are, however, not equally distributed across the city. The famous picture taken in 2004 by the photographer Tuca Vieira is still a mark of the socioeconomic inequalities that are easily distinguishable in the landscape (Figure 3). The city's morphology reflects the inequalities seen in several Latin American cities (Maricato, 2003). Poor settlements occupying marginalized areas evince the exclusionary urbanization process and the consequent precarious living conditions for a specific portion of the population (Pasternak, 2006).

The major definition of deprived areas in Brazil used in research, policies and plans was created by the IBGE under the generalized concept of 'Aglomerados Subnormais' - abbreviated as AGSN and direct translated as 'Substandard Settlements' - for the census in 1991 and it remains valid today (Pasternak & D'Ottaviano, 2016). The term is used to recognize and map deprived settlements at the national level by including them as an important part of the urban landscape and targets for pro-poor policies (Brasil, 2010). More than a proxy of slum household, it describes settlements with at least 51 dwelling units in a substandard form of urbanization, deprived of legal land ownership, within a disorderly layout or poor service provision (IBGE, 2010). According to the bureau, if a settlement has an irregular land registry and presents at least one of those criteria, it will be identified as a 'substandard settlement'. However, the AGSN definition encompasses all deprived settlement variations in Brazil as a single class, regardless of their local differences (D'Alençon et al., 2018). Thus, ignoring their heterogeneity and reinforcing the hierarchy between formal and informal areas as state by Pedro & Queiroz (2019).

3.2. Research methods

This section describes the overview of the methodology. Figure 4 illustrates the overall workflow in correspondence with the research objectives and questions. The literature review, described in Chapter 2, guided the development of the variables and the data collection process. Then, the research is comprised of two parts: data processing and unsupervised ML. After the open geodatabase for São Paulo is built, the RS- and GIS-based features are extracted, and the scale of analysis is selected so that all features are transformed, resampled and integrated accordingly. Once input data is prepared and analysed with EDA techniques, the second part develops the unsupervised learning model with optimization and feature importance techniques. The last part validates the model, analysing the relationships between the resulting cluster types and the spatial features used and assessing the applications of the unsupervised model to capture the intra-urban variations of the deprived areas in São Paulo. A parallel model only with census-based variables, the land-use layer and expert validation are used as reference data to validate the gridded results.

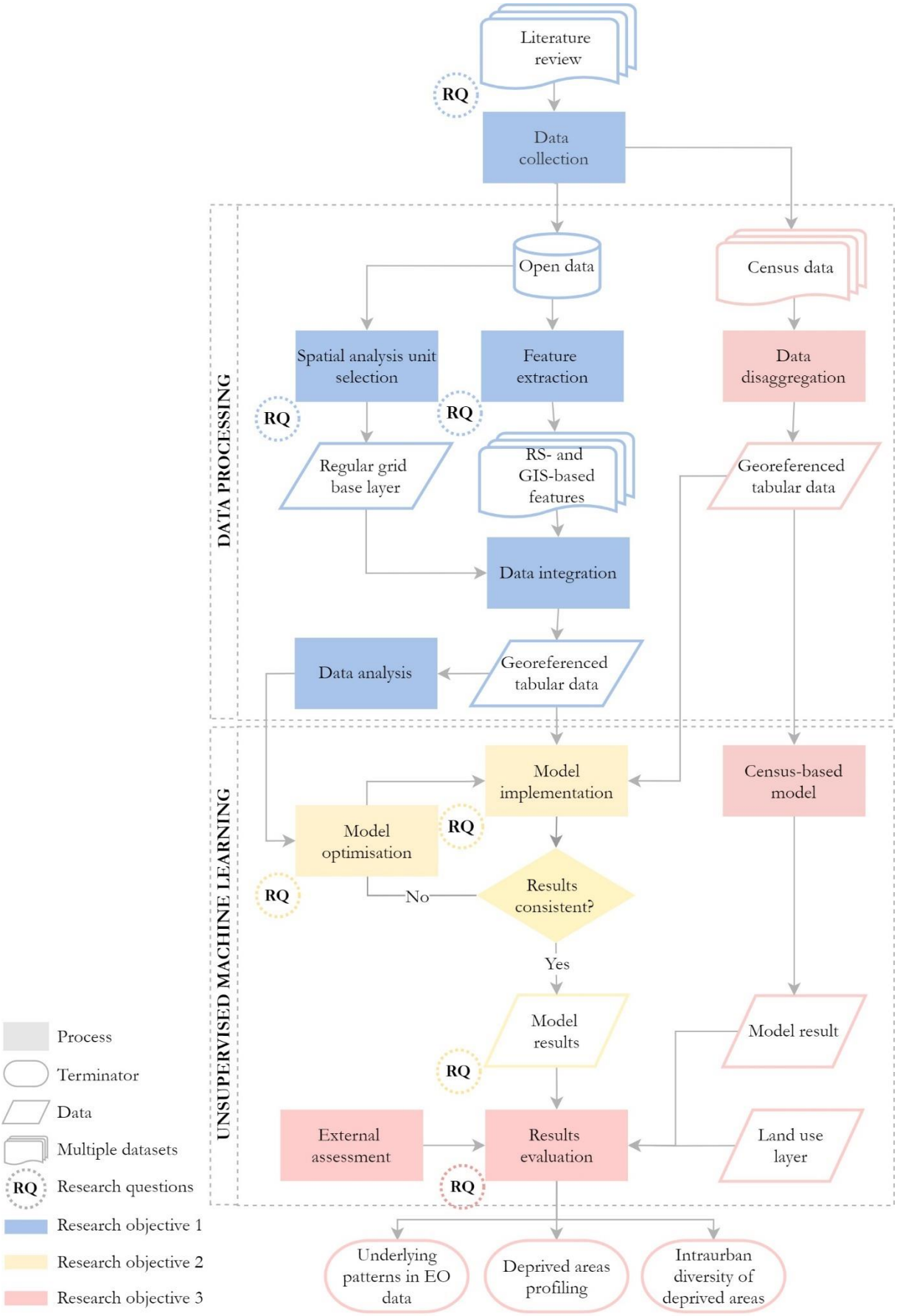


Figure 4. Methodology flowchart.

3.3. Data collection

Considering that this research uses a city-wide open data approach, the data collection process was guided by the following criteria: (1) all features must be spatial, quantitative and available for the entire study area; (2) all features should be provided from open data sources to deal with data scarcity; (3) all features should be collected ensuring systematic and constant monitoring (considering the limitations of the global repository, recurrent collected every five to seven years, or at least more frequent than ten years); (4) the features should be able to manifest the specificities of the city. Besides the given criteria, the observable properties identified in the literature (Chapter 2) also guided the data collection process. Figure 5 shows the two domains, built-up and land morphology, and the eight dimensions that structured this data collection process. They were built on previous studies and were responsible for guiding the extraction of 32 potential features from different open geodatabases. Annex 3 lists them, including their specification, data source, rationale and corresponding literature reference.

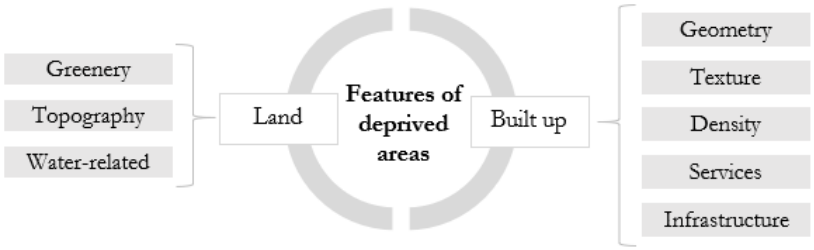


Figure 5. Diagram of the employed domains and dimensions of deprived areas employed.

In addition to the spatial features, this research also required the boundaries of the deprived settlements in São Paulo, delineated by IBGE (2019) using visual image interpretation and field survey. It was the first layer to be acquired as it is essential to define the spatial unit of analysis. Census data (2010), also provided from the IBGE web portal (2019b) and aggregated at the census tract level, is acquired for the model experiment. The 10-year gap is a limitation of the study, but it was not anticipated. The census operation was suspended due to the COVID-19 global pandemic. Moreover, the municipal land use layer is collected to be used as ancillary data for model validation (São Paulo, 2020). To maximize the research' reproducibility, Annex 4 also summarizes the Data Management Plan (DMP), containing information on the data collected with the open-source links, processing and modelling scripts and data products, openly available at GitHub[3].

3.4. Data processing

After data collection, several data processing steps were conducted to derive and integrate the input features. Sub-section 3.4.1 describes the inspection of the base layer and the selection of the optimal scale of analysis. The resulting base grid layer was used to extract the image features at the pixel level through different geoprocessing techniques, explained in detail in Section 3.4.2. Lastly, the feature input is integrated into a single input file to the model, standardized and assessed with EDA techniques.

3.4.1. Spatial unit of analysis

In this study, deprived settlements are perceived as part of a continuum on the urban landscape, rather expressed by continuous boundaries than discrete ones (Sankey, 2016). Thus, a pixel-based approach is provided and requires the identification of an optimal analytical unit. Considering the input spatial resolution – ranging from 10 to 100 m^2 -, the present study understands that the choice of this unit is not arbitrary, but rather contextual, relying on the scale and morphology of such deprived areas (Wang et al., 2019). For this, as mentioned in Section 3.3, the AGSN layer is used as a proxy for deprived areas in São Paulo. Each polygon was checked on top of ESRI world imagery map service. In accordance with Ferreira and Feitosa (2020), when compared to the AGSN layer of 2010, the newly released one has more coherent boundaries,

[3] This repository is openly available at GitHub for reference https://github.com/ltrentooliveira/MSc_Archive

justifying the choice of the layer (Figure 6). Despite topological corrections, the layer still faces boundaries inconsistencies, requiring careful correction. With regard to the work of Molenaar (2000) and the documentation provided by IBGE (2019a), the layer manifest extensional uncertainties as identified in Figure 7. Obvious digitation errors were corrected, e.g., polygons extrapolating to water where clearly no one lives in or large road networks in between settlements.

Figure 6. Comparison of AGSN 2010 and 2019. a) commission errors (residential/non-residential land uses); b) shifting polygons; c) sliver polygons; d) extrapolated to water; e) omission error.

Figure 7. Examples of related uncertainty that remains on the AGSN 2019. They indicate the inclusion of a) water/non-built-up areas; b) road network and non-residential areas.

After the inspection, the scale and shape of the polygons were analysed. Following the work of Taubenböck and Kraff (2014), Table 2 shows how high is the variability of the polygon's size in the city of São Paulo.

Table 2. Descriptive statistics of the 1,575 polygons of the AGSN layer regarding the area attribute.

Max	**869,888 m²**
Min	745 m²
Sum	61,724,714 m²
Std. Deviation	70,890 m²
Mean	39,190 m²

It is important to state that the presence of these large extent polygons contributes to the choice of a disaggregated approach, as a single settlement can manifest heterogenous aspects of deprivation. The shape

of the polygons also varies significantly and can influence the grid size choice. Prior studies adopted trial-and-error approaches to define the best grid cell size. Some, focused on pixel-based mapping and delineation, used from 15m (Mahabir et al., 2020) to 100m (Duque et al., 2017); others, focused on characterization, usually indicate finer granularity (Ajami et al., 2019). Based on this, regular grids of 10, 20, 50 and 100m were created and overlaid with the just inspected AGSN layer. By inspecting different settlement sizes, the final grid size is selected to represent the urban structure of the deprived settlements and open spaces between them in Section 4.1 (Kit et al., 2012).

3.4.2. Feature extraction process

After collecting the data, a series of processing steps were conducted to extract the 32 features at the pixel level and preparing them as input for the clustering model. Figure 8 describes the workflow, including the extraction of RS- and GIS-based features. As shown in the diagram, all layers are initially clipped to the study area, projected to the SIRGAS 2000 reference system, resampled and snapped to the 20x20 base grid. First, density and infrastructure features derived from the WorldPop are extracted. Second, the two topography features, mean DEM and slope values are extracted from the GDEM layer. Third, the spectral features are derived from the Sentinel 2A imagery. Mean statistics are calculated for the bands - 2 to 7, 11 and 12, due to their fine spatial resolution - using a 5x5 kernel. Bands 8 (near-infrared) and 4 (red light) are used to extract the Normalized Difference Vegetation Index (NDVI).

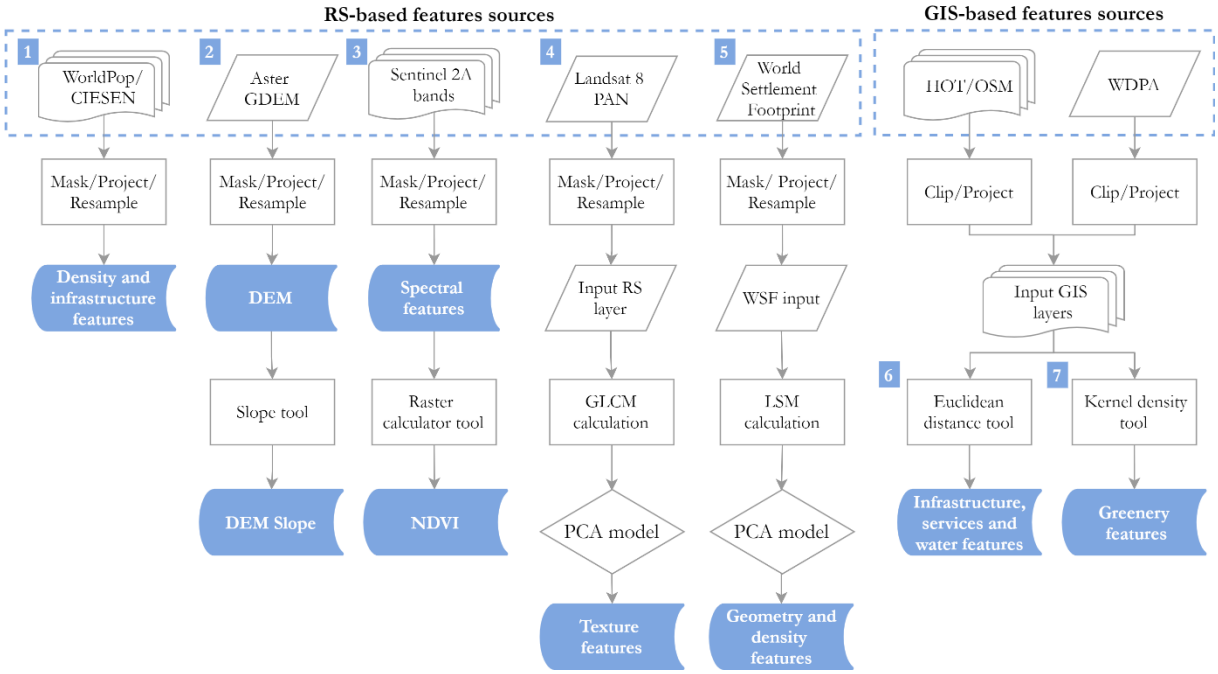


Figure 8. Processing steps of feature extraction. Blue shapes with bold text are the final stored features.

Steps four and five refer to the extraction of hand-crafted features that require more specific processing steps. As prior studies indicated, Gray-Level Co-occurrence Matrix (GLCM) measures the spatial relationship of the pixels (Mahabir et al., 2020). The panchromatic (PAN) band is chosen over a multispectral one because of its capability to detect brightness changes better and preserve contextual details at the chosen spatial resolution (STARS, n.d.). Referenced by Kohli et al. (2016), seven texture statistics, e.g., mean, variance, homogeneity, contrast, dissimilarity, entropy and second moment, were calculated using the *'glcm'* package in R Studio. The code required quantization levels, rotation direction and window size as entry parameters.[4] 32 grey scale values are chosen to balance information loss and processing time (Hall-

[4] https://github.com/ltrentooliveira/MSc_Archive/blob/main/GLCM.R

Beyer, 2017a). According to Wurm et al. (2017), some features are highly sensitive to unstructured or irregularly urban morphology, which is common in deprived areas. Therefore, the GLCM calculation should be over all directions (0, 45, 90 and 135 degrees).

Several studies suggested trial-and-error approaches empirically varying the moving sizes to decide the optimal one (Leonita et al., 2018). This research complied with Mahabir et al. (2020) because they used a similar granularity as employed here. Each feature was extracted at 3x3, 5x5, 7x7, 9x9 and 11x11 window sizes, generating 35 image features. Following the instructions of the study of Hall-Beyer (2017), Principal Component Analysis (PCA) is used to support the window size decision. PCA model calculates a covariance matrix of the input dataset, indicating the relationship between them (Hyvärinen, 2015). From that, the computation creates a line formed by the best projection of data points, capturing most of the variance in the data, named Principal Component (PC) (Jolliffe & Cadima, 2016). PCA discovers the dimension that maximizes the input variables' variance by isolating the variation for each individual PC (Hall-Beyer, 2017b). The PCs' interpretation relies on three main output metrics: *(1) correlation matrix*, the weight of each input feature for each PC*; (2) factor loadings*, which indicates the correlation between feature and PC; *(3) communality table*, that calculates how much of the variance in each of the feature is explained by each PC (Richardson, 2009). Based on these metrics, the features with very high collinearity, low communalities and loadings scores are removed from the model. The PCA iterations with each moving window are conducted using Python packages[5] (see Annex 5 for details on each experiment).

The fifth processing step encompasses the landscape metrics (LSM) extracted from the World Settlement Footprint layer (WSF). Class-level metrics can separately quantify the spatial patterns of each patch, e.g., homogenous regions on the landscape, describing the aggregation, subdivisions, area and shape distribution of their spatial structure (Jia, Tang, Xu, & Yang, 2019; Reis, Silva, & Pinho, 2016). Here, the WSF categorical raster expresses the built-up areas as a single class and delineate it as a patch mosaic in the landscape of São Paulo. Guided by Frazier and Kedron (2017), five metrics are calculated (patch area mean, shape index, fractal dimension index, patch density and aggregation index) in Fragstats using four moving windows (3x3, 5x5,7x7 and 9x9). Just as with the GLCM, the trial-and-error approach is assessed with PCA (Annex 5).

After extracting the RS-based features, OSM and Humanitarian OpenStreetMap Team (HOT) data are used to extract additional contextual and morphological characteristics of deprived areas using vector layers. HOT data were visually inspected, and it is more complete than OSM, with proper specification of features, and more reliable regarding urban facilities. Therefore, the OSM layers are used for the infrastructure features, but the HOT layers are also used to complement the services-related features.

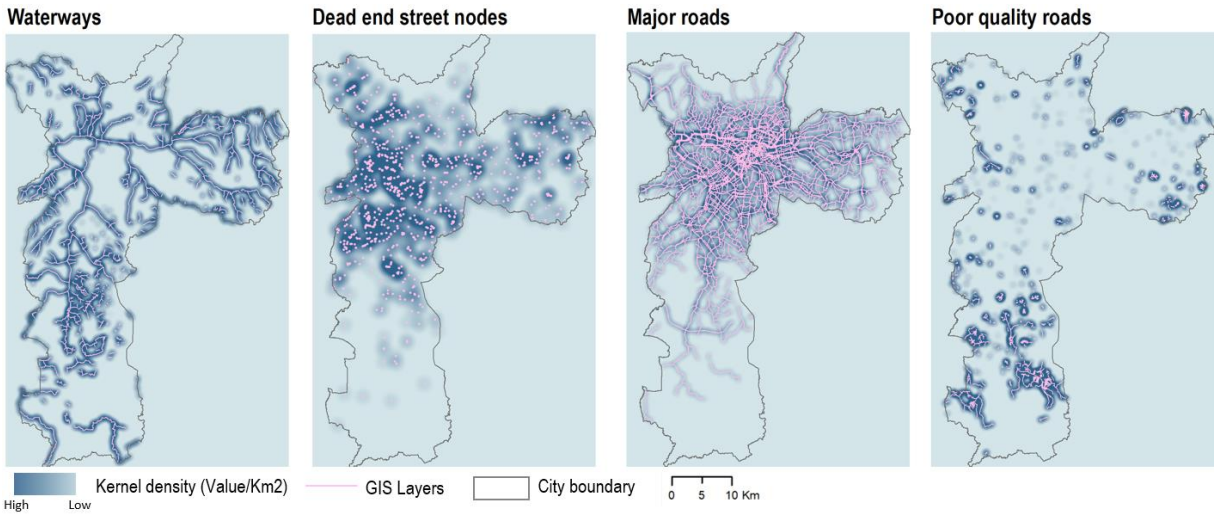


Figure 9. Density estimation: 1km search radius distance of each GIS-based feature.

[5] https://github.com/ltrentooliveira/MSc_Archive/blob/main/PCA_processing.ipynb

After data acquisition and inspection, data went through screening examination to select specific attributes using SQL query expressions (Annex 6). The World's Database of Protected Areas (WDPA) was used combined with the OSM data to derive the green areas. The features regarding accessibility to services, infrastructure, green or blue spaces features are calculated using Euclidean Distance to the nearest point of interest. In the seventh and last step, four GIS-based features are derived by calculating the kernel density. The work of Mahabir et al. (2020) is one of the few studies that clearly expressed the bandwidth used (1km). By using as a baseline, four bandwidths are tested (500m, 1000m, 1500m and 2000m). Figure 9 illustrates the chosen search radius, which was 1km to all four features, as the spatial structure of the features remains explicit with the smoothing effect (Leonita et al., 2018).

3.4.3. Data integration and analysis

After extracting the features, several steps were conducted to integrate and prepare the final input data to run the unsupervised learning model. A primary key is created following the label order from the grid's origin (lower left corner), and the grid polygon is converted to a point centroid layer. The cell values of each final input raster are extracted at the centroid location and recorded to the attribute table of the point class layer. Given the number of data points (149947) in the study area, this extraction and integration process was computationally intensive. Thus, the data points are split into seven batches, processing each at a time. Afterwards, the seven attributes table are remerged into a single georeferenced tabular file for EDA (see Section 4.2). The features are not expressed in the same unit, requiring standardization. Descriptive statistics are generated and visualized with box- and density plots to detect likely outliers and analyse data distribution, respectively. Next, null population values are handled as zero, and the Pearson correlation coefficient is computed to assess the statistical significance and multicollinearity of the processed features. These steps are conducted in R Studio, which code is openly available at GitHub repository for future reference[6].

3.5. Unsupervised machine learning

This section explains the process of building an unsupervised learning model. Like any other ML task, after data preparation and processing, the machine needs to learn an underlying model and iteratively improve, followed by model validation. The clustering analysis aims to build a model to characterize deprived areas in São Paulo without requiring labelled datasets using only open data sources.

3.5.1. Clustering algorithm choice

Identifying the best algorithm for the specific task and purpose depends on different criteria. Following Han et al. (2012), it is important to consider data input requirements, performance, dimensionality and sensitivity to the input. The first requires knowledge of the data used and their properties; thus, only algorithms that handle continuous numerical data could be selected. The second criterion is performance and scalability because of the high volume of data and large spatial extent of the study area. According to Alelyani et al. (2014), working with sub-samples in clustering does not indicate the representativeness of large datasets. Figure 10 shows how the processing time exponentially increases for several algorithms – the more sophisticated ones - adding impracticality regarding the capacity of the machine used for this research[7]. This computational limitation is even more emergent in LMIC municipalities, therefore a significant constraint to be dealt with. The figure indicates K-Means, DBSCAN and HDBSCAN as the best algorithms to support large datasets, without the expenses of RAM limit.

The three algorithms were experimented using readily available functions and packages in Python library. DBSCAN algorithm is a density-based algorithm, that is well documented in literature, but the code crashed recurrently for memory error. It could only be solved with parallel computing (Martino et al., 2017),

[6] https://github.com/ltrentooliveira/MSc_Archive/blob/main/EDA.R

[7] LENOVO _MT_20QU_BU_Think_FM_ThinkPad P1 Gen2 Intel Core i7-9750H CPU, 16GB RAM, Windows 10

hindering the usability of the model and the practical perspective proposed in this research, because the author is concerned by the technicalities and the available resources in LMICs.

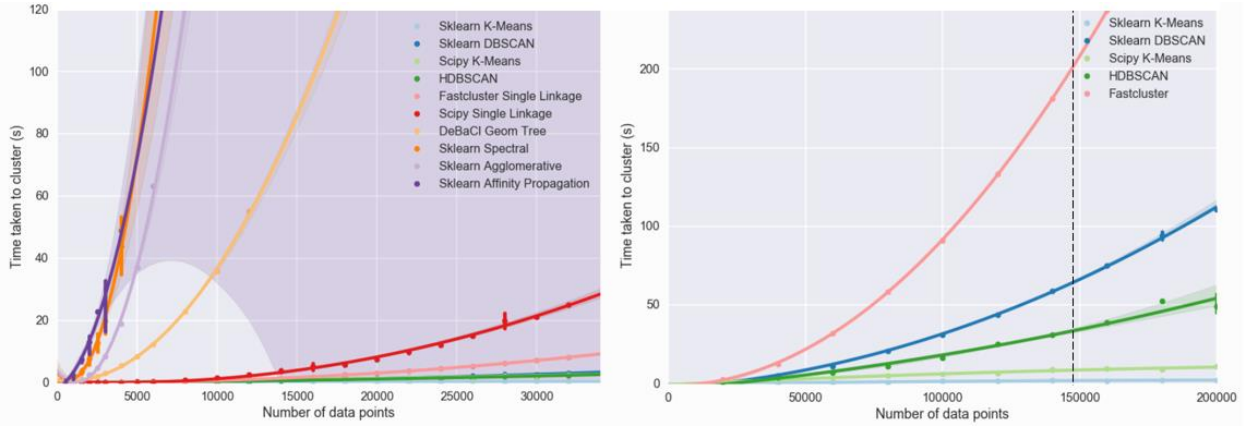


Figure 10. Performance comparison of clustering algorithms. The left graph compares smaller datasets and the right larger ones. Adapted from: McInnes, Healy, & Astels (2016).

HDBSCAN and K-Means algorithms ran on the available machine with the prepared dataset. Unlike the supervised ones, unsupervised ML tasks are not easily compared, because model assessment cannot be done with mathematical error metrics. Overall weaknesses and strengths of both algorithms are compiled from Campello, Moulavi, & Sander (2013) and Soni Madhulatha (2012) in Table 3. HDBSCAN can excel with fewer assumptions than K-Means regarding data uncertainties. However, the initialization parameters of HDBSCAN have very little documentation in the literature and are not intuitive for modellers. The 'min_cluster_size' sets the smallest cluster size possible, determining whether a group of points fall out of a cluster or are split to form other clusters. The 'min_samples' sets whether the cluster is considered noise or not, which is the biggest weakness of the model, because it is highly prone to disturb the model stability.
K-Means is the most popularly used and documented algorithm, and it only requires a predefined number of clusters as the input parameter. This can be a drawback, but there are heuristic techniques to help define it. Based on the scope of this research, contextual knowledge on deprived areas is provided and expected for transferability. Regarding the learning task, EDA can reduce possible biases due to data dependency. Lastly, k-means can be used in multiple programming languages and in consonance with different packages. Based on these arguments, the K-Means clustering algorithm is selected by the simplicity of the input parameters, its stability and compatibility with diverse open software and packages.

Table 3. Comparison of K-Means and HDBSCAN algorithms. Adapted from: Campello, Moulavi, & Sander (2013) and Soni Madhulatha (2012).

	KMEANS	**HDBSCAN**
Learning	Assign all points to clusters (noise points get lumped into clusters as well) Compromised with noise in data	Find clusters of varying density Account for noise in data (generate a separate cluster)
Parameters	Number of clusters	min_cluster_size and min_samples
Stability	Random initializations provide stability	Sensitive to initialization parameters
Performance	Exceptionally efficient	High efficiency

3.5.2. K-Means model implementation

K-Means is classified as a partitioning distance-based method that builds 'k' clusters dividing the data into groups, where a data point can only belong to one cluster (Han et al., 2012). Figure 11 illustrates how it works: (1) The model initializes with a certain partition based on the centroids of k clusters; (2) The set of data points are assigned to the closest cluster by measuring the mean-square distance (MSD) with respect to each defined centroid and new centroids are computed; (3) The centroids are relocated iteratively until

the clusters assignment no longer change; (4) steps 2 and 3 are repeatedly and iteratively calculated until final output. The quality of the clusters is measured by assessing whether that the data points close to one another are similar and the ones far are different in terms of the dataset attributes (Jain, 2008). To ensure these properties, the model calculates the MSD and assess it with the total within-cluster sum of square (aka WCSS, inertia or distortion score) that sum the distance from each point within a cluster from its centroid (Chiang & Mirkin, 2010). The algorithm chooses the cluster centroids with the lowest scores, indicating internal coherence and high compactness.

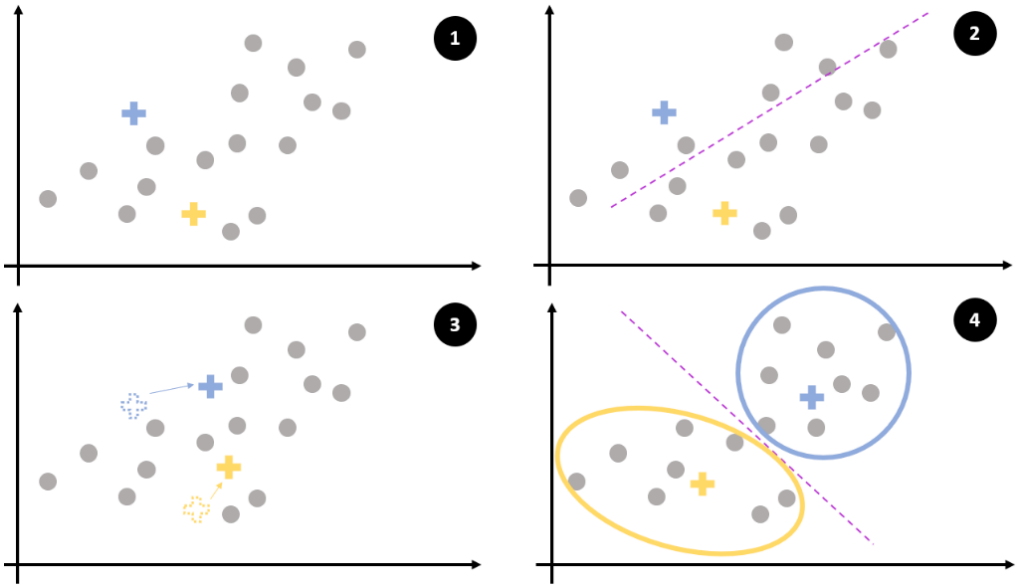


Figure 11. Simple illustration of how K-Means algorithm works with simple steps:(1) centroid initialization; (2) MSD calculation; (3) centroid relocation (4) final cluster.

By understanding how it works, the implementation of the algorithm is conducted. K-Means algorithm in Python requires few parameters setting: selecting a random seed to allow reproducibility and choosing the k value. The most widely used method to determine the optimal k value is the elbow method that plots the WCSS value as a function value of the number of clusters (Figure 12). The location of an 'elbow' on the graph, where WCSS values reduce exponentially and suddenly become constant, suggests the optimal number of clusters (Pedregosa et al., 2011).

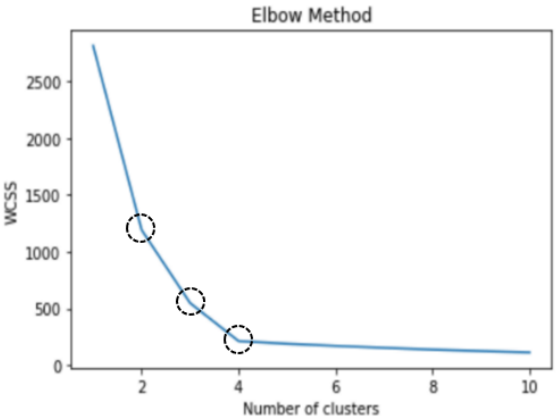


Figure 12. Example of elbow method.

In the figure, 2 to 4 are the eligible k values, while from 5 clusters on, there is a flattening of the function indicating that more clusters will not improve the partitioning task. It needs caution, because eventually, cohesive clusters can start partitioning by the exaggerated reduction of the WCSS (Han et al., 2012). In addition, it is also important to consider that more clusters, indicates more computational costs. For the implementation, multiple models are trained and in each successive model, the number of clusters increased until 10, computing and storing all the WCSS results, allowing further visualization.

3.5.3. K-Means model optimisation

Even with great fine-tuning capabilities, K-Means algorithm requires optimisation, and it is not an easy task.[8] A series of experiments are conducted with different features to check the model consistency and optimise the results. According to the work of Fränti and Sieranoja (2019), k-means robustness can be improved by repeating initializations and checking the clustering structure by adding/changing attributes; thus, all experiments are repeated 20 times to reduce erroneous clusters (Karmitsa, Bagirov, & Taheri, 2018).

Among the few available dimensionality reduction techniques with practical documentation described in Section 2.5, two tools are used to assess the model experiments: *(1) Pearson correlation matrix* diagnoses the presence of multicollinearity by recognizing highly correlated features (R>0.8); (*2) 'FeatureImpCluster' R package,* measures the importance per feature according to the calculation of a misclassification rate (Nugrahita & Surjandari, 2020). These techniques are used to select specific features for a model experiment aiming at increasing the model performance by avoiding biases and maintaining most of the feature' information.

In addition, the present research aims at increasing the interpretability of the results by understanding the significance of the features used for each resulting cluster (Solorio-Fernández, Carrasco-Ochoa, & Martínez-Trinidad, 2019). With this intention, the feature importance tool is also used in all k-means experiments. The permutation classification rate is calculated for each input, computing the number of wrong cluster assignments divided by the number of iterations (Fisher et al., 2019). The algorithm shuffles a feature value and if the model error increases, the higher the permutation score becomes and the more relevant the feature is (Thrun, Ultsch, & Breuer, 2020). The algorithm acts both like a wrapper and a filter method. It is used in conjunction with the clustering algorithm, evaluating the misclassification rates synchronously at each learning iteration which usually imply high computational costs, while simply scoring the features on a statistical basis, which reduces the computational time. Figure 13 shows the by-products offered by the tool. On the left, the overall mean misclassification rate per feature and the rate aggregated per cluster is on the right. Variables V1 and V2 are the most relevant features with the highest misclassification rates, but V3 and V4 also have impact on clusters 1 and 4.

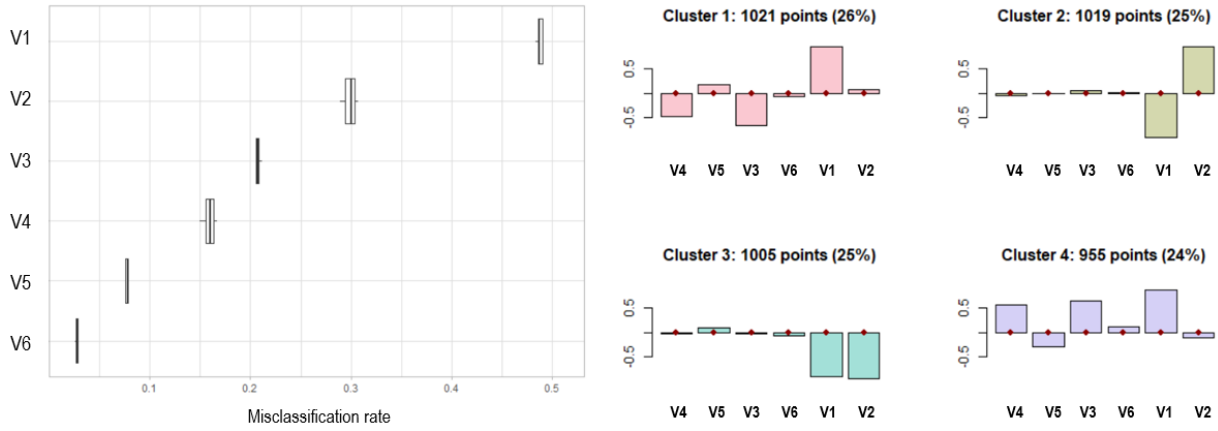


Figure 13. Example of two outputs of FeatureImpCluster tool. Left: overall mean misclassification rate per feature. Right: misclassification rate aggregated per cluster. Source: extracted from the developer's page.

Lastly, after assessing and selecting the most optimised model using solely spatial EO-based features, census-derived variables are added to the model. The aim is to increase the model performance and add different deprivation perspectives (Baud, Kuffer, Pfeffer, Sliuzas, & Karuppannan, 2010). The selection of the variables is guided by the IDEAMAPS project (Figure 14). Considering the added value of the census-

[8] The scripts used for model implementation and optimisation with the different input features can be found at https://github.com/ltrentooliveira/MSc_Archive/blob/main/kmeans_model.ipynb

derived features providing information at the household level, the variables aim to complement the model with dimensions not easily acquired by RS data. The 12 selected variables can be seen in Annex 7 and include crowding of the dwelling, employment level, ownership, drainage and waste management indicator groups. After the selection of the variables, some preparation steps are required to integrate the census-based features with the EO-based ones and basic areal interpolation techniques are applied to distribute the aggregated data – originally at enumeration tract-level - into the grid cells. The main processing steps are described on the flowchart in Annex 8.

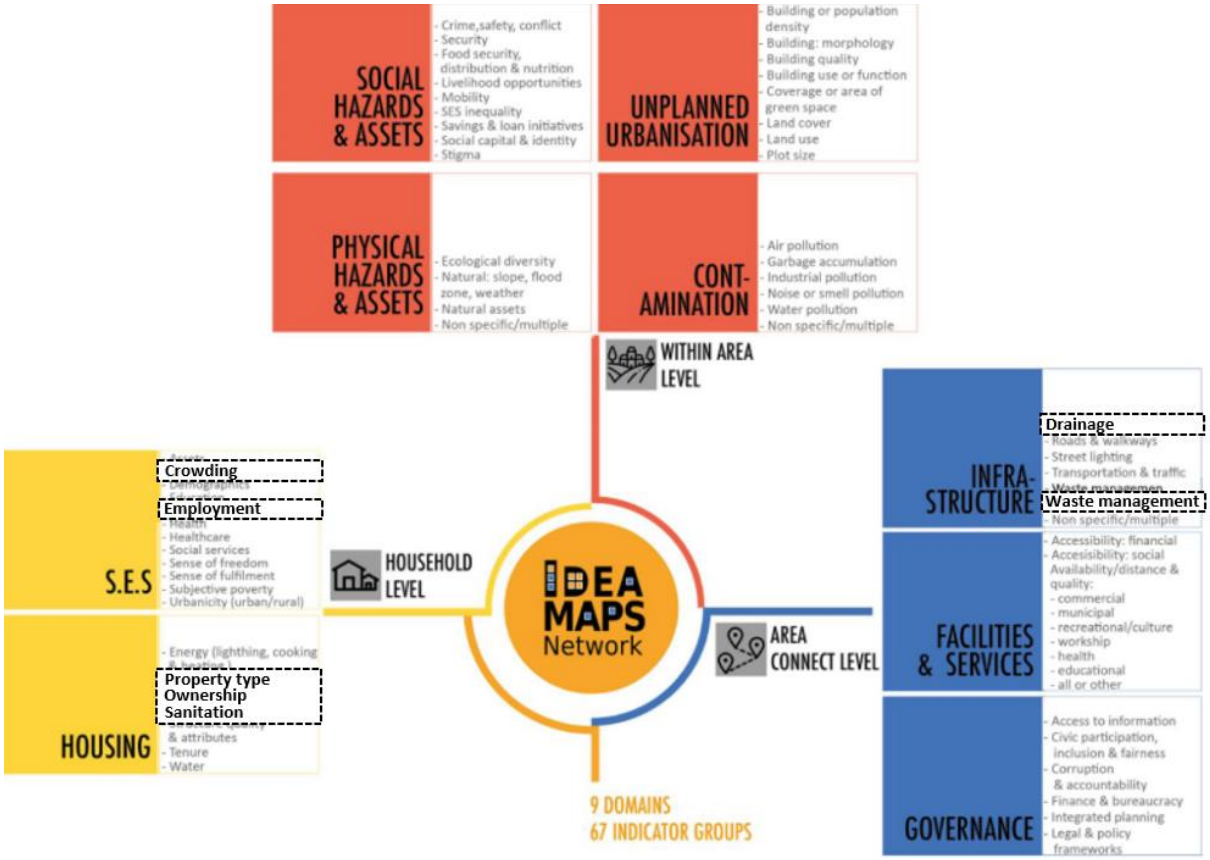


Figure 14. Selected variables from IDEAMAPS framework in black dashed boxes. Source: (Abascal et al., 2021).

3.5.4. K-Means results evaluation

Table 4 below presents the different validation steps conducted to assess the approach and the k-means model results. From the beginning, expert consultation has been used to assess the quality and reliability of the input information and methodology approach. Although the AGSN polygons were inspected using Google Earth, a local expert is also consulted to ensure the validity of the layer. The Brazilian expert on slum mapping with RS techniques confirmed that the 2019 layer is upgraded compared to 2010 and highlighted the conceptual constraints found in literature (Ferreira & Feitosa, 2020).

With respect to the clustering task itself, an expert on area deprivation mapping within ITC was consulted and provided insights on the transferability of the approach. An hour interview was conducted and the main discussion encompassed the importance of the features used and how their contextual aspect influences the model performance. Then, a new (online) meeting with the local expert was scheduled, where she highlighted the little influence of electricity provision on deprived areas for the city and the great importance of green areas, leading to the exclusion of the census variables related to electricity access and to the inclusion of band 5 in a model optimisation experiment.

Table 4. Conducted validation steps in sequential order.

	Validation step	Description
Approach	1. Field experts and local experts consultancies	• Assessment of AGSN layer • Insights on the model transferability • Insights on feature importance
Model results	2. Visual inspection	Assessment using Google Earth and Street View images
	3. AGSN layer overlay	Comparison between resulting cluster types and deprived settlement sizes
	4. Census-based model overlay	Comparison between cluster results with morphological features and socioeconomic ones
	5. Land use layer overlay	Comparison between results cluster types and different land use classes
	6. Expert validation	Assessment of results by local specialist

For the validation of the k-means results, this research evaluates them by different qualitative assessment approaches. The first one is a detailed visual inspection using Google Earth and Google Street View images, comparing how different the clusters are, spotting inconsistencies with ground truth data and providing support for profiling the different deprived settlement types.

Next, three auxiliary layers are used to identify possible relationships and assess the mapping capabilities of the clustering model using statistical assessment. Guided by the work of Friesen, Taubenböck, Wurm, and Pelz (2019), the mapped output is overlaid with the original AGSN layer to observe whether the size of deprived settlements is an aspect related to their precariousness. Then, in consonance with Wurm and Taubenböck (2018), a K-means model is trained only with the census-derived variables and the output map is overlaid to identify relationships between them, regarding morphological and socioeconomic variations.

The land use zoning map of the city is also compared, assuming possible mapping inconsistencies of the AGSN layer, especially in respect to the presence of non-residential land uses within the settlements. Lastly, the results of both models – the EO-based and the one combined with census variables - are assessed by a 15-years experienced local urban planner during a three hour online semi-structured individual interview. Maps, graphs, satellite imagery and street view images are shown, and the main discussion points refer to the emerging spatial patterns, the major influencing factors for their differences, how the clusters types can capture the socioeconomic aspects of the areas they occupy and the applicability potential of the approach. As the results are shown and key questions are provided (Annex 9), the open discussion leads to important feedback that are further presented in this report.

4. RESULTS

This chapter presents the results to identify and characterize types of deprived areas in São Paulo. Section 4.1 illustrates the results of the optimal unit of analysis for this study. Section 4.2 provides the PCA results from the GLCM and LSM features, selecting the optimal moving window and metrics for the model input. Section 4.3 analyses the EDA diagrams and further interprets the input features for the model, followed by the model implementation and optimization results. Section 4.5 presents the results of the optimised k-means model, characterizing each resulting cluster and providing the most significant variables for individual cluster types. In Section 4.6, the results are visually inspected, statistically assessed with other reference layers, and validated with expert knowledge.

4.1. Selecting the spatial unit of analysis

As explained in section 3.4.1, the selection of the spatial unit of analysis relies on the inspection of the AGSN polygons considering the different settlement sizes. Figure 15 illustrates the visual inspection with the four grid sizes. For small slum pockets (less than 1000m^2), the grid size of 10m depicts them well. For medium size (around 40000m^2), a grid size of 20 and 50m looks appropriate for regular shapes. When they have a more elongated shape, a grid size of 20m outperforms 50m. A 100m grid only performs well for very large slums (over 200000m^2), and there are only 51 out of 1,575 polygons, in total. Considering that 72% of the polygons fall under the average slum size and how irregularly shaped the AGSN polygons are, 20m is chosen as the optimal unit of analysis.

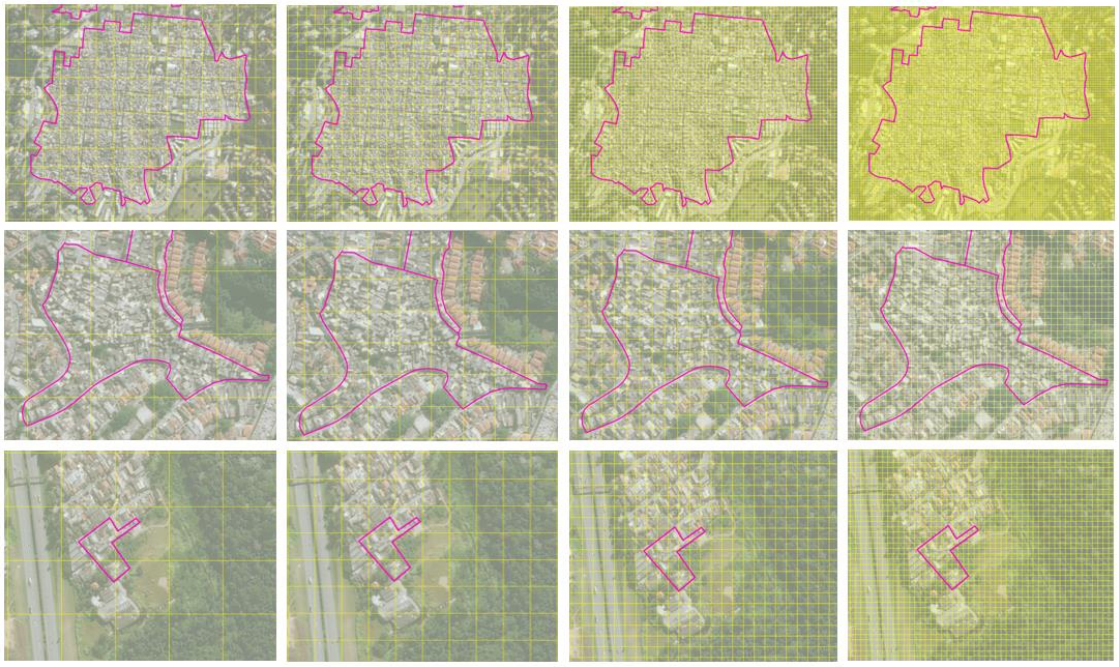


Figure 15. Grid sizes: from left to right: 100m, 50m, 20m, 10m. From top to bottom: large to small settlements.

After inspection, the regular grid base layer is refined: *(1) to reduce computational time.* Only the cells intersected by the AGSN polygons are sampled, which reduces the extent and the size of the file. With this sampling, the layer goes from 8,552,802 to 198,560 cells; *(2) to ensure the homogeneity of the sampled cells.* The layer presents non-homogenous pixels – mixed between deprived and non-deprived areas - mostly on the boundaries of the settlements, which can increase the information noise and mislead the interpretation of the features. To deal with this, Owen & Wong (2013) suggest establishing a threshold to sample the cells. A data-driven

comparison is performed in a macro-approach, comparing the total area of the AGSN layer (60,244,428m$^{2)}$ to the total grid area using three thresholds. Table 5 lists them, indicating that optimally only the cells with at least 50% of their area intersected by an AGSN polygon are included. Figure 16 illustrates how the sampled grid base layer encompasses even the smallest and elongated polygons.

Table 5. Results of data-driven approach with the comparison of area thresholds.

Threshold	Total number of cells	Total grid area (number of cells X cell size)
25%	166,794	66,717,600m^2
50%	149,947	59,978,800m^2
75%	134037	53,614,800m^2

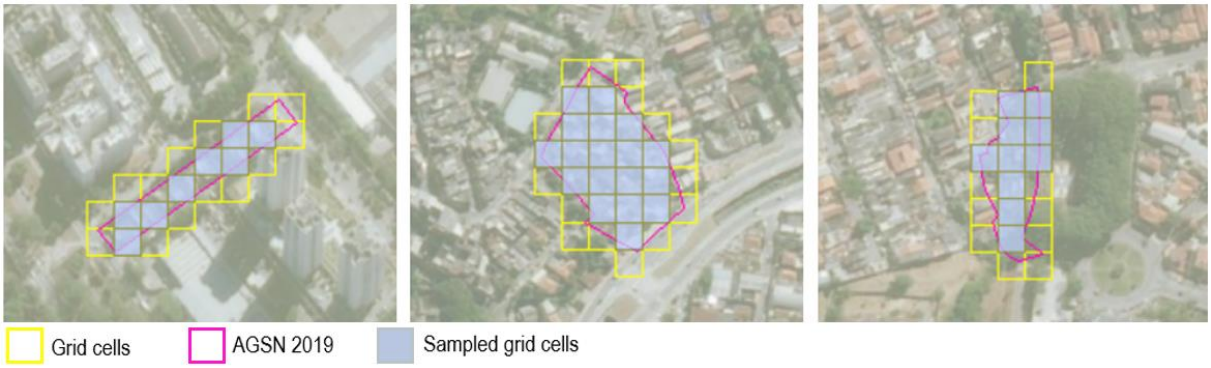


Figure 16. Examples for comparison between the grid base layer before and after sampling refinement processes.

4.2. Selecting hand-crafted features with PCA

The feature extraction process generated 35 GLCM features (seven metrics and six moving window sizes) and 20 LSM features (five metrics and four moving windows) according to the instructions described in Section 3.4.2 from previous studies. The results of the PCA experiments, shown in the Appendix (Annex 5), are interpreted according to the assessment criteria described on page 23.

For the GLCM features, the correlation matrix manifests highly correlated values for entropy and second moment and mean and variance when inspecting the outputs using all kernels as input. When considering the iteration per kernel dimension individually, the 5x5 moving window indicates less correlation among these features than other kernels, while maintaining high loading scores. In all iterations, second moment and contrast features show the lowest communality values and for most of the iterations, contrast accounted for the smaller proportion of model variation. Based on the criteria above, the 5x5 moving window is chosen by deriving seven GLCM features and the model is run once without variance, second moment and contrast features (details in Section 4.4).

Regarding the LSM features, in the model with all kernel sizes, Aggregation Index (AI) and Fractal Index (FRAC) are very highly correlated among each other and with Shape Index (SHAPE). As SHAPE is less correlated to the other two LSM, it is coherent to be maintained for further analysis. Thus, for the final model input, only SHAPE, PATCH_DENS (Patch Density) and AREA_MN (Patch Area Mean) are used as input features for the k-means model to be trained with. Considering each PCA iteration individually, the 5x5 kernel also shows less collinearity patterns when compared to the other kernels.

4.3. Analysing data descriptives

Figure 4 shows that after the selection of the spatial unit of analysis and the 32 features, the feature values are extracted at the grid base layer, integrated and standardized into a georeferenced data frame. This is the main input for the k-means model, and it is investigated with EDA techniques previous to the model implementation. First, the boxplot graphs spotted some extreme outliers in the contrast feature that are

removed and NoData values of population count feature that are handled as zero (Annex 10). Then, the descriptive of each feature are analysed to provide insights on the data patterns and check assumptions that can interfere with the model results. The histograms of the services-related features show a skewed distribution towards lower mean values but with high positive outliers, indicating that most of the data points are fairly close to facilities, but there are some outliers very far apart (Figure 17). The graphs also display a similar distribution to all spectral bands and since they use the same imagery as input, it indicates that the mean values are collinear – which is tested by the correlation matrix. This evidence justifies a model experiment without the spectral bands to test how their collinearity might impact the model results. Figure 17 also shows the strong relationship between higher VIIRS Night Time light (NTL) values and smaller distance to financial facilities, in line with Ghaffarian et al. (2018) and the high entropy and density values inferred in the work of Kuffer et al. (2017). In addition, a model experiment is also conducted without some RS-based features that are highly multicollinear (above 0.8).

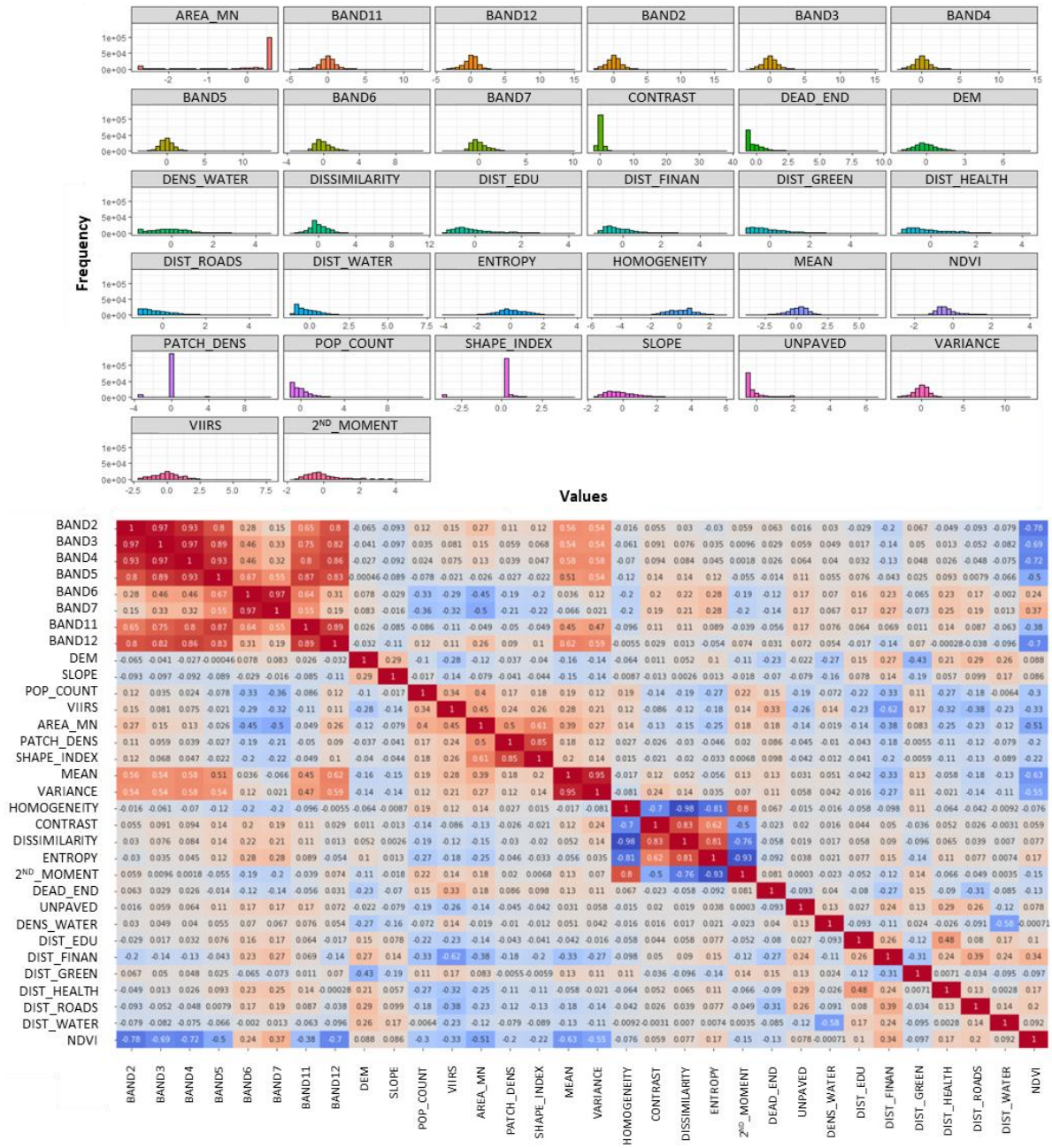


Figure 17. Histograms and Pearson Correlation Matrix.

4.4. K-means model implementation and optimisation

As explained in Section 3.5, K-Means is a data-dependent model, requiring techniques to evaluate its robustness and optimise its performance (Agarwal, Jakes, Essex, Page, & Mowforth, 2018). Seven experiments are conducted with different input features to check the model consistency and achieve the best possible results. Figure 18 shows the workflow of the seven trained and assessed models, identifying the features used in each model, the inclusion or removal of the features and the specific justifications.

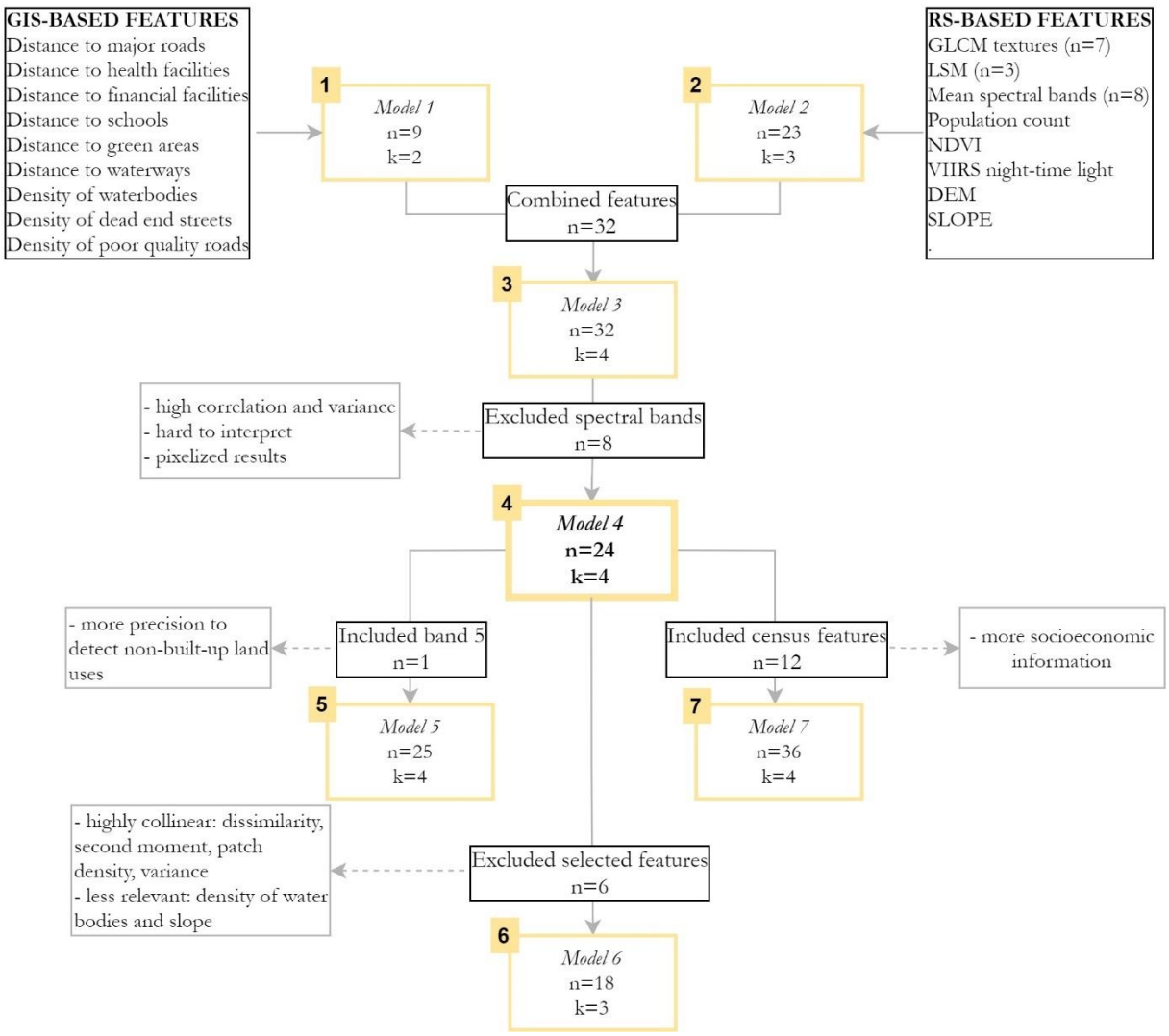


Figure 18. Flowchart of optimisation experiments. Initial 'n' means the number of features and 'k' means the number of clusters used in each model. Model 1 uses solely GIS-based features, Model 2 uses solely RS-based features, Model 3 combines GIS- and RS-based features, Model 4 in bold indicates the optimal model with all features except the spectral bands, Model 5 adds band 5 to Model 4, Model 6 uses only features sampled from features selection techniques and Model 7 combines features from Model 4 with census-based features.

The first two models are used to test the robustness of the algorithm and the sensitivity to the input. Model 1, which uses GIS-based features, provides very similar results in each initialization, showing some robustness. The same behaviour happens with Model 2, which uses RS-based features. Besides the mapped output, the resulting by-product of the feature importance function also shows consistency (Annex 11). In Model 1, distances to services features are very significant in all initializations, while in Model 2, the spectral bands 5, 4 and 3 occupy the highest misclassification scores. When their features are combined into Model 3, results indicate similar spatial patterns to Model 2, which is coherent as the k-means receive more features from it. By inspecting the results of Model 3, the spectral bands had a large variance and are highly correlated

even after standardization. Literature stress that the feature selection tools tend to fit together features with high variance and collinearity, indicating them as highly important to the model generating biased results (Nugrahita & Surjandari, 2020). Besides, spectral bands are not easily interpretable for urban mapping studies (Ozdemir, Mert, & Senturk, 2012). Based on this, Model 4 is trained without the eight spectral bands and inspection shows that results lose the pixelated effects present in Model 3. This pixilation worsens the analysis at the city scale, because the model has a good ability to grasp specific land cover details but poor cluster's generalizability, especially in peripheral areas (Figure 19). Then, Model 5 is trained using only the red edge band because the local expert stated the relevant relationship between the presence of vegetation and the deprived settlements in São Paulo. The results show that this model performs well for green areas but has problems to detect other non-residential land-use types. Figure 19 compares Models 3, 4 and 5 in a peripheral area exemplifying the pixelated output from Model 3 and the lower separability capacity of Model 5 to capture non-residential land uses in Cluster 2 in green.

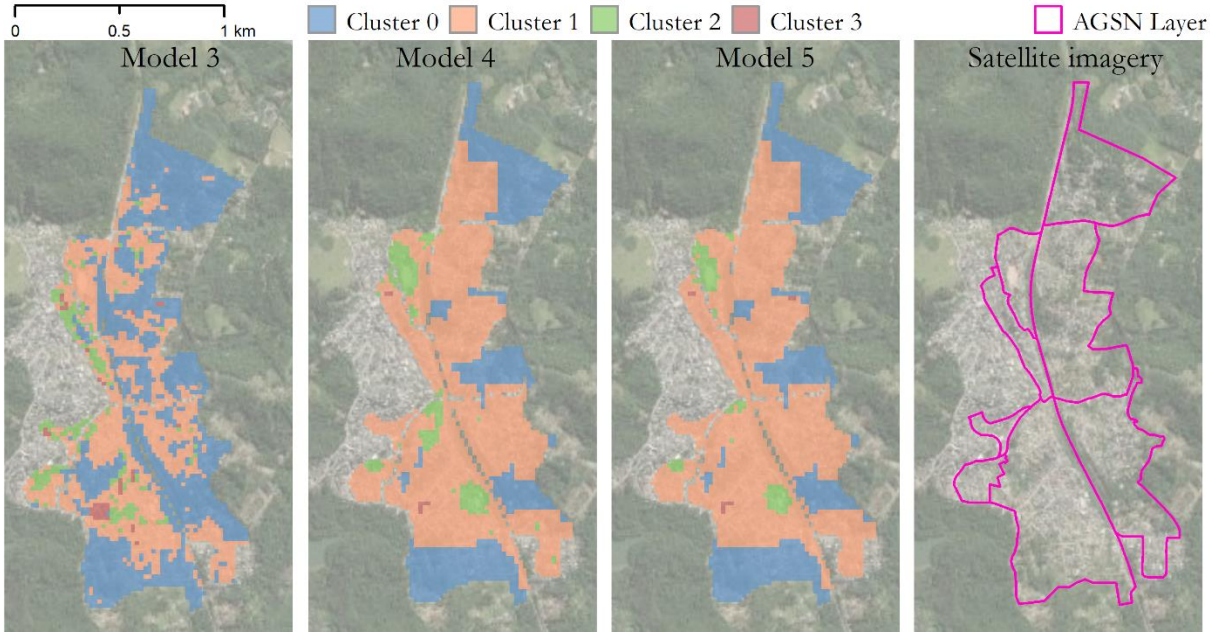


Figure 19. Comparison of clustering results obtained.

To reduce computational time and data redundancy – as several features are highly correlated - Model 6 combines the two feature selection techniques. Based on the correlation matrix in Figure 17, the highly correlated features and the lowest misclassification rates derived from the feature importance tool (Annex 12) are removed. The results show an important aspect of how the algorithm is capturing information. To capture finer-grained typologies and increase their separability, the model requires detailed information. Model 6 is more generalizable than Model 4 but covers finer differences that are meaningful, i.e., the removed features help distinguish the settlements and provide insightful information for the formation of the clusters. For instance, contrast and homogeneity features are removed for multicollinearity reasons, but they are responsible for assigning the non-residential land uses by capturing the fragmentation of the urban fabric, with irregular edges on the pixels. Figure 20 shows the information lost compromised the identification of definitive heterogeneities, especially regarding other built-up land uses. Cells belonging to Cluster 2, which allocates non-residential land uses for both models, is incorporated into Cluster 3 in Model 6, omitting important morphological heterogeneities.

Lastly, Model 7 includes 12 census-derived features to the 24 features from Model 4. By adding this information at the household level, the model can add other dimensions not covered by the morphological and environmental features extracted from RS- and GIS-based data. The results compared to the optimised Model 4 is carefully described in Sub-section 4.5.3.

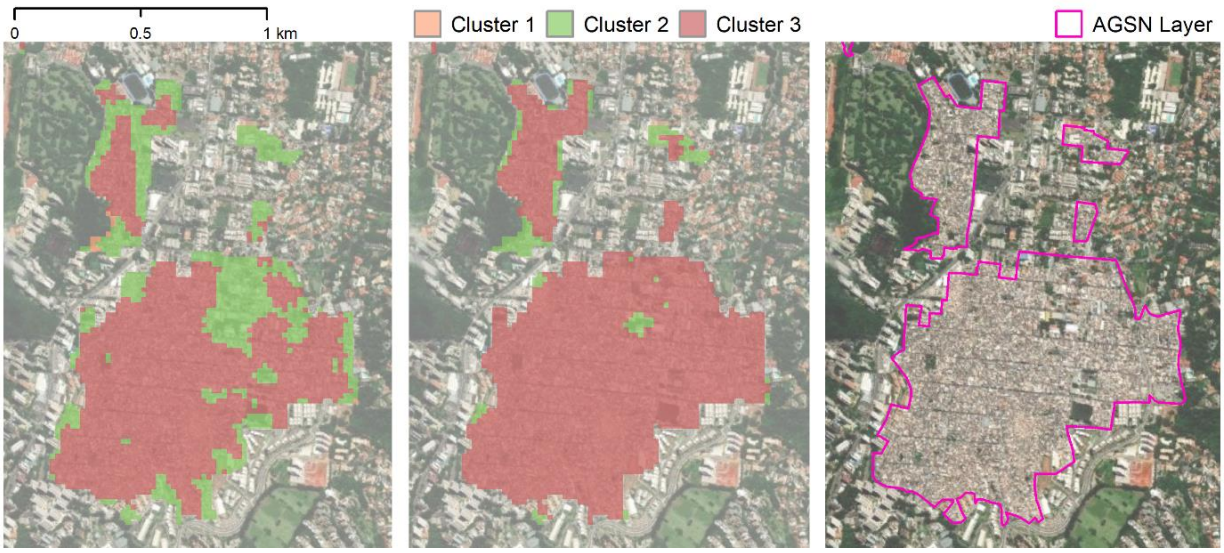


Figure 20. Comparison of Model 4 (left) and Model 6 (middle).

4.5. K-Means model results

This section addresses sub-objective 2, presenting a description of k-means Model 4' implementation and resulting clusters, as, based on the experiments above, it achieved the best performance. Figure 21 shows the WCSS metric calculation indicating 2, 3 and 4 as eligible k values. As Model 1 with only nine features resulted in two clusters, assigning 'k' as two might provide overgeneralized cluster results.

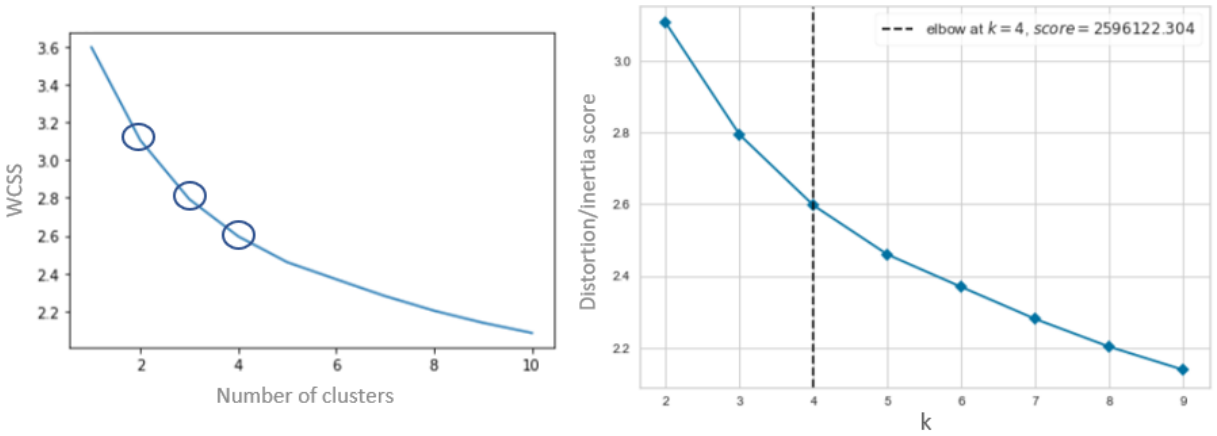


Figure 21. Results obtained from elbow method calculation from different implementation packages. Left: from sklearn package; Right: from yellowbrick.cluster package.

Even though the graph on the right indicates four as the optimal k value, this choice should also be based on subjective interpretation. Thus, the results with three and four clusters are mapped and analysed by visual inspection (Figure 22). Four clusters are chosen as the optimal because it indicates better separability between residential built-up and non-built-up areas, when compared to a k value of three. Green areas are separately assigned to a specific cluster – 1 in orange - indicating better separability.

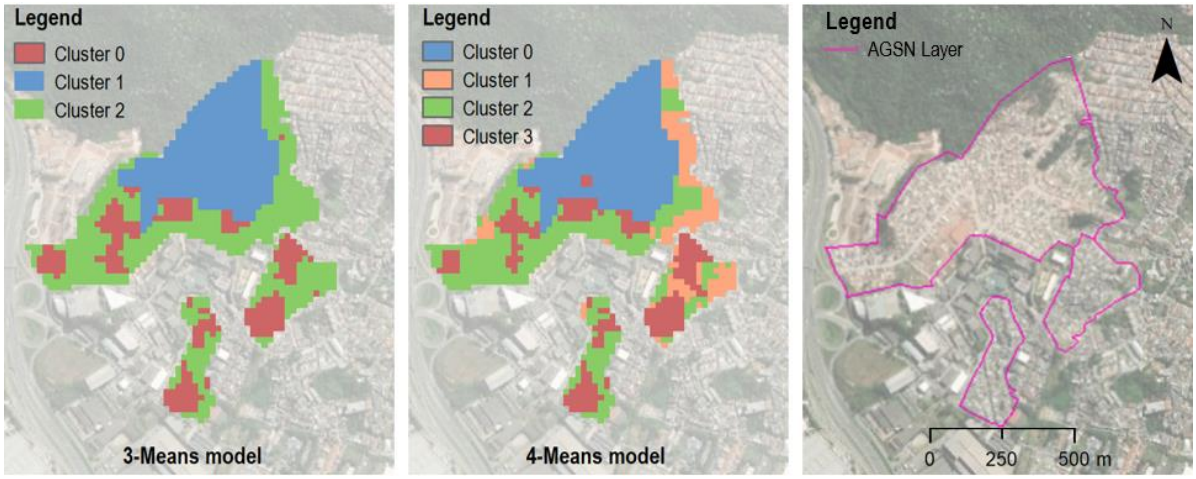


Figure 22. Comparison of results. In the model with 4 clusters, green areas are separated into their own cluster.

4.5.1. Analysing the emerged types of deprived areas in São Paulo

Figure 23 shows the four cluster types of deprived areas distinguished for the city of São Paulo at the 20m grid-level applied on imagery. As stated in Table 6, Clusters 0 and 1 are the ones with less allocated cells and are predominantly located in the city's outskirts. In contrast, Cluster 3 has the highest cell assignment and is more centrally located and Cluster 2 has diversified locations in the urban fabric. Visual inspection indicates that the spatial distribution patterns of Cluster 1 is very near non-built-up areas/rural areas. Conversely, Cluster 3 is the most urban type, even when assigned in settlements outside the central areas. Cluster 2 is often shaped linearly, which might be related to occupations near water- and roadways.

Table 6. Number of cells per cluster.

Cluster	Number of cells	Percentage of total cells
0	9,749	6.5%
1	31,059	20.7%
2	46,638	31.1%
3	62,501	41.7%

The characterization of the emerged clusters and the relationships between them and each feature are analysed with violin plots, illustrated in Figure 24. Settlements assigned to Cluster 0 are characterized by low accessibility to health and finance facilities. As they are located in peripheral areas, occupying elevated topography, they are further from waterways and present a higher DEM mean. They are assigned with the lowest VIIRS values. Thus, lower socioeconomic status, but also the lowest population counts. The lower density of dead-end streets also indicates little access to infrastructure. This cluster also presents the highest NDVI values, indicating the existence of dense and healthy vegetation.

Regarding texture features, Cluster 0 has the lowest mean and variance values suggesting the presence of bare soil or non-built-up areas but also slum-like morphology areas, which is not easily discriminated by satellite imagery (Kuffer, Pfeffer, Sliuzas, et al., 2016). Visual inspection reveals more bare soil land, or what appear to be construction sites than slum-like patterns (e.g., dense, small and irregular buildings). The very low LSM values refer to their peripheral location, close to unpopulated and highly vegetated areas, where the WSF layer cannot identify the presence of buildings or areas where the official census allocates low population counts. These characteristics depict tiny settlements scattered in the peri-urban areas that could be new deprived areas appearing. Thus, Cluster 0 is labelled as "Infant settlements in open spaces".

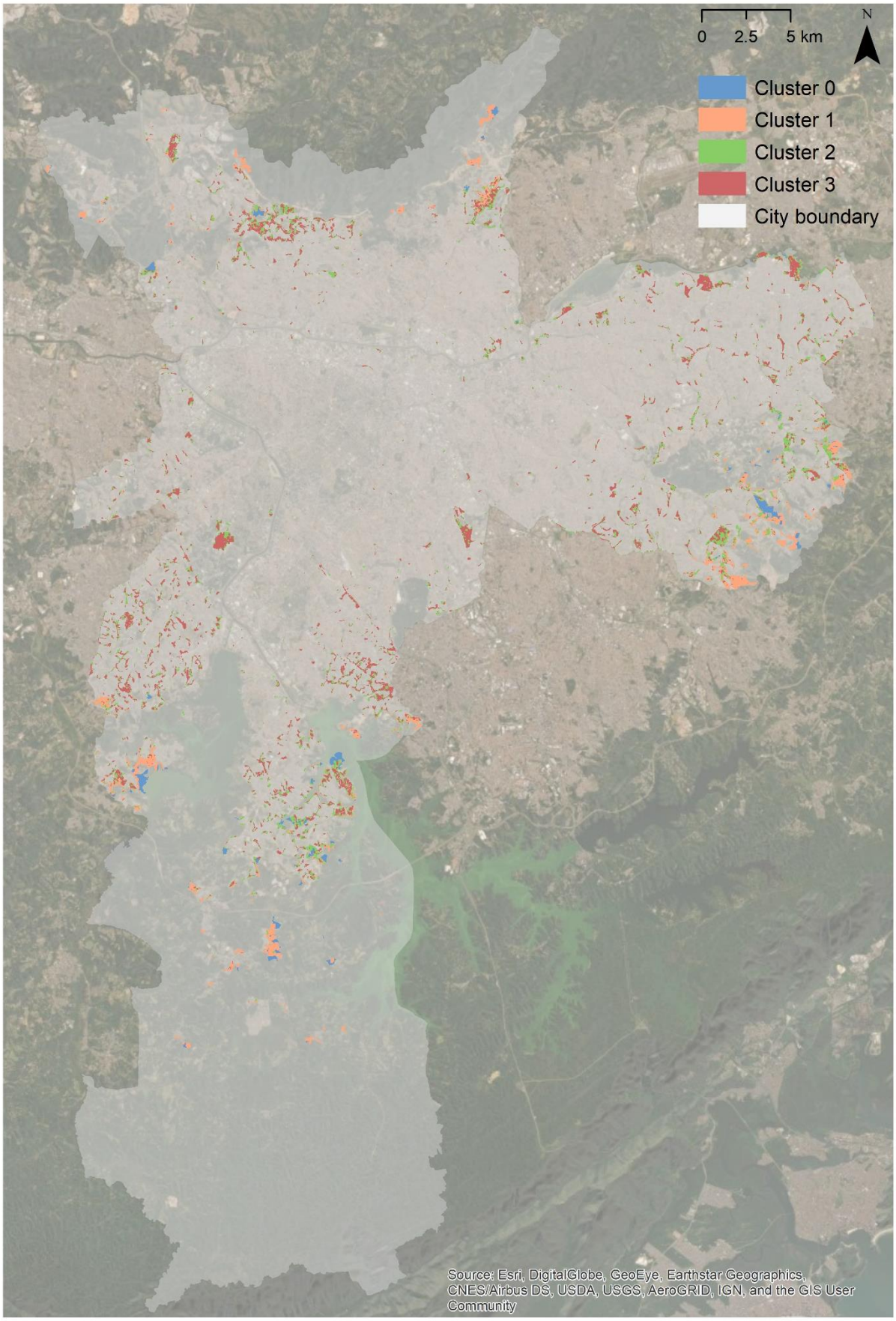


Figure 23. Clustering map of the types of deprived areas in São Paulo.

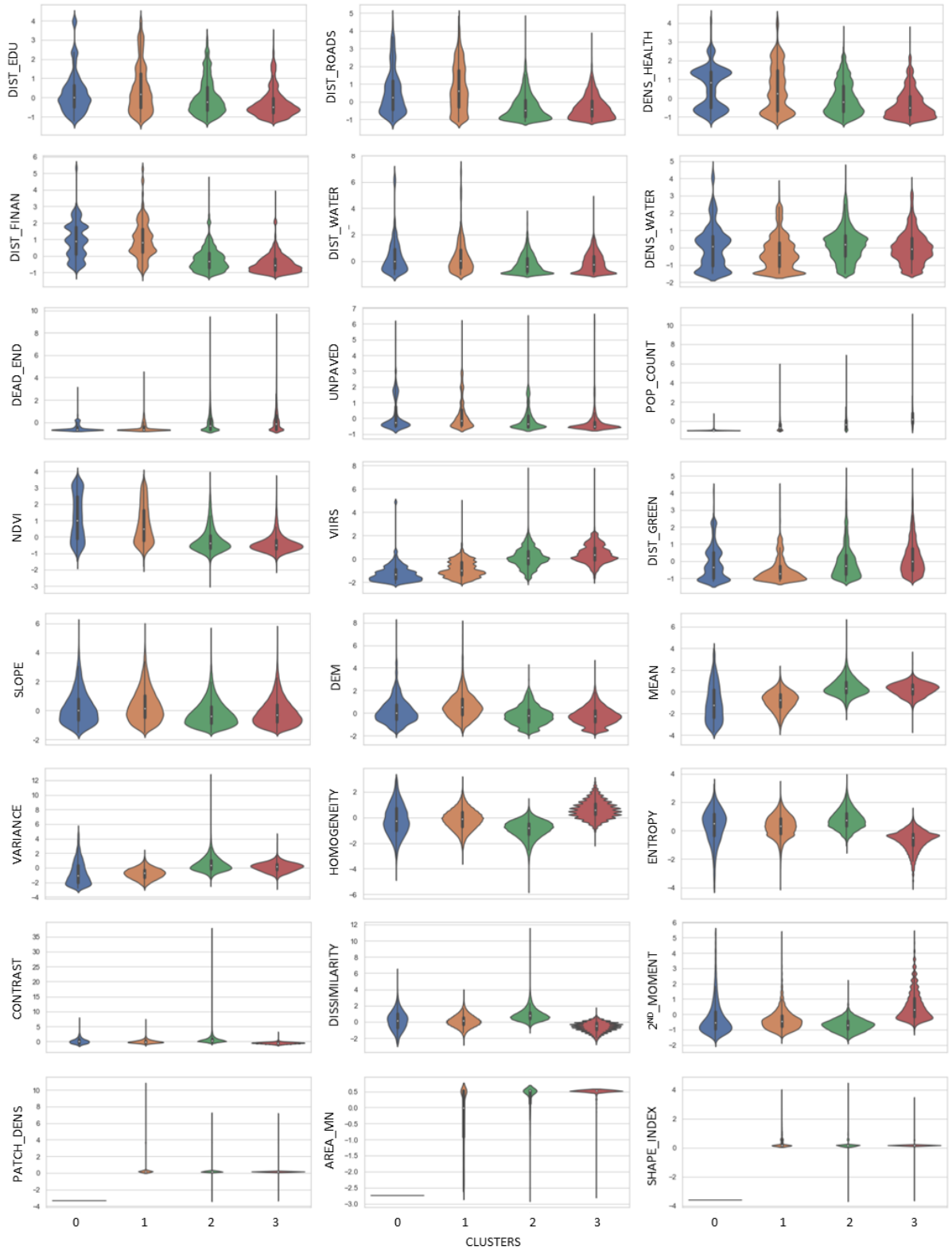


Figure 24. Violin plots with standardized features values per cluster type.

It is possible to see similarities on Clusters 0 and 1, such as the high mean values on accessibility- and topography-related features, but there are also crucial differences. In Cluster 1, the settlements have the lowest accessibility to services, are even further from major roads and have the highest mean density of poor-quality roads. Considering that these settlements have considerably higher population counts than Cluster 0 and apparently less infrastructure, Cluster 1 tends to reflect more precariousness.

The lack of basic infrastructure is also reflected in the proximity to large conservation areas and in the highest DEM and slope values. Texture features generate mixed information, as pixels have high homogeneity and entropy levels, depicting both spatial disorder and uniformity. Visual inspection indicates that this ambiguity might be related to a more organic layout following the landscape, e.g., hills and water bodies, and the vegetated areas that are falsely included in the AGSN layer.

Considering patch density, which consists of both built up and non-built-up pixels, Cluster 1 shows more spatial heterogeneity of the pixels, indicating that settlements tend to be more fragmented. The patch area feature also has greater variability, showing less uniformity in the urban morphology. Higher built-up shape index also implies more irregular structure and less compactness, suggesting a more sprawled form, thus an earlier stage of development (Jochem et al., 2020; Sliuzas et al., 2010). Based on this, Cluster 1 is titled "Unordered and poorly consolidated settlements".

For Cluster 2, cells are closer to services, facilities, roads and waterways. This study included only main water veins, which are mostly channelized along major roadways in non-residential nuclei. Higher density of dead-end streets and population with lower DEM values show higher built-up density in low-lying areas, which, in agreement with the water-related features values, can indicate higher flood risk zones. This cluster type has higher mean and variance values that, together with the lowest homogeneity, suggest higher pixel variability, more edges and less compactness, thus a more regular layout, e.g., nonresidential buildings. This is coherent when associated with the highest contrast, entropy and dissimilarity values, since more formal areas have sharper edges between the building and its surroundings.

The skewed distribution of mean and variance feature values also indicate the presence of both formal and informal morphology appearance in this deprived settlement type. Lower NDVI values and high VIIRS values also contribute to this description, as for instance, industrial areas are less vegetated but wealthier than formal residential areas. Higher shape index values indicate complexity, and the skewed distribution of the patch mean area suggests variability in the patch size. These suggest different stages of development or fragmentation of the urban fabric. Thus, Cluster 2 can be labelled as "Less deprived settlements connected to non-residential areas".

Lastly, settlements assigned as Cluster 3 are characterized by central locations close to financial and education facilities and with the highest density of dead-end streets and population values. The cells present the lowest entropy, dissimilarity and contrast whilst having the highest homogeneity and second moment values. Together with high mean values, texture metrics indicate pixel uniformity, thus a more dense and compact urban fabric and slum-like morphology patterns.

LSM features values shows an overall high patch area indicating less fragmented and dense morphology, while the skewed distribution of shape index relates to more complex and mature settlements. Therefore, Cluster 3 is denominated "Densely urbanized and mature settlements with irregular layout".

It is important to state that the adoption of these titles stresses the main features of the emerging cluster types providing a comprehensive description for the reader. Mean feature values are compiled and presented in Figure 25 to facilitate the comparison among the clusters. For most of the features, Clusters 2 and 3 have considerably similar feature values, except texture features, emphasizing the compactness and uniformity of the settlements in Cluster 3. Conversely, Cluster 0 and 1 have more similar characteristics, but differences are also clearly spotted, especially regarding LSM and NDVI values.

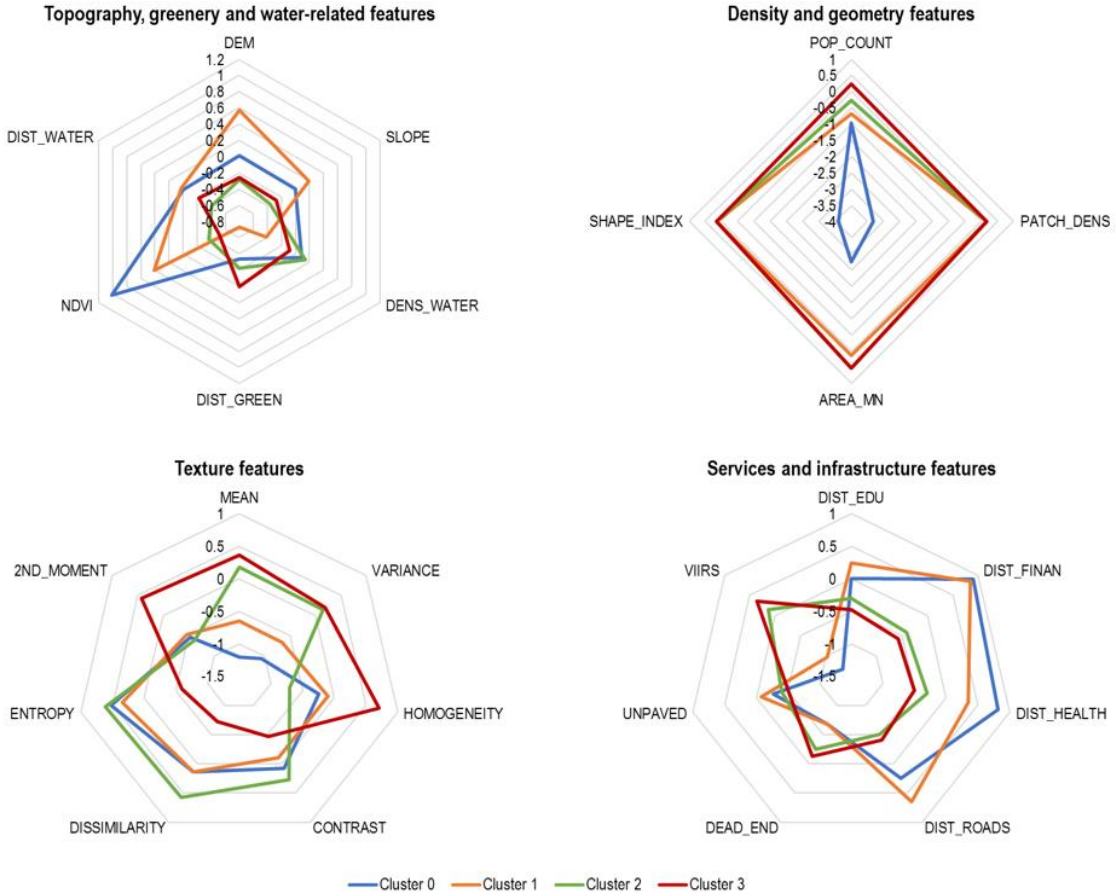


Figure 25. Radar graphs profiling the emerged clusters according to the mean values of each feature.

4.5.2. Assessing feature importance

Besides helping with the model optimisation, the feature importance tool is used to answer the research question on the most significant features to capture deprivation in São Paulo. Overall, the most significant features are three highly correlated ones (entropy, homogeneity and dissimilarity), additional to VIIRS and distance to finance facilities. Figure 26 shows the misclassification rate per cluster and Table 7 summarizes the magnitude of the positive and negative scores of the most important features.

There is a slight prevalence of RS-based features, which is reasonable as they account for 2/3 of features. As the LSM features are correlated and with very low values, these are the features with the highest scores for Cluster 0, together with NDVI and distance to finance facilities. For Cluster 1, accessibility to services, DEM, VIIRS, mean and variance features are the most relevant features, suggesting settlements with low income, sprawled layout, near green areas and the periphery. Cluster 2 and 3 are inversely influenced by the texture features. Distinctively high positive scores to contrast, dissimilarity and entropy in Cluster 2, are related to the presence of formal/ non-residential buildings and consequent pixel complexity. Meanwhile, Cluster 3 presents high negative scores on the same features evincing the most pixel levels are homogenous, hence urban structures with slum-like morphology.

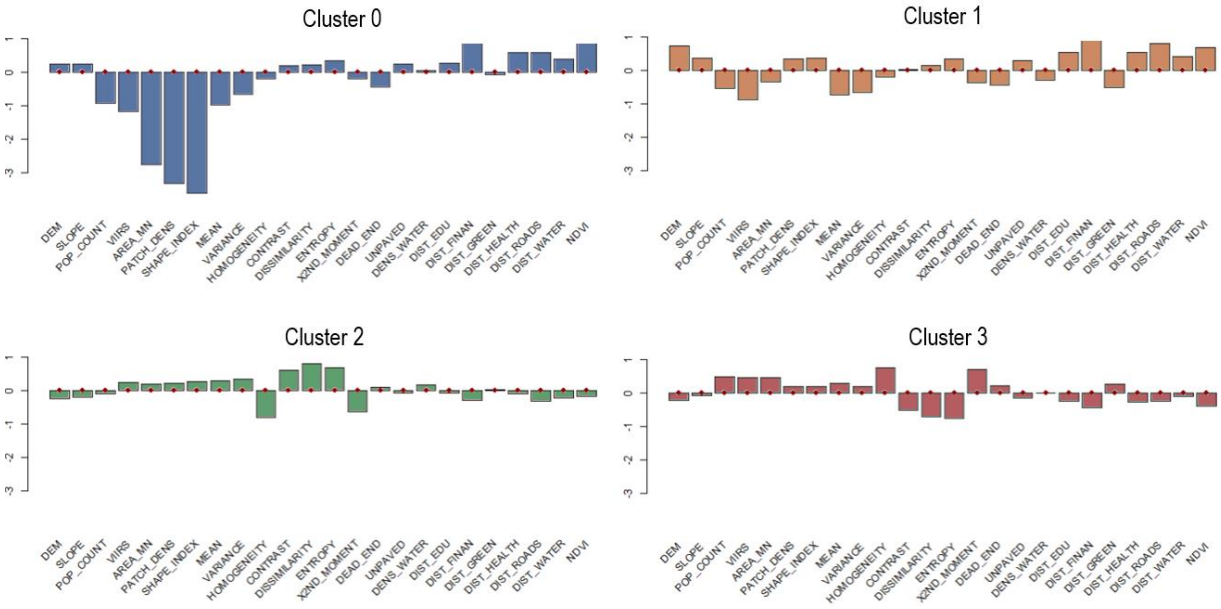


Figure 26. Visualization of EO features using FeatureImpCluster in R per cluster type.

Table 7. Summary of most important features per cluster type in decrescent order of relevance.

Magnitude	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Positive	Distance to financial facilities NDVI Distance to roads Distance to health facilities	Distance to financial facilities Distance to roads NDVI DEM	Dissimilarity Entropy Contrast	Homogeneity Second moment Population count VIIRS Built up mean area
Negative	Shape index Patch density Built up mean area VIIRS	VIIRS Mean Variance Distance to green	Homogeneity Second moment	Entropy Dissimilarity Contrast

4.5.3. Combining census data

This subsection brings an additional perspective to the model by including information at the household level and answers one of the research questions by analysing its contribution to the model. The Sankey diagram represents the cluster cells' flow from the chosen Model 4 only with EO-based features to Model 7, including the census-derived features (Figure 27). The graph shows some consistency and correspondence maintained in the major flows, but there are also explicit divergencies. Half of the cells allocated to Cluster 0 in Model 4 is redistributed by Model 7 into 1. By inspection, Cluster 0 in Model 7 mainly assigns cells with zero or very low population

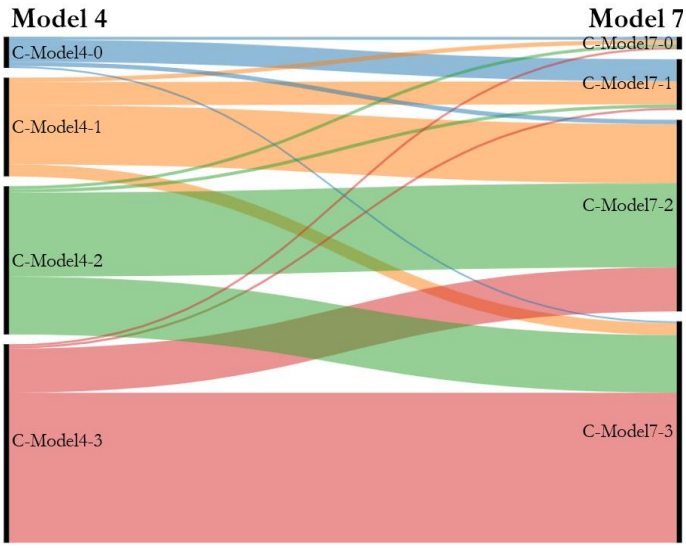


Figure 27. Sankey diagram comparing Models 4 and 7.

allocated by the census, mostly non-built-up areas. This assumption is based on the temporal mismatch between the AGSN layer from 2019 used in the analysis and the census information from 2010. The census-based features have very low population values, indicating that precarious settlements mapped by the IBGE in 2019 are not yet considered in 2010. A major flow of Cluster 1 in Model 4 is being assigned by Model 7 to Cluster 2, which reduces the separability for non-residential land uses. This can decrease the effectiveness of Model 7 to distinguish different land uses compared with Model 4. Figure 28 also indicates this with higher values on the distance to services and lower VIIRS values compared to the radar charts of Model 4 (see Figure 25). Comparing the mean feature values of Cluster 1 in Models 4 and 7, the assigned settlements show higher deprivation levels for Model 7. For instance, in terms of RS-based features, Model 7 presents higher mean and variance, indicating higher built-up complexity and lower mean patch area, which relates to less fragmented structures and more sprawled occupation. By census-based means, the lowest values on garbage collection (HH_GARB), access to water (HH_WATER), average income (AV_INC) and homeownership (HH_PRIV) indicate the highest level of precariousness and unconsolidated development.

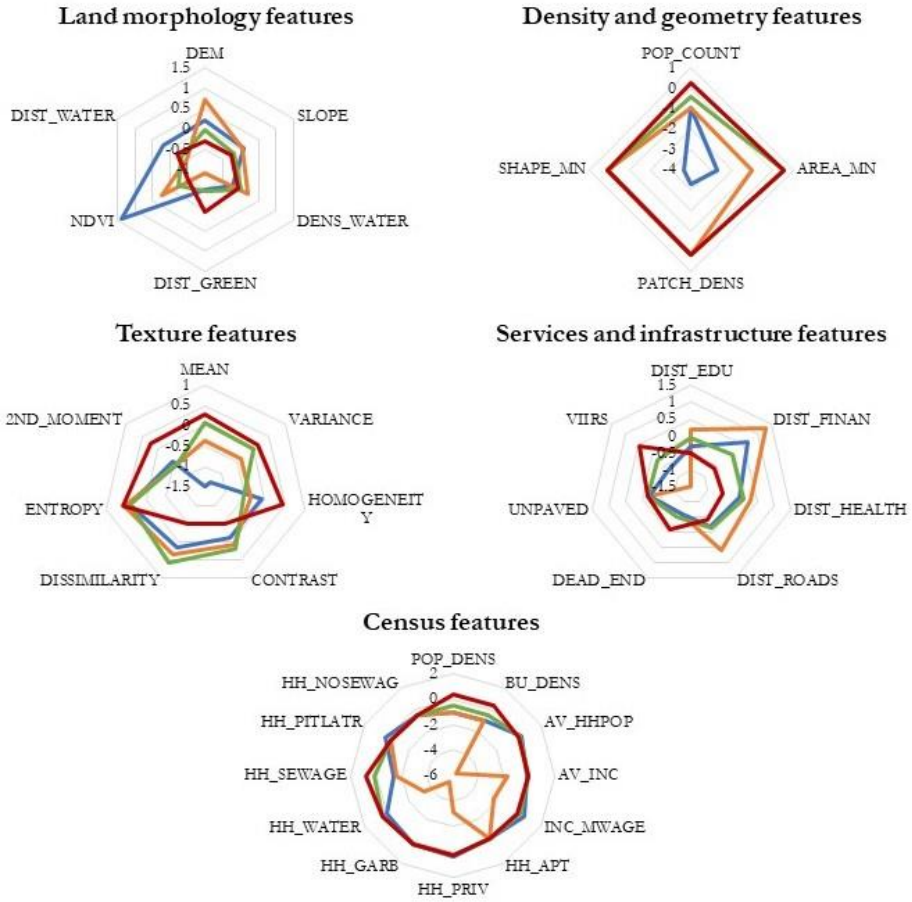


Figure 28. Spider graphs profiling the emerged clusters according to the mean values of each feature.

Cluster 3 in Model 7 shows the least deprived type according to the census-derived features. It presents higher entropy values that suggest irregular edges and more pixel variability, which adds a different dimension than Model 4. It also indicates that Clusters 2 and 3 in Model 7 have more similar characteristics than depicted in Model 4. The census-derived feature values suggest that their main difference relates to the higher built-up and population densities of Cluster 3. Figure 29 compares Models 4 and 7 highlighting the shortcomings of adding the census-derived features. Example 1 illustrates how Model 7 depicts the discrete boundaries of the census enumeration units, which adds spatial fallacies to the disaggregated approach. Example 2 shows Cluster 3 being incorporated by Cluster 2 in Model 7, which can reduce the model ability to capture specific land uses.

Figure 29. Comparison of model outputs. Left: Model 4; Middle: Model 7; Right: Satellite imagery.

4.6. Results evaluation

Validation of the k-means results (Model 4) follows a series of qualitative assessments. First, the model is visually described and evaluated with satellite and street-view images. Next, it is statistically and quantitively assessed with reference data and finally validated by a local expert.

4.6.1. Visual assessment

This first section provides an in-depth description of the emerged cluster types, evaluating them visually by comparing them with satellite and street-view images. Though field verification would be optimally conducted - not possible due to COVID travel restrictions - imagery can show the morphological differences and how they are depicted on the ground, providing support to the model validity.

Figure 30 presents a zoom into the largest deprived settlement in São Paulo, called Paraisópolis, located in the central area, southwest of the city core. It is a very densely urbanized area with a fairly regular layout and some infrastructure provision. Examples A and B show two street corners with distinctive morphological patterns, that the model can differentiate in the same settlement. Thus, it can emphasize the relevance of using a grid-based analysis to capture the intra-urban variations of deprived areas.

Image A exhibits, as representative features of Cluster 3, narrow streets, dense and small buildings occupying the entire plot with predominantly residential use. Meanwhile, Cluster 2, in example B, is characterized by non-residential or mixed land uses, larger building sizes and less precarious infrastructure as the presence of side- and crosswalks.

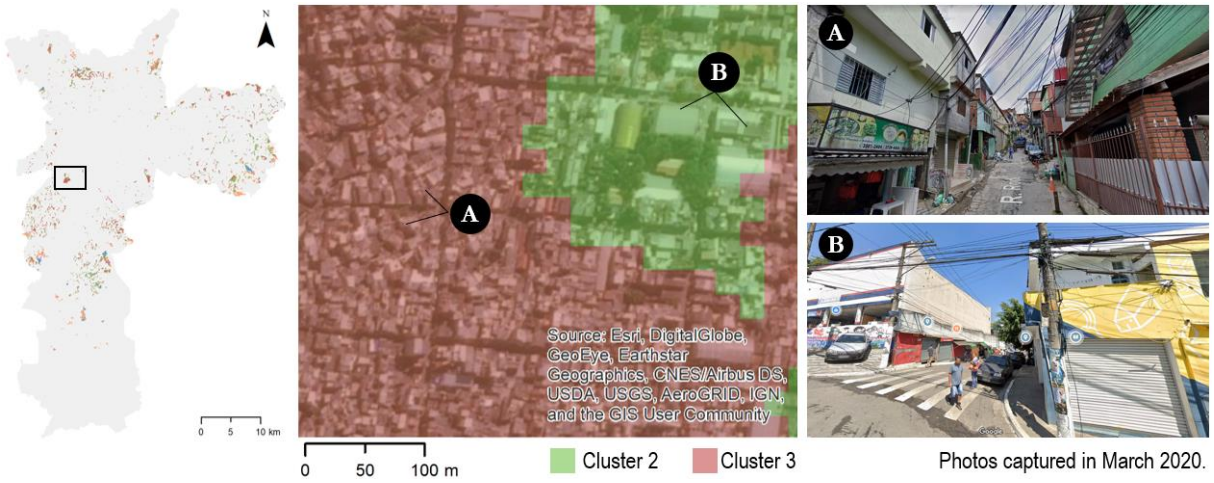


Figure 30. Visual assessment of a central area using satellite and street-view images. Source: (Google Maps, n.d.).

Figure 31 presents a deprived settlement named Jardim da Serra, in the far north of the city, showing the assignment of Clusters 0 and 1 in low-accessibility and density areas. While image C shows a settlement in hilly and vegetated terrain, with precarious buildings and infrastructure, assigned as Cluster 1, image D displays Cluster 0 with only a few houses on the top of a smaller hill in front of Cluster 1 on the background. These examples reflect the hazardous element of Clusters 1 due to its topographic location and the infant stage of development of Cluster 0. There are low population counts and most of the surrounding is comprised by open green spaces with little or no urbanization resources.

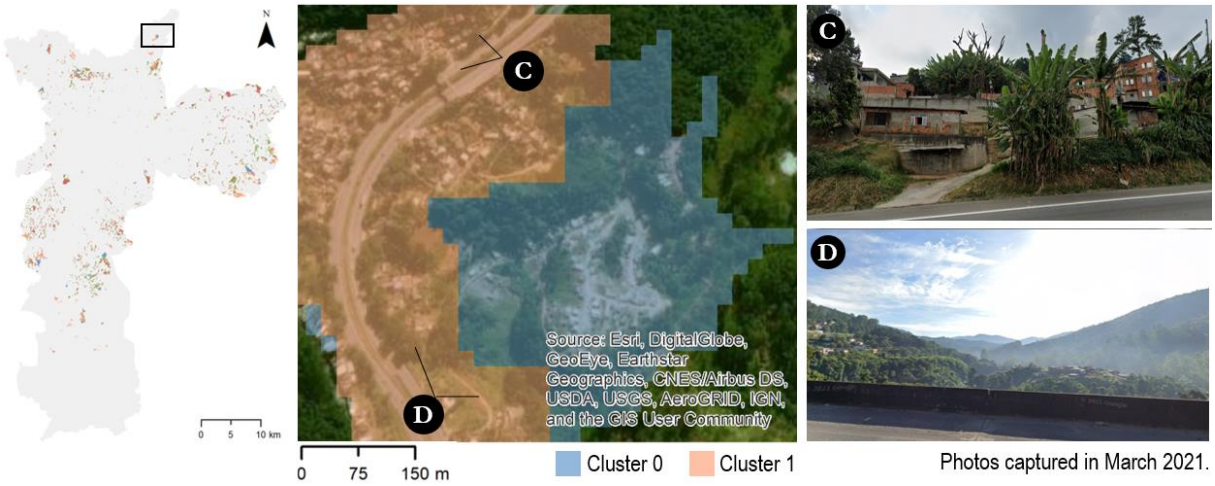


Figure 31.Visual assessment of a northern area using satellite and street-view images. Source: (Google Maps, n.d.).

In Figure 32, image E shows a settlement called Chácara Florida, in the far south of the city, showing that the model can catch changes that from manual image interpretation might appear similar. Clusters 1 and 2 have similar spatial morphology. Still, from street-view, the buildings from Cluster 2 have a more formal aesthetic. In contrast, the immediate units assigned to Cluster 1 already show buildings with lower height and raw construction materials.

Image F is used to reflect the degree of intra-cluster separability. The site assigned as Cluster 2 has a higher level of precariousness compared to the one in image E, in terms of infrastructure provision, environment and dwelling unit. These differences point to the morphological nuances existing in a single cluster and how the model is applying the clusters' portioning. The model can differentiate larger building structures – visually recognized as a mechanical shop and a bar - assigning as Cluster 2 just as in Figure 30 and assign low-density areas with vegetation to Cluster 1, just as in Figure 31.

Figure 32. Visual assessment of southern area using satellite and street-view images. Source: (Google Maps, n.d.).

The nuances described above are visually assessed via street view images in Figure 33, using two examples of each cluster in different locations. While the building sizes and topography can vary in Cluster 0, the settlements are regularly marked by open surroundings, poor or no infrastructure provision and low built-up density. In contrast, Cluster 1 is commonly located in hilly terrain, with poor service provision, while varying the provision of basic infrastructure – highly dependent on the location, i.e., central or peripheral. According to the land use, settlements are assigned as Cluster 2 swing between medium to high-density built-up morphology, varying from small to large buildings. Overall, these areas are less deprived of infrastructure and service provision, such as public transportation. In Cluster 3, the most predominant feature is the high population density with multistory buildings over three floors. The nuances often happen in services and infrastructure provision, because some settlements already participate in upgrading programs.

Figure 33. Examples of deprived settlements in each cluster type. From upper left to bottom right: Clusters 0, 1, 2 and 3 with their respective representative colours, settlement name and capturing date. Source: (Google Maps, n.d.).

4.6.2. Comparison with AGSN boundary

The AGSN polygons have a wide range of area extents, indicating that deprived settlements in São Paulo can vary significantly in size. Table 8 provides an overview from summary statistics of the area of each polygon. The polygons are overlaid with the output map, and the analysis reveals information about the number, distribution and size patterns of these deprived settlements. Since the research uses a gridded approach and different cluster types can be assigned to a single settlement, a 50% area threshold was established to define when a settlement belongs to each cluster type. Three thresholds are explored (50%, 60% and 75%) but 50% was optimal to keep most of the information. Table 9 shows the statistical analysis results, comparing the clusters assignment from Model 4 according to five categories of settlement sizes, derived from Table 8 and defined in labels of Table 9. For instance, tiny settlements are smaller than Quartile 1 and huge settlements are larger than the maximum area statistics value. Table 9 confirms that more than 50% of the deprived settlements are tiny and small occupations and only 7% are assigned as huge-sized. It also indicates that Cluster 0 is mainly allocated for large settlements and Cluster 2 is predominantly assigned for tiny ones, making sense with the clusters' features. While Cluster 0 is located in the periphery, surrounded by non-built-up areas incorporated as commission errors, Cluster 2 is often located in dense non-residential regions. Meanwhile, Cluster 3 is more likely to be allocated in small- and average-sized settlements. As stated in Table 6, it accounts for over 40% of all cells, thus encompassing the most recurrent settlement size. The analysis of Cluster 1 suggests another level of precariousness to this cluster type, because the settlements are not only in the periphery but also tiny and scattered on the landscape.

Table 8. Area statistics summary of the AGSN polygons. All items in m².

Quartile 1	7,668
Quartile 3	42,748
Interquartile range	35,080
Max	95,368
Min	0
Median	19,186

Table 9. Inspection of the number of settlements per cluster type according to five categories of settlement sizes.

		AGSN settlement sizes				
Settlements per cluster type	**Cluster**	**Tiny (a<Q1)**	**Small (Q1<a<median)**	**Average (median<a<Q3)**	**Large (Q3<n<max)**	**Huge (n>max)**
	0	2	0	3	7	7
	1	26	47	46	34	34
	2	177	126	100	49	24
	3	186	201	200	135	43
	Sum	391	374	349	225	108
	Relative	27%	26%	24%	16%	7%

4.6.3. Comparison with a census-based model

There is no direct reference data to validate the results found in this research. Thus, a clustering model using only the 12 census variables (Annex 7) of Model 7 is built, named 'Census-based model', serving as an estimation of the current living conditions. The model is trained with two clusters, derived from the elbow method, and the result is mapped and compared in Figure 34. The illustration below indicates that 2/3 of the cells are assigned as 'less deprived' and occupy central areas, while 1/3 are the 'more deprived' ones and are mostly located in the periphery. The clusters are labelled 'Less deprived' and 'More deprived', considering the mean feature values for each cluster. The radar graph shows the differences between them (Figure 35).

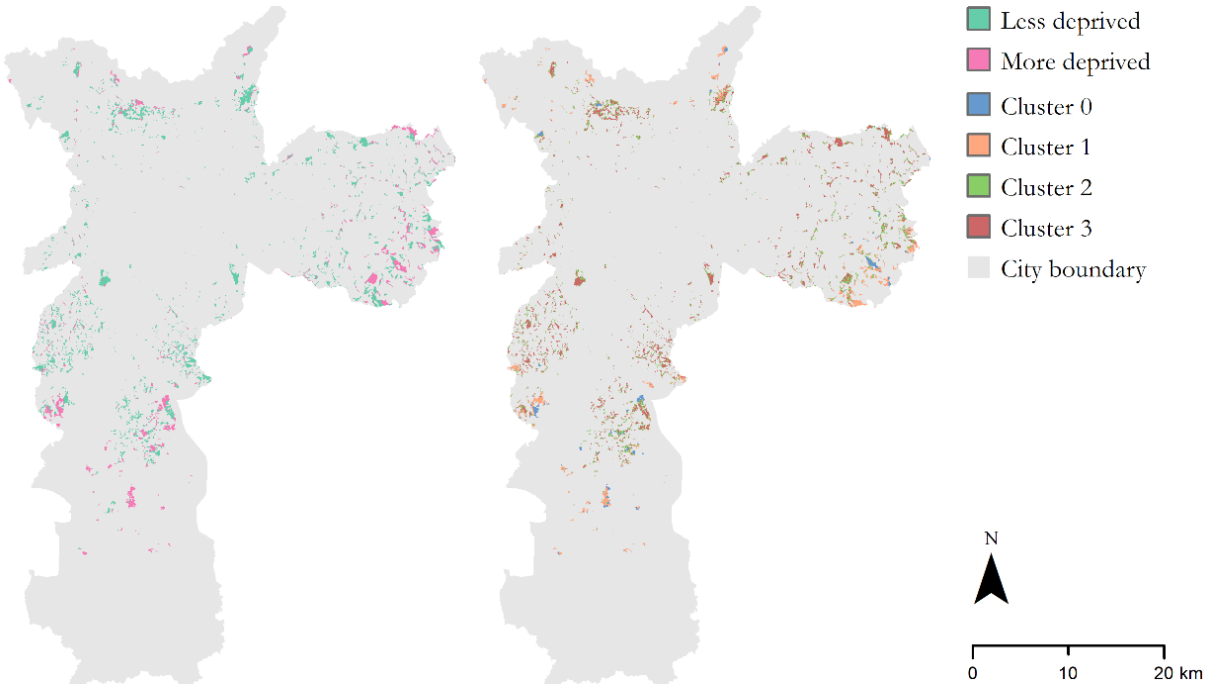


Figure 34. Comparison between the Census-based model (left) and Model 4, EO-based (right).

Figure 35. Left: Radar graph with mean features values of census-derived model. Right: Stacked bar graph showing the distribution of each cluster type per census-based category.

The more deprived cells are characterized by more households earning less than minimum wage (INC_WAGE) and lower average values of income (AV_INC). They refer to less dense and populated settlements (lower values of POP_DENS and BU_DENS) with lower access to basic infrastructure. The difference between values related to sanitation features is very striking between the two clusters. The cluster presents lower values for households with adequate sewage connected to the infrastructure system (HH_SEWAGE) and higher values for households with sewage discharged in septic, river or lake (HH_NOSEWAG).

The stacked bar chart, also in Figure 35, shows the comparison of the distribution of the cluster results per census-derived clusters. Among all, Cluster 1 in Model 4 seems to be the more deprived one. Based on the census features used, these settlements are identified by sparse urbanization, peripheral location without the basic service provision and lower access to opportunities, which aligns with the characteristics captured by Model 4, described in subsection 4.5.1. Though not considering physical characteristics and urban form as EO-based features depicted, the census-based model also detects higher levels of precariousness in Cluster 1 compared to the other three clusters.

4.6.4. Comparison with land use layer

This sub-section overlays the results of Model 4 with the municipal land use map. To facilitate comparison, the initially 30 land use categories are aggregated into six broader classes: (a) *Mixed*-use that includes residential and commercial and residential and services; (b) *Conservation areas* that encompass natural parks but also areas for agriculture and agrotourism activities; (c) *High-income residential* uses encompass formal areas with different densities; (d) *Low-income* areas that are subjected to public upgrading plans and interventions; (e) *Commercial and industrial land uses*; (f) *Transport/corridor* land uses include adjacent areas to structuring roadways with medium density aiming at requalifying urban spaces of the city. The stacked bars in Figure 36 visualize the different land uses across the four cluster types of Model 4.

Cluster 0 is highly occurring in conservation areas, which are often located in the outskirts of São Paulo. Only a few cells are located in residential land uses, which is another indication of the inclusion of open spaces in the AGSN layer. The number of cells in the commercial land uses was surprising. By visual inspection, most of the cells allocated in this land-use are found in large industrial areas and few cells are found in commercial areas. This makes sense with the peripheral location Cluster 0 occupies. In Cluster 1, most cells are located in conservations areas, low-income residential and mixed-use, respectively. This relates to the early-stage development characteristic of the type, often scattered adjacent to non-built-up areas.

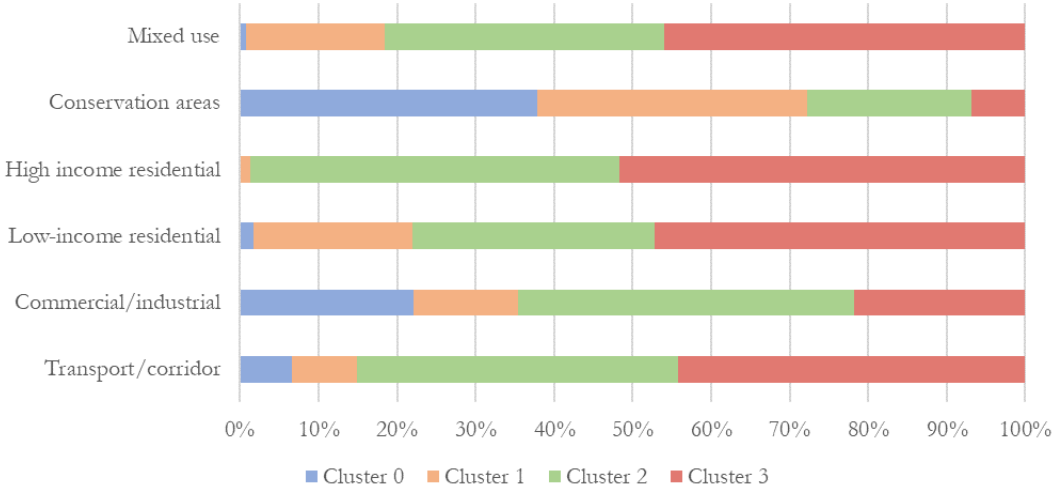


Figure 36. Stacked bars showing cells count per cluster type on the different aggregated land uses categories.

In Cluster 2, a relatively even distribution can be noticed among different land-use types, which is compelling to their diverse location. Most cells occur in transport/corridor, mixed, commercial and industrial land uses, highlighting their presence in non-residential areas as Model 4 suggests. Since the number of cells per cluster differs considerably, Figure 37 presents the relative values per cluster. It shows how most of the cells are still occurring in low-income residential areas, evincing a more consolidated occupation when compared to the previous cluster types. For Cluster 3, most cells occur in residential, mixed and transport/corridor land uses, respectively, with only a few cells in conservation and industrial areas. It agrees with the central distribution of the settlements. Figure 37 also shows that 80% of the cells occupy residential land uses. Overall, the results outline that Cluster 0 and 1 often occupy open green spaces in the periphery and less structured areas to receive residential land uses. This indicates features of early-staged pocket slums with a lack of basic service and infrastructure provision, which are important aspects of area deprivation.

Meanwhile, Clusters 2 and 3 often belong to low-income residential areas and mixed land uses. Since most cells occupying high-income residential areas belong to Cluster 3, and there is no direct information on the density or building height of this land use, it is inferred that this cluster is the densest one. Formal central areas in São Paulo do not provide intra-urban open spaces to allow informal occupation of the settlements and if the AGSN layer maps them, they have more than 51 dwellings.

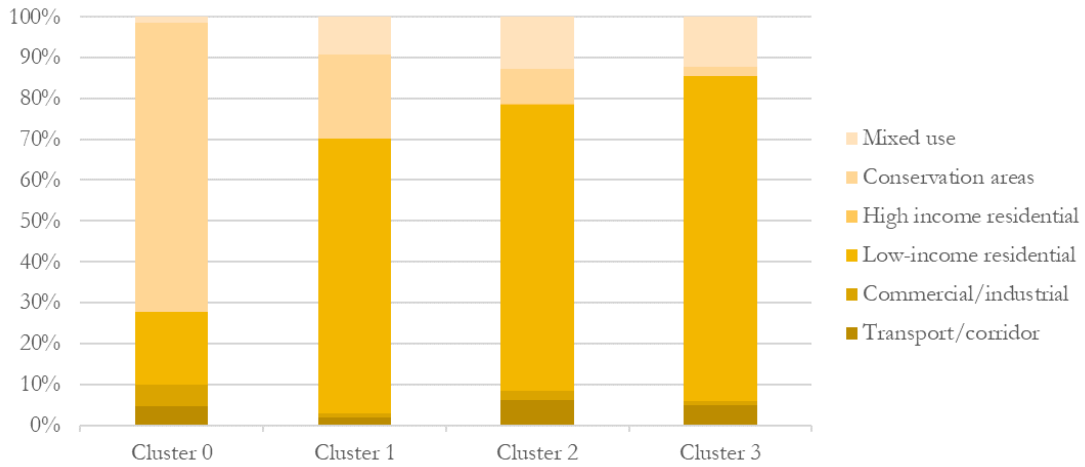


Figure 37. Stacked bars showing the aggregated land uses per cluster type. The high-income residential land use cannot be seen in graph because the percentage values are too small.

4.6.5. Expert validation

Lastly, the results of Model 4 are presented to a 15-years experienced urban planner, who is head of the housing department in the municipality of São Paulo, through an online semi-structured interview. Overall, the expert confirmed that the emerged cluster types could capture the diversity of the deprived settlements of São Paulo. There are different levels of precariousness between the neighbourhoods, with evident contrast between central and peripheral areas and the different stages of development depicted by the model. She confirmed that Cluster 0 and 1, the peripheral settlements sparsely distributed in the landscape, are occupations in infant stage of development, and they often face the most precarious conditions. In the past decade, there are several newly occupied areas, scattered in the greenery with little or no service and infrastructure provision. The possibility of capturing these developing deprived settlements with EO was seen as of great usefulness for the municipality.

In addition, she highlighted the significance of the model differentiating the settlements within or adjacent to open green spaces through Cluster 2, because of the commission issues from the AGSN layer. Settlement sizes are overly estimated, and the model can fairly capture this, though with potential for improvement. It is valid to notice the experiment (Model 5) with spectral band 5 as an attempt to increase this separability between built-up and non-built-up areas. The local specialist also implied the interference of the land morphology to the location of the settlements. In the city centre, most deprived areas occupy low-lying areas and waterways, which are highly polluted in São Paulo or industrial areas. In the periphery, these settlements occupy hillsides, settle within/nearby conservation and rural areas and are often less populated, as most job opportunities are enclosed in the city centre.

Besides, as reported in the previous subsection, the expert knowledge acknowledges the existence of other land-use types mapped by the AGSN layer, often assigned to Cluster 2, and confirms the model validity to this differentiation between these areas and the residential ones. The expert also indicates that the results of the clustering model presented are promising and can be useful to not only São Paulo but other municipalities in Brazil, even suggesting further presentation of the results to the national demographic bureau (IBGE), especially over a full automatization of the workflow.

5. DISCUSSION

This research aimed to build an unsupervised machine learning model to characterize deprived areas at the city level based on their morphological and environmental features using solely open geodata. This chapter provides an interpretation of the findings of this study, starting with a summary of the research outcomes and their prospective applications. Based on the specific research objectives and questions stated in sub-section 1.3, the following sections reflect the main spatial characteristics of deprivation in Sao Paulo and the proposed methodology. Lastly, the pros and cons of the approach and limitations of the study are stated.

5.1. Applications of the proposed model

Through this study, the clustering model distinguished four types of deprived areas in the city of São Paulo from the extracted morphological features. The intraurban diversity is spatially assessed through internal and external validation. From these assessments, it is possible to notice a good correspondence of the modelling results and diversity of deprivation patterns in the city. Table 10 profiles each cluster type describing their typical structure based on their most important features. It summarizes the results comprehensively for end-users, facilitating their information for the decision-making process.

Table 10. Profiling of cluster types with their main characteristics. Source of Ground Views: (Google Maps, n.d.)

Cluster type 0: Infant settlements in open spaces	
Peripheral, low access to services and infrastructure Low-density, low-income settlements Fragmented and scattered in open green spaces	
Cluster type 1: Unordered and poorly consolidated settlements	
Low accessibility to services and infrastructure Steep slope, irregular and poor consolidation Low-medium population density, lowest income level Presence of vegetated areas	
Cluster type 2: Less deprived settlements connected to non-residential areas	
Medium-high population density Fragmentation of residential and mixed land uses Better accessibility to service and infrastructure	
Cluster type 3: Densely urbanized and mature settlements with irregular layout	
Higher population density, higher income levels Complex layout, mature development Better accessibility to services and infrastructure Low lying areas, further from green areas	

The applications of the resulting typologies are manifold. The gridded output can be used as complementary references to land use maps. Often, the dynamics of land-use change are assessed by the market logic, but together with the outcomes of this research, zoning maps can be structured by focusing on the living conditions of the deprived population, aiming at an equity logic that considers the urban poverty patterns (Schindler, 2017). The typologies can also inform disaster risk policies, demonstrating when the most

vulnerable population is located and guide the distribution of resources to the most urgent situations (Ghaffarian et al., 2018). Moreover, the captured spatial patterns can provide insightful information for direct pro-poor policies, e.g., slum upgrading programs. Variations on location, topography, density, layout, accessibility and infrastructure provision described for each cluster type can help target each settlement's demands accordingly, promoting tailored interventions for both complex mature settlements and newly occupied areas (Olthuis et al., 2015). Besides, the typologies highlight investment gaps between central and peripheral areas, suggesting capacity strengthening or decentralization of resources (Friesen et al., 2019).

5.2. Deprived areas in São Paulo

This research acknowledges the inherent contextual and complex nature of deprived settlements (Kuffer, Pfeffer, et al., 2017) and the consequent impossibility of establishing a reference standard to be plead worldwide (Gevaert et al., 2019). While current literature aims at a universal concept for deprived areas, which often leans on their similarities and overlooks their diversity (Ajami et al., 2019), the present study does not attempt to tackle this wicked conceptualization problem. Instead, it proposes more understanding of the local manifestation of deprivation and poverty by developing a list of potential spatial features that can reflect the different living conditions of deprived communities in São Paulo, as well as a workflow for extracting these features using solely freely available and consistently updated data.

In this context, one of the great difficulties for characterization studies to be replicable and practical for LMICs is the lack of consistency and availability of data. Traditionally, research uses census data with aggregated information at the administrative unit level. Specifically in Brazil, the country has a good database from the last conducted census in 2010, with spatial and tabular data with full spatial coverage. However, besides happening only once a decade, which is incoherent with the constant changes and dynamics of deprived areas, the census tract units vary considerably in size, which can generate stigmatisation of the deprived settlements. Considering these issues, the GIS- and RS-based features used are acquired from open geodatabases to ensure their applicability in other contexts. Datasets were considered regarding their source, spatial coverage, resolution and biases (this one in particular for crowd-sourcing data).

Based on previous studies, the characteristics of deprived areas were conceptualized as built-up and land morphology domains and their respective dimensions. Their employment did not only guide the feature extraction process but mainly improved the understanding of how the morphological characteristics of the settlements vary across different locations and affect the planning and intervention decisions. For instance, the high density of dead-end streets is related to the high built-up density, especially in Cluster 3, which increase the costs of intervention correlated to the infrastructure. On the other hand, even with a sparser layout, the hilly and peripheral aspect of Cluster 1 might require even more costly infrastructure investments from the municipality. The knowledge on these nuances reveals that the significance of the dimensions differs from settlement to settlement, which can support planning and decision-making processes, and that the derived features are reflective of the living conditions of these communities in São Paulo.

Notwithstanding the relevance of Model 4' outcomes and the challenges on census data, this research also understands that some socioeconomic characteristics of deprived areas are not depicted by EO-based data. The work from Thomson et al. (2020) already pointed out how the deprivation levels of the dwelling and its neighbourhood can vary considerably. Thus, besides the GIS- and RS-based features, this study also incorporates census-derived features in Model 7, including information at the household, that provides an additional and important perspective to the characterization analysis. As presented in Section 4.5, combining these features articulated both physical and social characteristics of the deprived areas. For instance, the census-derived features incorporated information on coverage of basic provision of the sewage system and home ownership, which are fundamental aspects of deprivation in São Paulo (Pasternak & D'Ottaviano, 2016). These characteristics include a new layer of deprivation and stress aggravated levels of precariousness, especially for Cluster 1, which can be crucial information for decision-makers. Moreover, it also improved the model ability to detect non-built-up areas by increasing the separability of Cluster 0 with the low to zero

population values. Although Model 7 could capture more nuances of deprivation than the model with only EO-based features, it is important to emphasize the remaining challenges and implications of adding census data to ML models. First, Section 4.5 shows that the data aggregation generate ecological fallacies, producing misleading results according to the administrative census boundaries that clearly appear in the gridded outputs (Kohli et al., 2016; Wardrop et al., 2018). Second, as census layers were produced in 2010 and the AGSN layer was updated in 2019, the census does not count several new deprived settlements, thus several cells (3229) containing zero information, which also compromise the results. It might be interesting to repeat the approach after the update of the census data.

Moreover, this research explored strategies to sensibly select a unit of analysis considering not only the integration of EO-derived features with different spatial scales but also the local heterogeneities of deprived settlements. Slum and deprivation studies that provide disaggregated outputs rarely document their grid size choices, but those that primarily rely on the resolution of the datasets and the sensitivity to the scale of analysis. As this study aims for an intra-city analysis, the grid size is carefully decided to retain most of the information of the settlements without creating omission issues. A 20m resolution might not be considered finer-grained in several slum studies (Ajami et al., 2019; Duque et al., 2017), but the present results show sufficient level of detail at the city scale. Thus, as Huang and Zhang (2013) stated, this work also demonstrated that the current granularity of open global RS datasets can provide accurate results and be integrated with other global products, e.g., population maps as WorldPop data. Yet, it can affect the spectral information extracted, especially in pocket-size areas or weirdly shaped ones. In this context, the increasing availability of VHR data could bring great advantages for deprivation mapping studies. Nevertheless, employing a grid-based approach has practical values in this research. The results of Model 4 argue that spatially disaggregated outputs better reflect the local heterogeneity of deprived settlements in São Paulo, notably in medium to large settlements, avoiding stigmatisation of the areas. Besides, it also allows comparison with other relevant maps for an informed decision-making process. For example, in the case of Cluster 1, which is peculiarly located in steeper slope regions, the results can be overlaid with hazard maps and increase the interpretability of where the most exposed and vulnerable population is, ensuring their prioritization in Disaster Risk Management (DRM) plans and programs.

5.3. Capturing intra-urban diversity with an unsupervised learning model and open datasets

The present study is able to successfully capture the diversity of deprived areas for the city of São Paulo, by harvesting large sets of open data and adopting an unsupervised learning method. These two core choices of the proposed approach aim at tackling two fundamental gaps found in the literature (Thomson et al., 2020): *(1) the lack of scalability*. As stated before, the conceptualisation issues are affected by the diverse nature of deprived areas. The ambiguity of definitions hinders the capacity of the current supervised modelling approaches to capture the diversity of deprived areas, because of the large training and testing datasets required (Gevaert et al., 2019). In EO terms, this often means large datasets of expensive VHR data that are computationally demanding, i.e., an inaccessible and unaffordable scenario for most LMICs. Considering this, the unsupervised algorithm does not require training data and the pool of features developed in this research from different open data sources enable scalability of the modelling approach. To extrapolate it for a regional or even national scale would likely require reducing the resolution of the analysis. (2) *the lack of transferability*. Most characterization studies are tailored to a specific context and are not tested for different locations. Due to time constraints, this study could not reproduce the approach for another Brazilian city. Still, all the decisions were taken in order to allow transferability of the model, by using the open datasets. The model is still context-based and requires tailoring of the features accordingly, but the pool of features found in literature is meaningful for São Paulo and the eight dimensions applied has generalization potential. Nevertheless, when applying the model to a different context, it is worth stressing that each city has its own specificities. While characteristics can be applicable to other Latin American cities with similar morphological profiles (Kohli et al., 2013), others might not, especially regarding features under the land

morphology domain. For different cities different numbers of clusters and types of clusters are expected in relation to their local context of deprivation, but still the workflow developed here remain the same. Thus, either way, the relevance of this analysis relies on its transferability and scalability aspects, which can push a step forward in understanding deprivation on larger scales.

In this research, a k-means model was trained to characterize deprived areas, and the selection of the algorithm relates to the second sub-objective. In respect of the reproducibility of the approach and the efficiency of algorithms with large datasets, K-Means and HDBSCAN were tested, and K-Means is chosen over HDBSCAN for the simpler initialization with no input calibration required. The algorithm creates this characterization map rapidly highlighting the intra-urban variations, but results greatly depend on the one input necessary, the 'k' value. Even with the existing data-driven methods, e.g., elbow method, the resulting clusters need to be interpreted, assessing its purity and completeness. Visual inspection indicates that the four clusters provided a balanced result, preventing small clusters from becoming noise.

Besides the manual choice of the 'k' value, other procedures were taken to increase the effectiveness of the k-means, such as running the model 20 times with the same initialization seed, analysing the data descriptively and standardizing the features values. However, an intriguing result of this research concerns the challenge that k-means often faces with high dimensional data. Dimensionality reduction techniques were studied in literature (Section 2.5) and based on the constraints, two tools were applied: Pearson Correlation coefficient and the feature importance tool from R package 'FeatureImpCluster'. The results show that the removed features accounted for important characteristics of the settlements for the proposed model; thus, excluding these dimensions worsen the cluster's separability. In other applications, e.g., on larger national scale where datasets would be considerably large to deal with such multivariate input, these tools can be very beneficial.

The R package tool is also used to generate insights on their significance. GIS- and RS-based features are responsible for capturing different aspects of deprivation. The emerging clusters indicate more precarious conditions in the periphery, related to the statistical significance of the features in services and infrastructure dimensions. Distance to financial facilities and distance to major roads are the most important GIS-based features and they present relevance in the four cluster types. Texture features can capture the differences between the two denser settlement types, evincing the slum-like morphological patterns from Cluster 3, and can indicate land-use variations in Cluster 2. DEM features capture the topography nuances of the landscape, indicating differences not only between central settlements in low lying areas and periphery, but also within the peripheral ones. NDVI and LSM values are likely to be responsible for the great separability between the two peripheral clusters. Besides, the algorithm that developed the WSF layer is sensitive to the presence of built-up areas, constantly overpredicting these areas (Palacios-lopez et al., 2019). In turn, most of the lowest values are assigned to a single cluster, which provides insights into the recurrent presence of open spaces upon these settlements. From all features, the density of water bodies is the least relevant one in this analysis, and the moving window size (1km) might be related to the model's ability to capture the underlying pattern from the data. These findings suggest different levels of importance of a feature for characterising deprived areas within the same city, thus adding to previous work that already indicated different feature' importance across cities (Kohli et al., 2012; Mahabir et al., 2020).

5.4. Pros and cons of the approach

Table 11 lists the pros and cons of the proposed approach. The first advantage concerns the adoption of an unsupervised model to capture the diversity of deprived areas. Especially with the context of COVID-19, there is a rising demand for data-driven approaches that do not require surveying. Acquiring training data is particularly resource-intensive in LMICs and deprivation studies demands large sets of data. Second, more than avoid spatial fallacies, the gridded output allows scalability of the model and easily comparability with other reference data. Third, solely open datasets support the scalability and transferability of the method to different cities across the globe. However, this great advantage has its counterpart. The data cleaning,

preparation and feature extraction processes are quite demanding, especially under the open data perspective. Each dataset was checked for quality issues via careful inspection, determining the most-time consuming process of this study. This was necessary to reduce uncertainties as most as possible and maintain most of the information of the model. Inconsistencies and biases exist, and researchers need to acknowledge and address these constraints, which are regarded as acceptable. The present research hopes for more open access availability of high-quality datasets to promote more consistent results to LMICs. The second con relates to the level of automatization applied to the model. Due to time constraints, the programming languages and platforms were not unified into a single code from data integration, analysis to visualization. The author understands that the full automatization of the modelling steps has utmost importance for practicality and convenience to end-users. Nonetheless, this is a completely manageable issue. The third con refers to the need for a delineation layer for deprived areas and is not a disadvantage per se. From the beginning, this research did not attempt to map settlements, but to offer an alternative approach for their characterization. A possible solution is to develop an unsupervised classification approach to both identify and characterize deprived areas (Thomson et al., 2021). Even today there is research in progress in Lastly, it is important to stress that machine learning models only works because the data used is meaningful to solve the task, which is to characterize the deprived areas. This means that the k-means algorithm provided useful results because there were indeed relationships between the input data and the output.

Table 11. Summary of pros and cons of the proposed approach.

Pros	**Cons**
Unsupervised modelling does not require labelled data for training, that is usually resource-intensive and often unmanageable for LMICs	Data processing steps can be time-consuming
The spatially disaggregated approach allows flexibility for the output, scalability of the model and avoid stigmatisation of the settlements	Little automatization of the approach can hinder the practical applicability of the model
Open datasets allow transferability of the approach fighting the data poverty issues of LMICs	Requires the reference data of deprived settlements

5.5. Limitations

The limitations of this analysis concern the input data and the processing steps. The first input constraint is the AGSN layer. As discussed in Section 3.4.1, the layer has omission issues, excluding settlements with less than 50 households, which does not worsen the achieved results directly, but can induce misconceptions on the level of heterogeneity on the ground. The second relates to the coarse granularity of open datasets. For instance, the spatial resolution of the datasets varies from 10 to 100m, except from the VIIRS, which original resolution is 500m. The resampling process offers significant spatial biases, especially down-sampling. Third, even though the crowdsourcing GIS-based features were carefully checked for major quality issues, inconsistencies still exist, which can influence the reliability of the extracted features to some extent.

Regarding the conducted processing steps, the first limitation concerns the different thresholds applied to the base grid layer to retain most of the homogeneous pixels (purely deprived). Removing the pixels with more than 50% of intersection with the non-deprived areas prevent noisy results but can also indicate information loss. The second constraint relates to the moving window calculation for some handcrafted features. The 5x5 kernel can generate statistical biases (Frazier & Kedron, 2017), because for every extracted centroid, the 25 pixels within the kernel (100x100m) are assigned with the same value, even though remaining with the same resolution. In this sense, the use of 10m and 15m resolution imagery for LSM and GLCM calculation, respectively, is a limitation of this research as already stated in previous studies (Duque et al., 2017; Kemper, Mudau, Mangara, & Pesaresi, 2015; Kohli et al., 2013). Optimally, more VHR imagery will be soon freely available to address this shortcoming.

6. CONCLUSION

6.1. Summary of findings

This study demonstrated the potential of unsupervised learning models for characterizing deprived areas based on morphological and environmental features using solely open data. The approach proposed an alternative method to capture the diversity of deprived areas with disaggregated outputs and used the city of São Paulo as case study. The key findings according to each sub-objective are as follows:

The *first research sub-objective* seeks to derive spatial features for characterizing deprived areas. This study has shown the physical characteristics can be captured from open GIS- and RS-based features. Based on literature, eight dimensions (i.e, texture, geometry, density, spectral, service, infrastructure, greenery, topography and water-related) were used to guide the extraction process of 48 features, from which 32 were used, and reflect the nuances of the settlements within the city. By using open datasets, the research expressed how they can generate relevant information on deprived areas, even with existing limitations, e.g., inconsistencies and low resolution. The increasing availability of open-source data at a global scale and the improvement of the spatial resolution of RS data can significantly contribute to mapping and characterisation studies. In addition, this research also tested, through a data-driven approach, the optimal unit of analysis to integrate these large datasets and provide a disaggregated output. It was shown that the 20x20m resolution grids could best capture spatial similarities and commonalities in a very meaningful way to end-users, which can easily acknowledge the existing complexity within the deprived settlements.

The *second sub-objective* aims at employing the unsupervised learning model and efficiently capturing the intra-urban diversity of deprived areas. The current research stands out to other existing deprivation studies exactly because it employed unsupervised machine-learning. Certainly, considering the contextual and complex nature of deprived settlements, a basic understanding of the study area is necessary to select the appropriate features. However, no training and testing data are required for the model, which removes a highly resource-intensive process, particularly for LMICs. The multiple experiments and validation procedures have shown the promising performance of the selected k-means algorithm with the large datasets and large contribution of the coupled feature importance tool ('FeatureImpCluster' in R). The four emerged clusters shown how the settlements are not homogenous, even within a single city. The results also reveal how the spatial patterns of the clusters are related to the significance of the features to the model. Due to the statistical significance of service- and texture-based features, the most precarious conditions are manifested in Cluster 1, located on the periphery and has medium population density. Comparing the results of Model 4, with only EO data and Model 7, combining census-derived features, reveals a significant contribution of social characteristics to the approach. Notwithstanding the limitations of aggregated information at the household level and its temporal inconsistency, the inclusion of socioeconomic features reveals different perspectives of deprivation that are not easily depicted by the open EO datasets alone. For instance, the estimated level of deprivation of Cluster 1 is shown even higher considering the conditions of sewage system and homeownership.

The *third and last sub-objective* refers to the added value and applicability of the proposed clustering approach for informing decision-makers, which was addressed in the Discussion chapter. This study profiled each emerged cluster type and assessed the ability of the model with the respective spatial features to characterize deprived areas in São Paulo. By using the feature importance tool, this research could illustrate the main relevant differences among the four clusters. Cluster type 0 is marked by zero to low population density values, fragmented urban layout and high density of green areas. These characteristics are important to inform the location of infant settlements that require specific policies to be controlled as they often occupy conservation areas. Cluster type 1 reflects settlements with lack of services and infrastructure provision and the location in steeper slopes. The inclusion of the census-derived features specifically added relevant information to this cluster, evincing higher precariousness, which can help shape better pro-poor policies

and target resources. Cluster type 2 differs from the others with the highest urban fragmentation and variability, indicating the presence of non-residential land uses. This suggests bordering zones that already provided by basic services and infrastructure and require specific urban plans to sew the urban fabric and smoothen the differences between deprived and non-deprived areas. Cluster type 3 is marked by the highest population and built-up densities, complex layout and better accessibility to facilities. Considering the global pandemic, such areas should be urgently targeted as the virus spread is significantly faster in highly concentrated zones as this settlement type. Also, considering the level of service provision in both Clusters 2 and 3, they are more prone to the presence of deprived households within a non-deprived neighbourhood (Thomson et al., 2020), requiring also specific policy-making.

The information above, extracted from the different morphological features and confirmed by the expert validation, showed how the emerged clusters can assess important characteristics of these settlements and qualify their differences, thus proving the unsupervised model's potential for informed decision-making. Besides, the disaggregated model outputs can be applied in different strategic urban planning areas, such as zoning or disaster risk, but also with little adaptation, implemented in other Brazilian cities. By solely relying on open data and a clustering algorithm, the main advantage of the approach is its scalability capacity, pushing the application of the model towards mapping approaches above the city scale, which is essential to the global development goals.

6.2. Recommendation for future work

While the present study proves that unsupervised models and open data hold great potential and applicability to characterize deprived areas, more insights are necessary on the use of open and large datasets and on the heterogeneity of deprived areas within and across cities of the globe. Looking ahead, recommendations for future studies are listed below, considering this research as part of a larger framework that is worth exploring.

- Even though the solely EO-based model here presented does differentiate the morphological and environmental aspects of multiple deprivation in São Paulo, the integration of social characteristics enriches the model. In this sense, different census-derived features should also be experimented with. Studies have shown that schooling level and birth rates are important poverty indicators in Brazil and might be a suitable candidate for additional experiments (Bacelar & da Cunha, 2019).
- Unfortunately, this approach was not an entirely automatized workflow, which hampers the real applicability of the model and underlines room for further improvement. First, it would be interesting to automatize the extraction of the spatial features through geopandas packages instead of relying on GIS built tools, because they require less computational capacity in a high-level interface. Second and utmost necessary, it would be to unify the different code scripts (EDA, clustering model and feature importance tool) instead of using different programming softwares. Developing a fully automated process can not only facilitate transferability and scalability, but also ensure the applicability of this pilot study to municipalities without requiring advanced programming skills by the local experts.
- As an experimental research, the approach was only applied in São Paulo, but as stated previously, the model has promising potential for transferability to other Brazilian cities. It is equally encouraged to scale the model up to regional or national scale, and even to replicate it to other global contexts exploring how to transfer it, e.g., the tailoring process of the different features required to discover deprivation patterns in other cities.

7. LIST OF REFERENCES

Abascal, Á., Rothwell, N., Shonowo, A., Thomson, D., Elias, P., Elsey, H., … Kuffer, M. (2021). "Domains of Deprivation Framework" for Mapping Slums, Informal Settlements, and Other Deprived Areas in LMICs to Improve Urban Planning and Policy: A Scoping Review. *Preprints*, (March), 1–37. https://doi.org/10.20944/preprints202102.0242.v1

Agarwal, S., Jakes, S., Essex, S., Page, S. J., & Mowforth, M. (2018). Disadvantage in English seaside resorts: A typology of deprived neighbourhoods. *Tourism Management*, *69*, 440–459. https://doi.org/10.1016/j.tourman.2018.06.012

Aguilar, R., & Kuffer, M. (2020). Cloud Computation Using High-Resolution Images for Improving the SDG Indicator on Open Spaces. *Remote Sensing*, *12*, 1–17. https://doi.org/10.3390/rs12071144

Ajami, A., Kuffer, M., Persello, C., & Pfeffer, K. (2019). Identifying a slums' degree of deprivation from VHR images using convolutional neural networks. *Remote Sensing*, *11*(11). https://doi.org/10.3390/rs11111282

Alelyani, S., Tang, J., & Liu, H. (2014). Feature Selection for Clustering: A Review. In *Data Clustering: Algorithms and Applications* (pp. 1–37). https://doi.org/10.1201/9781315373515-2

Bacelar, S. G., & da Cunha, J. M. P. (2019). Moradia na Favela : Uma Visão Sociodemográfica dos Aglomerados Subnormais em Campinas. *Anais XVIII ENANPUR*, 20. Retrieved from http://anpur.org.br/xviiienanpur/anais

Bastos da Cunha, M., Firpo de Souza Porto, M., Pivetta, F., Zancan, L., Santos Francisco, M., Brum Pinheiro, A., … Calazans, R. (2015). O desastre no cotidiano da favela: reflexões a partir de três casos no Rio de Janeiro [An everyday disaster in favela: reflections based on three cases in Rio de Janeiro]. *O Social Em Questão*, *33*, 95–122.

Batty, M. (2019). Urban analytics defined. *Urban Analytics and City Science*, *46*, 403–405. https://doi.org/10.1177/2399808319839494

Baud, I., Kuffer, M., Pfeffer, K., Sliuzas, R., & Karuppannan, S. (2010). Understanding heterogeneity in metropolitan india: The added value of remote sensing data for analyzing sub-standard residential areas. *International Journal of Applied Earth Observation and Geoinformation*, *12*(5), 359–374. https://doi.org/10.1016/j.jag.2010.04.008

Brasil. (2010). *Guia para o Mapeamento e Caracterização de Assentamentos Precários*. Brasilia - DF.

Cai, D., Zhang, C., & He, X. (2010). Unsupervised feature selection for Multi-Cluster data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 333–342. https://doi.org/10.1145/1835804.1835848

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Pei J., Tseng V.S., Cao L., Motoda H., & Xu G. (Eds.), *Advances in Knowledge Discovery and Data Mining*. (Vol. 7819, pp. 160–172). https://doi.org/10.1007/978-3-642-37456-2_14

Carr-Hill, R. (2013). Missing Millions and Measuring Development Progress. *World Development*, *46*, 30–44. https://doi.org/10.1016/j.worlddev.2012.12.017

Chakraborty, A., Wilson, B., Sarraf, S., & Jana, A. (2015). Open data for informal settlements: Toward a user's guide for urban managers and planners. *Journal of Urban Management*, *4*(2), 74–91. https://doi.org/10.1016/j.jum.2015.12.001

Chiang, M. M. T., & Mirkin, B. (2010). Intelligent choice of the number of clusters in k-means clustering: An experimental study with different cluster spreads. *Journal of Classification*, *27*(1), 3–40. https://doi.org/10.1007/s00357-010-9049-5

D'Alençon, P. A., Smith, H., Álvarez de Andrés, E., Cabrera, C., Fokdal, J., Lombard, M., … Spire, A. (2018). Interrogating informality: Conceptualisations, practices and policies in the light of the New Urban Agenda. *Habitat International*, *75*(April), 59–66. https://doi.org/10.1016/j.habitatint.2018.04.007

Denaldi, Rosana Petrarolli, Juliana Gomes Gonçalves, Gilmara da Silva de Morais, G. M. (2018). Tecidos Urbanos E A Identificação De Assentamentos Precários Na Região Metropolitana Da Baixada Santista. *URB Favelas*, 21. Retrieved from http://www.sisgeenco.com.br/sistema/urbfavelas/anais2018a/ARQUIVOS/GT2-917-9-20190102174927.pdf

Duque, J. C., Patino, J. E., & Betancourt, A. (2017). Exploring the potential of machine learning for automatic slum identification from VHR imagery. *Remote Sensing*, *9*(9), 1–23.

https://doi.org/10.3390/rs9090895
Ebert, A., Kerle, Æ. N., & Stein, Æ. A. (2009). *Proxies and spatial metrics derived from air- and spaceborne imagery and GIS data*. 275–294. https://doi.org/10.1007/s11069-008-9264-0
Engstrom, R., Ofiesh, C., Rain, D., Jewell, H., & Weeks, J. (2013). Defining neighborhood boundaries for urban health research in developing countries: A case study of Accra, Ghana. *Journal of Maps*, *9*(1), 36–42. https://doi.org/10.1080/17445647.2013.765366
Ferreira, N. D. J., & Feitosa, F. da F. (2020). Cartografias das Favelas : Uma Análise Comparativa. *I Seminário Nacional – Urbanismo, Espaço e Tempo. Temática: Cidade, Pandemia e Cotidiano.*, 5. Belo Horizonte: Revista Políticas Públicas e Cidades.
Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81. Retrieved from http://jmlr.org/papers/v20/18-760.html
Fränti, P., & Sieranoja, S. (2019). How much can k-means be improved by using better initialization and repeats? *Pattern Recognition*, *93*, 95–112. https://doi.org/10.1016/j.patcog.2019.04.014
Frazier, A. E., & Kedron, P. (2017). Landscape Metrics : Past Progress and Future Directions. *Current Landscape Ecology Reports*, 63–72. https://doi.org/10.1007/s40823-017-0026-0
Friesen, J., Taubenböck, H., Wurm, M., & Pelz, P. F. (2019). Size distributions of slums across the globe using different data and classification methods. *European Journal of Remote Sensing*, *52*(sup2), 99–111. https://doi.org/10.1080/22797254.2019.1579617
Gevaert, C. M., Kohli, D., & Kuffer, M. (2019). Challenges of mapping the missing spaces. *2019 Joint Urban Remote Sensing Event, JURSE 2019*, (May), 2019–2022. https://doi.org/10.1109/JURSE.2019.8809004
Ghaffarian, S., Kerle, N., & Filatova, T. (2018). *Remote Sensing-Based Proxies for Urban Disaster Risk Management and Resilience : A Review*. https://doi.org/10.3390/rs10111760
Goldblatt, R., Stuhlmacher, M. F., Tellman, B., Clinton, N., Hanson, G., Georgescu, M., … Balling, R. C. (2018). Using Landsat and nighttime lights for supervised pixel-based image classification of urban land cover. *Remote Sensing of Environment*, *205*(December 2017), 253–275. https://doi.org/10.1016/j.rse.2017.11.026
Google Maps. (n.d.). Visual assessment of a central area of São Paulo using reference satellite and street-view images. Retrieved July 11, 2021, from https://www.google.nl/maps/place/São+Paulo,+Brazilië/@-23.6821604,-46.8754915,105442m/data=!3m2!1e3!4b1!4m5!3m4!1s0x94ce448183a461d1:0x9ba94b08ff335bae!8m2!3d-23.5557714!4d-46.6395571
Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031
Gram-Hansen, B. J., Azam, F., Helber, P., Coca-Castro, A., Bilinski, P., Varatharajan, I., & Kopackova, V. (2019). Mapping informal settlements in developing countries using machine learning and low resolution multi-spectral data. *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 361–368. https://doi.org/10.1145/3306618.3314253
Grippa, T., Georganos, S., Zarougui, S., Bognounou, P., Diboulo, E., Forget, Y., … Wolff, E. (2018). Mapping Urban Land Use at Street Block Level Using OpenStreetMap , Remote Sensing Data , and Spatial Metrics. *ISPRS International Journal of Geo-Information*. https://doi.org/10.3390/ijgi7070246
Hall-Beyer, M. (2017a). *GLCM Texture: A Tutorial v. 3.0* (3rd ed.). https://doi.org/10.13140/RG.2.2.12424.21767
Hall-Beyer, M. (2017b). Practical guidelines for choosing GLCM textures to use in landscape classification tasks over a range of moderate spatial scales. *International Journal of Remote Sensing*, *38*(5), 1312–1338. https://doi.org/10.1080/01431161.2016.1278314
Han, J., Kamber, M., & Pei, J. (2012). Cluster Analysis: Basic Concepts and Methods. In *Data Mining* (3rd ed., pp. 443–495). https://doi.org/10.1016/B978-0-12-381479-1.00010-1
Huang, X., & Zhang, L. (2013). *An SVM Ensemble Approach Combining Spectral , Structural , and Semantic Features for the Classification of High-Resolution Remotely Sensed Imagery*. *51*(1), 257–272.
Hyvärinen, A. (2015). A unified probabilistic model for independent and principal component analysis. In *Advances in Independent Component Analysis and Learning Machines*. https://doi.org/10.1016/B978-0-12-802806-3.00003-8
Ibes, D. C. (2015). A multi-dimensional classification and equity analysis of an urban park system: A novel methodology and case study application. *Landscape and Urban Planning*, *137*, 122–137.

https://doi.org/10.1016/j.landurbplan.2014.12.014
IBGE. (2019a). Aglomerados Subnormais | IBGE. Retrieved June 3, 2021, from https://www.ibge.gov.br/geociencias/organizacao-do-territorio/tipologias-do-territorio/15788-aglomerados-subnormais.html?=&t=o-que-e
IBGE. (2019b, August 23). SP Capital. Retrieved June 4, 2021, from Censo Demográfico 2010 website: https://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/Resultados_do_Universo/Agregados_por_Setores_Censitarios/
IBGE. (2020). *Estimativas da população residente para os municípios e para as unidades da federação brasileiros com data de referência em 1º de julho de 2020*. Retrieved from https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=2101747
Jain, A. K. (2008). Data Clustering : 50 Years Beyond K-means. *ECML PKDD*, 3–4.
Jasiewicz, J., & Stepinski, T. F. (2013). Geomorphons-a pattern recognition approach to classification and mapping of landforms. *Geomorphology*, *182*, 147–156. https://doi.org/10.1016/j.geomorph.2012.11.005
Jia, Y., Tang, L., Xu, M., & Yang, X. (2019). Landscape pattern indices for evaluating urban spatial morphology – A case study of Chinese cities. *Ecological Indicators*, *99*, 27–37. https://doi.org/10.1016/j.ecolind.2018.12.007
Jochem, W. C., Leasure, D. R., Pannell, O., Chamberlain, H. R., Jones, P., & Tatem, A. J. (2020). *Classifying settlement types from multi-scale spatial patterns of building footprints*. *0*(0), 1–19. https://doi.org/10.1177/2399808320921208
Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis : a review and recent developments. *The Royal Society*. http://dx.doi.org/10.1098/rsta.2015.0202
Karmitsa, N., Bagirov, A. M., & Taheri, S. (2018). Clustering in large data sets with the limited memory bundle method. *Pattern Recognition*, *83*, 245–259. https://doi.org/10.1016/j.patcog.2018.05.028
Kemper, T., Mudau, N., Mangara, P., & Pesaresi, M. (2015). Towards an automated monitoring of human settlements in South Africa using high resolution SPOT satellite imagery. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *40*(7W3), 1389–1394. https://doi.org/10.5194/isprsarchives-XL-7-W3-1389-2015
Kit, O., Lüdeke, M., & Reckien, D. (2012). Texture-based identification of urban slums in Hyderabad, India using remote sensing data. *Applied Geography*, *32*(2), 660–667. https://doi.org/10.1016/j.apgeog.2011.07.016
Kohli, D., Sliuzas, R., Kerle, N., & Stein, A. (2012). Computers , Environment and Urban Systems An ontology of slums for image-based classification. *Computers, Environment and Urban Systems*, *36*(2), 154–163. https://doi.org/10.1016/j.compenvurbsys.2011.11.001
Kohli, D., Stein, A., & Sliuzas, R. (2016). Uncertainty analysis for image interpretations of urban slums. *Computers, Environment and Urban Systems*, *60*, 37–49. https://doi.org/10.1016/j.compenvurbsys.2016.07.010
Kohli, D., Warwadekar, P., Kerle, N., Sliuzas, R., & Stein, A. (2013). Transferability of object-oriented image analysis methods for slum identification. *Remote Sensing*, *5*(9), 4209–4228. https://doi.org/10.3390/rs5094209
Kraff, N. J., Wurm, M., & Taubenbock, H. (2020). Uncertainties of human perception in visual image interpretation in complex urban environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 4229–4241. https://doi.org/10.1109/JSTARS.2020.3011543
Kuffer, M., Orina, F., & Sliuzas, R. (2017). *Spatial Patterns of Slums : Comparing African and Asian Cities*. 5–8.
Kuffer, M., Pfeffer, K., & Sliuzas, R. (2016). Slums from space-15 years of slum mapping using remote sensing. *Remote Sensing*, *8*(6). https://doi.org/10.3390/rs8060455
Kuffer, M., Pfeffer, K., Sliuzas, R., & Baud, I. (2016). Extraction of Slum Areas From VHR Imagery Using GLCM Variance. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(5), 1830–1840. https://doi.org/10.1109/JSTARS.2016.2538563
Kuffer, M., Pfeffer, K., Sliuzas, R., & Baud, I. (2017). Capturing the diversity of deprived areas with image-based features: The case of Mumbai. *Remote Sensing*, *9*(4). https://doi.org/10.3390/rs9040384
Kuffer, M., Thomson, D. R., Boo, G., Mahabir, R., Grippa, T., Vanhuysse, S., … Kabaria, C. (2020). The role of earth observation in an integrated deprived area mapping "system" for low-to-middle income countries. *Remote Sensing*, *12*(6). https://doi.org/10.3390/rs12060982
Leonita, G., Kuffer, M., Sliuzas, R., & Persello, C. (2018). Machine learning-based slum mapping in support of slum upgrading programs: The case of Bandung City, Indonesia. *Remote Sensing*, *10*(10). https://doi.org/10.3390/rs10101522

Leyk, S., Gaughan, A. E., Adamo, S. B., De Sherbinin, A., Balk, D., Freire, S., … Pesaresi, M. (2019). The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. *Earth System Science Data*, *11*(3), 1385–1409. https://doi.org/10.5194/essd-11-1385-2019

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys*, *50*(6). https://doi.org/10.1145/3136625

Lilford, R., Kyobutungi, C., Ndugwa, R., Sartori, J., Watson, S. I., Sliuzas, R., … Ezeh, A. (2019). Because space matters: Conceptual framework to help distinguish slum from non-slum urban areas. *BMJ Global Health*, *4*(2). https://doi.org/10.1136/bmjgh-2018-001267

Lucci, P., Bhatkal, T., & Khan, A. (2016). *Are we underestimating urban poverty?* London.

Mahabir, R., Agouris, P., Stefanidis, A., Croitoru, A., & Crooks, A. T. (2020). Detecting and mapping slums using open data: a case study in Kenya. *International Journal of Digital Earth*, *13*(6), 683–707. https://doi.org/10.1080/17538947.2018.1554010

Maricato, E. (2003). Metrópole, legislação e desigualdade. *Estudos Avançados*, *17*(48), 151–167.

Martínez, J., Pfeffer, K., & Baud, I. (2016). Factors shaping cartographic representations of inequalities. Maps as products and processes. *Habitat International*, *51*, 90–102. https://doi.org/10.1016/j.habitatint.2015.10.010

Martino, A., Rizzi, A., & Mascioli, F. M. F. (2017). Efficient approaches for solving the large-scale k-medoids problem. *IJCCI 2017 - Proceedings of the 9th International Joint Conference on Computational Intelligence*, (Ijcci), 338–347. https://doi.org/10.5220/0006515003380347

McInnes, L., Healy, J., & Astels, S. (2016). *The HDBSCAN Clustering Library*. Retrieved from https://www.ibm.com/docs/en/wsd?topic=modeling-hdbscan-node

Molenaar, M. (2000). Three conceptual uncertainty levels for spatial objects. *ISPRS IC WG IV/III.1 - GIS Fundamentals and Spatial Databases*, 670–677. Retrieved from http://www.isprs.org/proceedings/XXXIII/congress/part4/670_XXXIII-part4.pdf

Mudau, N., & Mhangara, P. (2021). *Investigation of Informal Settlement Indicators in a Densely Populated Area Using Very High Spatial Resolution Satellite Imagery*.

Nobrega, R. A. A., O'Hara, C. G., & Quintanilha, J. A. (2008). An object-based approach to detect road features for informal settlements near Sao Paulo, Brazil. *Lecture Notes in Geoinformation and Cartography*, *0*(9783540770572), 589–607. https://doi.org/10.1007/978-3-540-77058-9_32

Nugrahita, R., & Surjandari, I. (2020). Identify product families using cluster analysis : case study in Passenger Car Radial (PCR) tire product. *IOP Conference Series: Materials Science and Engineering PAPER*. https://doi.org/10.1088/1757-899X/909/1/012057

Olthuis, K., Benni, J., Eichwede, K., & Zevenbergen, C. (2015). Slum Upgrading: Assesing the importance of location and a plea for a spatial approach. *Habitat International*, *50*, 270–288. https://doi.org/10.1016/j.habitatint.2015.08.033

Owen, K. K., & Wong, D. W. (2013). An approach to differentiate informal settlements using spectral , texture , geomorphology and road accessibility metrics. *Applied Geography*, *38*, 107–118. https://doi.org/10.1016/j.apgeog.2012.11.016

Ozdemir, I., Mert, A., & Senturk, O. (2012). Predicting landscape structural metrics using aster satellite data. *Journal of Environmental Engineering and Landscape Management*, *20*(2), 168–176. https://doi.org/10.3846/16486897.2012.688371

Palacios-lopez, D., Bachofer, F., Esch, T., Heldens, W., Hirner, A., Marconcini, M., … Reinartz, P. (2019). *New Perspectives for Mapping Global Population Distribution Using World Settlement Footprint Products*.

Pasternak, S. (2006, June). São Paulo e suas favelas. *Pós Revista*, 176–197. https://doi.org/10.11606/issn.2317-2762.v0i19p176-197

Pasternak, S., & D'Ottaviano, C. (2016). Favelas no Brasil e em São Paulo: avanços nas análises a partir da Leitura Territorial do Censo de 2010*. *Cadernos Metrópole*, *18*(35), 75–100. https://doi.org/10.1590/2236-9996.2016-3504

Patel, A., Koizumi, N., & Crooks, A. (2014). Measuring slum severity in Mumbai and Kolkata: A household-based approach. *Habitat International*, *41*, 300–306. https://doi.org/10.1016/J.HABITATINT.2013.09.002

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). 2.3. Clustering — scikit-learn 0.24.2 documentation. Retrieved June 11, 2021, from Scikit-learn: Machine Learning in Python website: https://scikit-learn.org/stable/modules/clustering.html#k-means

Pedro, A. A., & Queiroz, A. P. (2019). Slum: Comparing municipal and census basemaps. *Habitat*

International, *83*(October 2018), 30–40. https://doi.org/10.1016/j.habitatint.2018.11.001
Pfaffel, O. (2021). *Feature Importance for Partitional Clustering*. CRAN.
Reis, J. P., Silva, E. A., & Pinho, P. (2016). Spatial metrics to study urban patterns in growing and shrinking cities. *Urban Geography*, *37*(2), 246–271. https://doi.org/10.1080/02723638.2015.1096118
Richardson, M. (2009). *Principal Component Analysis*. Retrieved from http://www.dsc.ufcg.edu.br/~hmg/disciplinas/posgraduacao/rn-copin-2014.3/material/SignalProcPCA.pdf
Roy, D., Bernal, D., & Lees, M. (2020). An exploratory factor analysis model for slum severity index in Mexico City. *Urban Studies*, *57*(4), 789–805. https://doi.org/10.1177/0042098019869769
Roy, D., Lees, M. H., Palavalli, B., Pfeffer, K., & Sloot, M. A. P. (2014). The emergence of slums: A contemporary view on simulation models. *Environmental Modelling and Software*, *59*, 76–90. https://doi.org/10.1016/j.envsoft.2014.05.004
Sankey, T. T. (2016). Scale, Effects. In *Encyclopedia of GIS* (pp. 1–8). https://doi.org/10.1007/978-3-319-23519-6_1161-2
São Paulo. (2020). Sistema de Consulta do Mapa Digital da Cidade de São Paulo. Retrieved June 6, 2021, from GeoSampa Mapa website: http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx
São Paulo, S. (2019). *Metodologia para Identificação e Caracterização de Assentamentos Precários em Regiões Metropolitanas Paulistas*. São Bernardo do Campo, SP.
Schindler, S. (2017). Towards a paradigm of Southern urbanism. *Https://Doi.Org/10.1080/13604813.2016.1263494*, *21*(1), 47–64. https://doi.org/10.1080/13604813.2016.1263494
Scrucca, L. (2016). Identifying connected components in Gaussian finite mixture models for clustering. *Computational Statistics and Data Analysis*, *93*, 5–17. https://doi.org/10.1016/j.csda.2015.01.006
Sliuzas, R., Kuffer, M., & Masser, I. (2010). The spatial and temporal nature of urban objects. In T. Rashed & C. Jürgens (Eds.), *Remote Sensing of Urban and Suburban Areas* (Vol. 10, pp. 67–84). https://doi.org/10.1007/978-1-4020-4385-7_5
Sliuzas, R., Mboup, G., & de Sherbinin, A. (2008). *Report of the expert group meeting on slum identification and mapping*. Retrieved from https://www.researchgate.net/publication/271074739_Report_of_the_Expert_Group_Meeting_on_Slum_Identification_and_Mapping
Small, C. S. (2014). Mapping Urban Growth and Development As Continuous Fields in Space and Time. *Geography Department University of Sao Paulo*, *1*, 155–179. https://doi.org/10.11606/rdg.v0i0.555
Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2019). A review of unsupervised feature selection methods. *Artificial Intelligence Review*, *53*(2), 907–948. https://doi.org/10.1007/s10462-019-09682-y
Soni Madhulatha, T. (2012). AN OVERVIEW ON CLUSTERING METHODS. *IOSR Journal of Engineering*, *2*(4), 719–725.
STARS, P. (n.d.). Multispectral and panchromatic images. Retrieved June 9, 2021, from Knowledge Portal - Magazine website: https://www.stars-project.org/en/knowledgeportal/magazine/remote-sensing-technology/introduction/multispectral-and-panchromatic-images/
Taubenböck, H., & Kraff, N. J. (2014). The physical face of slums : a structural comparison of slums in Mumbai , India , based on remotely sensed data. *Journal of Housing and the Built Environment*, *29*(1), 15–38. https://doi.org/10.1007/s 10901-013-9333-x
Taubenböck, H., Kraff, N. J., & Wurm, M. (2018). The morphology of the Arrival City - A global categorization based on literature surveys and remotely sensed data. *Applied Geography*, *92*(September 2017), 150–167. https://doi.org/10.1016/j.apgeog.2018.02.002
Taubenböck, H., Wurm, M., Geiß, C., Dech, S., & Siedentop, S. (2019). Urbanization between compactness and dispersion: designing a spatial model for measuring 2D binary settlement landscape configurations. *International Journal of Digital Earth*, *12*(6), 679–698. https://doi.org/10.1080/17538947.2018.1474957
The World Wide Web Foundation. (2016). *ODB Global Report Third Edition*. Retrieved from http://opendatabarometer.org/doc/3rdEdition/ODB-3rdEdition-GlobalReport.pdf
Thomson, D. R., Kuffer, M., Boo, G., Hati, B., Grippa, T., Elsey, H., … Sliuzas, R. (2020). Need for an Integrated Deprived Area "Slum" Mapping System (IDEAMAPS) in Low- and Middle-Income Countries (LMICs). *Social Sciences*. https://doi.org/10.3390/socsci9050080
Thomson, D. R., Merodio, P., Kuffer, M., Ramirez, A., Juearez, J., & Jacquin, C. (2021). *Toolkit: Operationalising the IDEAMAPS network*. Retrieved from https://ideamapsnetwork.org/wp-

content/uploads/2021/04/IDEAMAPS_in_Government_2021.pdf
Thrun, M. C., Ultsch, A., & Breuer, L. (2020). *Explainable AI Framework for Multivariate Hydrochemical Time Series*. (November), 1–29. https://doi.org/10.20944/preprints202011.0451.v1
UN-Habitat. (2003). *The Challenge of Slums - Global Report on Human Settlements*. Retrieved from https://unhabitat.org/the-challenge-of-slums-global-report-on-human-settlements-2003
UN-Habitat. (2015a). *SLUM ALMANAC | Tracking Improvement in the Lives of Slum Dwellers*. Nairobi.
UN-Habitat. (2015b). *SUSTAINABLE DEVELOPMENT GOAL 11: Make Cities and Human Settlements Inclusive, Safe, Resilient And Sustainable*. New York, NY, USA.
United Nations. (2018). *The Sustainable Development Goals Report 2018*. New York.
Van Dijk, M., Moorthy, I., Nguyen, B., See, L., & Fritz, S. (2019). Tracking poverty using satellite imagery and big data. *The International Institute for Applied Systems Analysis*, (December), 1–16. Retrieved from http://pure.iiasa.ac.at/id/eprint/16240/
Wang, J., Kuffer, M., & Pfeffer, K. (2019). The role of spatial heterogeneity in detecting urban slums. *Computers, Environment and Urban Systems*, *73*(April 2018), 95–107. https://doi.org/10.1016/j.compenvurbsys.2018.08.007
Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., … Tatem, A. J. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(14), 3529–3537. https://doi.org/10.1073/pnas.1715305115
Wurm, M., & Taubenböck, H. (2018). Detecting social groups from space – assessment of remote sensing-based mapped morphological slums using income data. *Remote Sensing Letters*, *9*(1), 41–50. https://doi.org/10.1080/2150704X.2017.1384586
Wurm, M., Taubenböck, H., Weigand, M., & Schmitt, A. (2017). Slum mapping in polarimetric SAR data using spatial features. *Remote Sensing of Environment*, *194*, 190–204. https://doi.org/10.1016/j.rse.2017.03.030

8. APPENDIX

ANNEX 1 – 2019 AGSN layer definition.

The AGSN ("Substandard Settlements") layer comprises an irregular occupation of public and private land with housing purposes, characterized by organic layouy, lack of basic services and restricted ownership regulations. The layer was created before the Census of 2020 was conducted – that never happened to the global COVID-19 pandemic – as an attempt to help the fight the health crises in Brazil that is commonly more severe in precarious areas due to the population densities and lower sanitation structures (IBGE, 2019a). The layer was built based on the mapping process of the 2010 census. Two methods were used in combination to update the layer, field survey and open satellite and street-view images. The layer provides information on the shape and extent but also estimation on the occupied dwelling.

ANNEX 2 – Typologies built in the MAPPA project for the Metropolitan Region of São Paulo.

Table created by author, adapted the variables from Brasil (2010) and the characteristics from Denaldi, Petrarolli, Gonçalves e de Morais (2018).

Typologies	**Relevant variables**			
	Occupation	**Locational trend**	**Urban pattern**	**Infrastructure**
Hillslopes	Various stages, various density	Steep terrain	Irregular blocks (border vs. internal)	Poor quality roads
Wetlands	High density and precariousness	Along waterways and mangroves	Irregular layout	Lack of basic infrastructure
Irregular occupation	High density	Hazard prone	Irregular layout, occupy entire plot	Poor quality roads
Regular occupation	High density	Low slope	Small plots with regular layout	Basic infrastructure
Sparse occupation	Lower degree of consolidation	Conservation areas or along roads or trails	Poor layout definition	Lack of basic infrastructure

ANNEX 3 – List of EO-based features found in literature.

Dimension	**Feature**	**Rationale**	**References**
Texture	GLCM-texture measures	Asymmetry of size and shape of rooftops can distinguish deprived areas	(Duque et al., 2017; Mahabir et al., 2020; Owen & Wong, 2013)
Density	Population count	Building layout can vary depending on the stage of development of the deprived area and can be a proxy of adequate living space	(Kohli et al., 2016; Sliuzas, Mboup, & de Sherbinin, 2008; Van Dijk, Moorthy, Nguyen, See, & Fritz, 2019)
	Built up density		
Geometry	Built up shape index	Shape and layout of buildings can indicate morphological patterns of deprived areas	(Grippa et al., 2018; Leonita et al., 2018; Hannes Taubenböck, Wurm, Geiß, Dech, & Siedentop, 2019)
	Built up mean area		

Spectral response	Raw spectral bands 2-7 and 11-12	Raw image bands may be sensitive to morphological variations	(Duque et al., 2017; Mahabir et al., 2020)
Infrastructure	Distance to major roads	Deprived areas have lower accessibility	(Kuffer, Pfeffer, & Sliuzas, 2016; Lilford et al., 2019)
	Unpaved road density	Proximity to low-quality roads can indicate low economic status	(Kohli et al., 2012; Nobrega, O'Hara, & Quintanilha, 2008)
	VIIRS night-time light	Energy access can measure poverty rates	(Ghaffarian et al., 2018; Van Dijk et al., 2019)
	Dead end nodes density	Slums often have low internal road access	(Wurm & Taubenböck, 2018)
Services	Distance to schools	Dwellers of deprived areas often have lower access to service provision and job opportunities	(Kohli et al., 2013; Kuffer et al., 2020; Lilford et al., 2019; Mahabir et al., 2020)
	Distance to health facilities		
	Distance to finance facilities		
Greenery	NDVI	Vegetation can indicate different stage of development and location of slums	(Kuffer, Pfeffer, Sliuzas, et al., 2016; Owen & Wong, 2013)
	Proportion of green areas	Sparse deprived settlements are often close to green amenities.	(Ebert, Kerle, & Stein, 2009; Ghaffarian et al., 2018)
Topography	Slope	High exposure to hazard is common to deprived areas; they are often located in low-lying or steep slope regions	(Ebert et al., 2009; Jasiewicz & Stepinski, 2013; Olthuis et al., 2015)
	DEM		
Water-related	Distance to nearest waterbody	Deprived areas are often concentrated near water banks and/or more exposed to flood risk	(Kohli et al., 2016; Kuffer, Pfeffer, & Sliuzas, 2016; Mahabir et al., 2020)
	River network density		

ANNEX 4 – Data Management Plan (DMP) summary.

Source layers, model input, scripts and data products are stored and available at GitHub platform (https://github.com/ltrentooliveira/MSc_Archive), ensuring replicability of the analysis.

Data	Specification	Year
Landsat-8 imagery (BOA reflectance) [9]	Band: Panchromatic / Resolution: 15m Cloud coverage:0.4%	2013
Sentinel-2A imagery (BOA reflectance) [10]	Bands: 2, 3,4,5,6,7,11,12 / Resolution: 10m Cloud coverage:1.2:%	2014
WorldPop Constrained Population Count layer[11]	Resolution: 3 arc-second UN adjusted /Projection-based estimates	2020
WorldPop resampled VIIRS night-time lights layer	Resolution: 3 arc-second Resampled from 15-arc second	2016

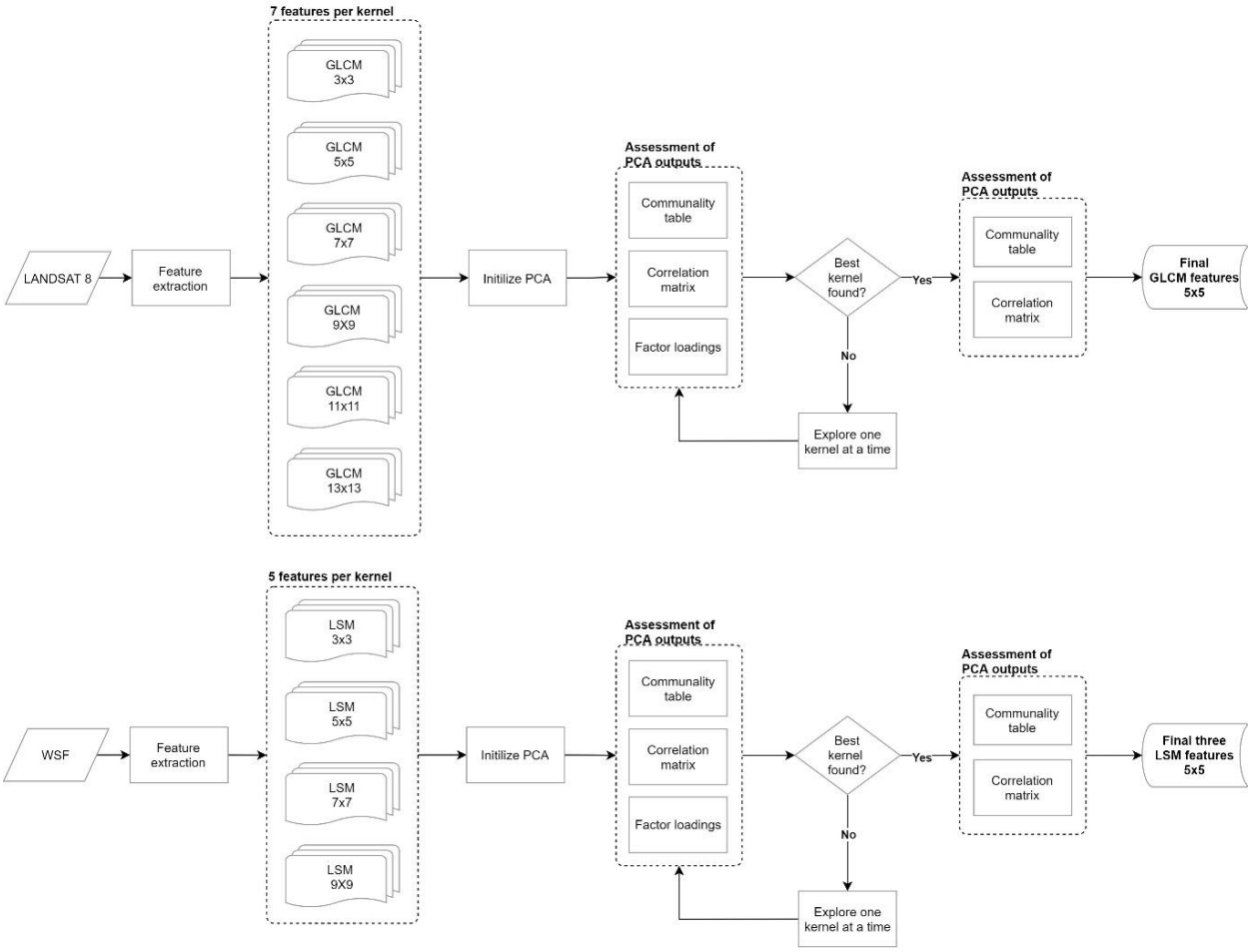
[9] Available at https://earthexplorer.usgs.gov/
[10] Available at https://scihub.copernicus.eu/dhus/#/home
[11] Available at https://www.worldpop.org/geodata/summary?id=18539

World Settlement Footprint (WSF) layer[12]	Resolution: 10m Derived from Sentinel-1 radar and Landsat-8	2015
Humanitarian OpenStreetMap Team (HOTOSM) layers[13]	Shapefile (point)	2020
OpenStreetMap (OSM) layers[14]	Shapefile (point and line)	2020
ASTER Global Digital Elevation Model (GDEM) [15]	Resolution: 30m stereo-pair images	2019
World Database on Protected Areas (WDPA) layer[16]	Shapefile (polygon)	2020
Brazilian census – AGSN layer[17]	Shapefile	2019
Brazilian census data[18]	Shapefile and tabular formats	2010
Land use layer[19]	Shapefile	2016

Code	**GitHub access link**
GLCM calculation	https://github.com/ltrentooliveira/MSc_Archive/blob/main/GLCM.R
EDA	https://github.com/ltrentooliveira/MSc_Archive/blob/main/EDA.R
PCA	https://github.com/ltrentooliveira/MSc_Archive/blob/main/PCA_processing.ipynb
Feature importance	https://github.com/ltrentooliveira/MSc_Archive/blob/main/FeatureImp_tool.R
K-means model	https://github.com/ltrentooliveira/MSc_Archive/blob/main/kmeans_model.ipynb

[12] Available at https://springernature.figshare.com/articles/dataset/World_Settlement_Footprint_WSF_2015
[13] Available at https://data.humdata.org/organization/hot
[14] Available at https://www.openstreetmap.org/
[15] Available at https://earthexplorer.usgs.gov/
[16] Available at https://www.protectedplanet.net/country/BRA
[17] Available at https://www.ibge.gov.br/geociencias/organizacao-do-territorio/tipologias-do-territorio
[18] Available at https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html
[19] Available at http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx#

ANNEX 5 – Workflow of PCA iterations to choose the best moving window for hand-crafted features.



ANNEX 6 – SQL expressions to select appropriate attributes from GIS layers.

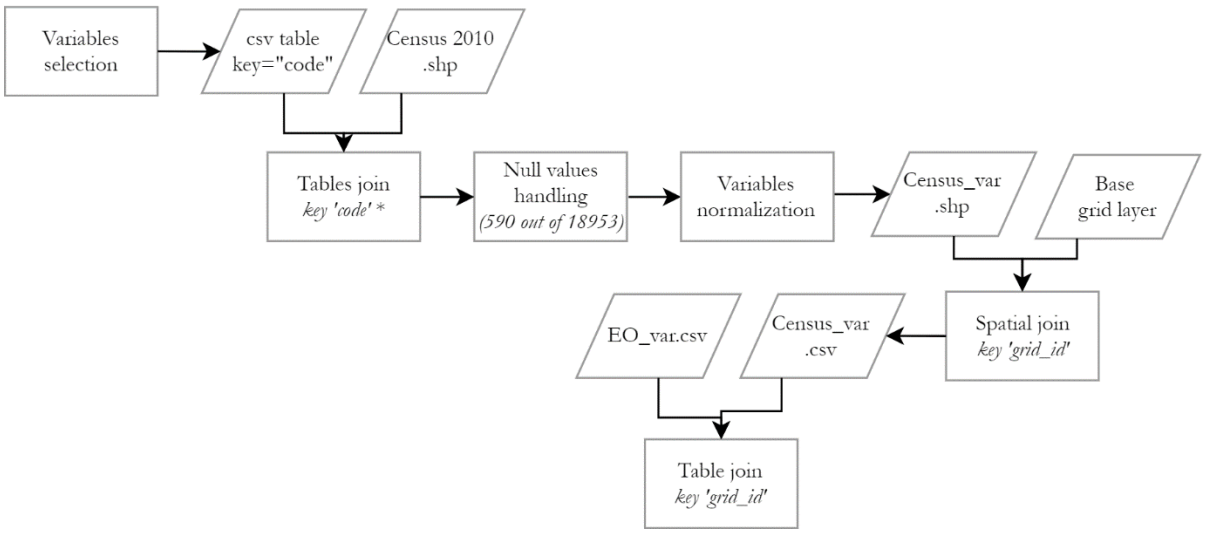
Feature name	Data sources	SQL expression
Distance from major roads	OSM	"type" = 'primary' OR "type" = 'primary_link' OR "type" = 'secondary' OR "type" = 'secondary_link' OR "type" = 'service' OR "type" = 'services' OR "type" = 'tertiary' OR "type" = 'tertiary_link' OR "type" = 'trunk' OR "type" = 'trunk_link'
Density of poor quality roads	OSM	"surface" = 'paving_stones' OR "surface" = 'unpaved' OR "surface" = 'dirt' OR "surface" = 'gravel' OR "surface" = 'mud' OR "surface" = 'pebblestone' OR "surface" = 'pedras_para_pavimentação' OR "surface" = 'sand' OR "surface" = 'ground' OR "surface" = 'earth' OR "surface" = 'compacted' OR "surface" = 'fine_gravel'
Distance to education facilities	HOTOSM	"amenity" = 'school' OR "amenity" = 'university' OR "type" = 'amenity' OR "amenity" = 'kindergarten'
Distance to health facilities	HOTOSM	"amenity" = 'clinic' OR "amenity" = 'clinica' OR "type" = 'hospital'

ANNEX 7 – List of census-derived variables.

From (IBGE, 2019b), features 01to 04 are derived from census variables from table 'Básico_SP'; feature 05 from table 'DomicílioRenda_SP' and feature 06 to 12 from table 'Domicilio01_UF'. The variables mentioned in the processing column are used to normalize some features and/or construct others.

Initial	Feature name	Description	Processing
F01	POP_DENS	Population density	V002/total tract area
F02	AV_HHPOP	Average number of residents per households	V003 (no processing)
F03	BU_DENS	Dwelling/building density	V001/total tract area
F04:	AV_INC	Average monthly income of head of household per dwelling unit (V005)	V005 (no processing)
F05	INC_MWAGE	Proportion of households with monthly income per capita of up to 1 minimum wage	(V005 + V006 + V007 + V008)/V001
F06	HH_APT	Proportion of apartment-type dwellings	V005/V002
F07	HH_PRIV	Proportion of private owned households	(V006 + V007)/V002
F08	HH_GARB	Proportion of dwelling units with adequate waste disposal	V035/V002
F09	HH_WATER	Proportion of dwelling units with adequate water provision	V012/V002
F10	HH_SEWAGE:	Proportion of dwelling units with adequate sewage network	V017/V002
F11	HH_PITLATRINE	Proportion of dwelling units with sewage discharged in pit latrine	V019/V002
F12	HH_NOSEWAGE:	Proportion of dwelling units with sewage discharged in septic in river or lake	V021/V002

ANNEX 8 – Flowchart of pre-processing steps for census-derived model.

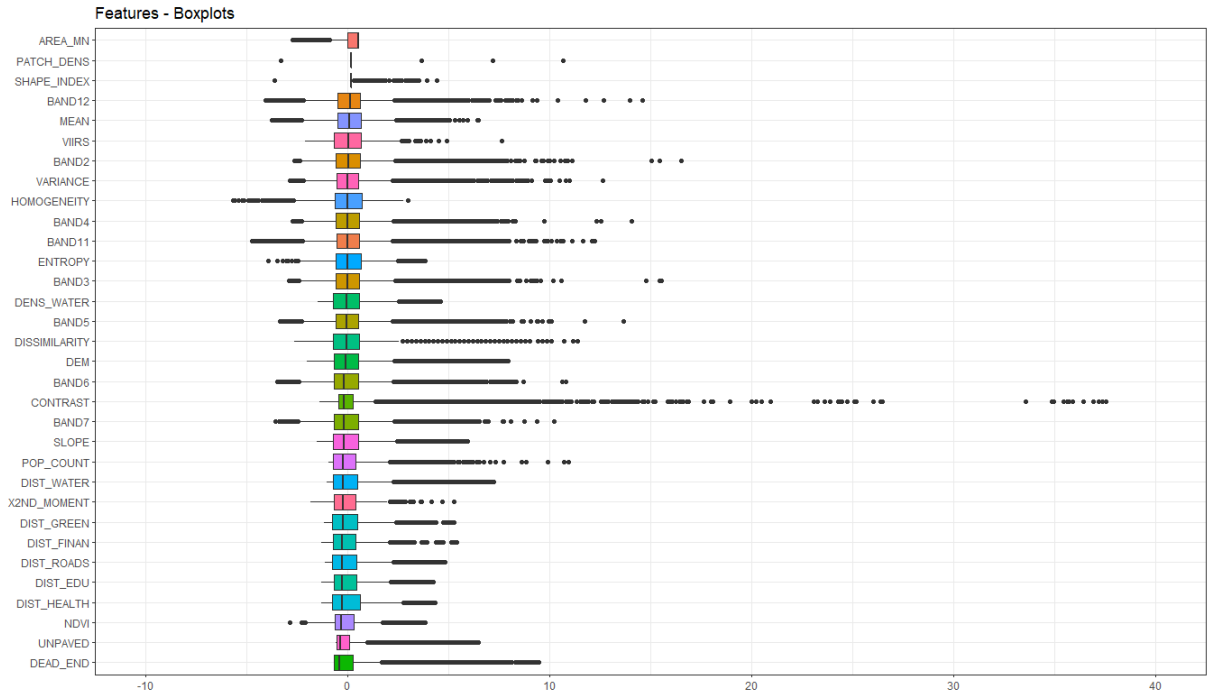


ANNEX 9 – Semi-structured interview with local urban planner.

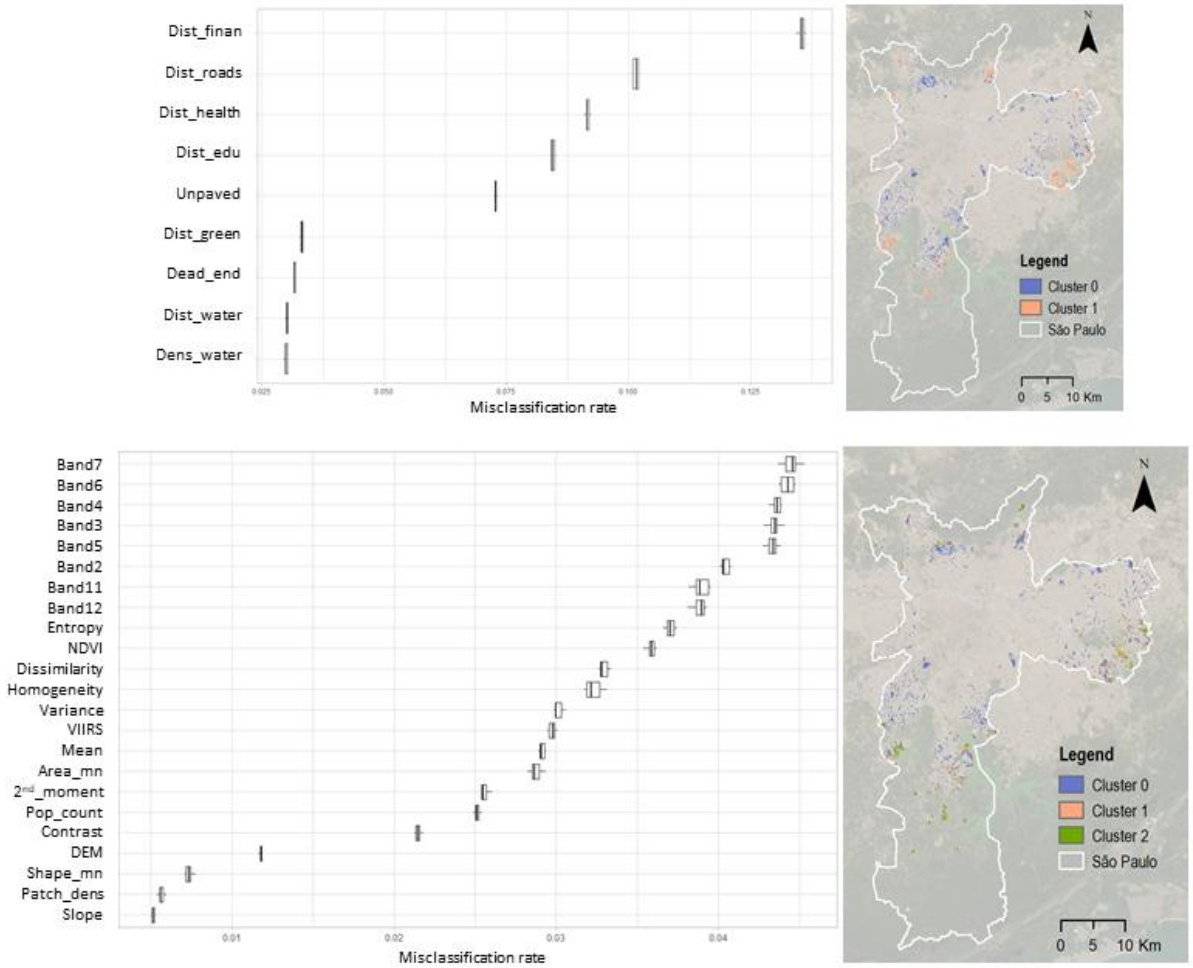
Four pre-determined questions that guided the interview with the local specialist and a snapshot of the presentation displayed during this interview.

- Are the results coherent with the spatial patterns on the ground?
- Which influencing factor is more important for each cluster?
- How do the clusters relate to the socioeconomic aspects of the areas they occupy?
- If it makes sense, how feasible the approach in terms of practicalities?
- What type of analysis would be interesting through the lens of urban planning strategies?

ANNEX 10 – Boxplots of 32 processed input features.



ANNEX 11 – By-product of the k-means coupled with the feature importance tool of Models 1 and 2.



ANNEX 12 – Output of feature selection tool ('FeatureImpCluster' package) as a by-product of the Model 3.