UNIVERSITY OF SÃO PAULO
SCHOOL OF ENGINEERING

CAIO DE BORTHOLE VALENTE PIERONI

Analysis of travel patterns from precarious settlements transit users in São Paulo through smart card data mining

São Paulo

2018

CAIO DE BORTHOLE VALENTE PIERONI

Analysis of travel patterns from precarious settlements transit users in São Paulo through smart card data mining

Dissertation submitted to the School of Engineering at the University of São Paulo for a Master of Science degree.

Supervisor: Prof. Dr. Mariana Abrantes Giannotti

São Paulo

2018

CAIO DE BORTHOLE VALENTE PIERONI

Analysis of travel patterns from precarious settlements transit users in São Paulo through smart card data mining

Dissertation submitted to the School of Engineering at the University of São Paulo for a Master of Science degree.

Area: Transportation Engineering

Supervisor: Prof. Dr. Mariana Abrantes Giannotti

São Paulo

2018

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

São Paulo, ______ de ____________________ de __________

Assinatura do autor: ________________________

Assinatura do orientador: ________________________

Catalogação-na-publicação

Pieroni, Caio de Borthole Valente
Análise de padrões de viagens de usuários de transporte público de assentamentos precários em São Paulo através da mineração de dados de bilhetagem / C. B. V. Pieroni -- versão corr. -- São Paulo, 2018.
154 p.

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

1.Engenharia de Transportes 2.Planejamento de Transportes 3.Mobilidade Urbana 4.Transporte Público 5.Mineração de Dados I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Transportes II.t.

ACKNOWLEDGMENTS

First of all, I would like to thank my parents, Celso and Cristina, those responsible for making who I am today. I have nothing but gratitude for you.

My warm and sincere thanks to my loving partner Marina, who supported me in the most difficult moments (not a few) and helped me on a daily basis with her smile and her partnership. This thesis is also yours.

Thank you to my Professor and Supervisor Mariana for the knowledge, patience and very useful advices, and also for opening a whole new world of opportunities for me. My sincere thank you.

Finally, thank you to all my family (specially my brother Bruno) and colleagues from Poli, from São Bernardo, from work(s), who, in some extent, all made part of this path with me.

“*Essentially, all models are wrong; but some are useful*”

(George Edward Pelham Box)

RESUMO

Título: Análise de padrões de viagens de usuários de transporte público de assentamentos precários em São Paulo através da mineração de dados de bilhetagem. Dissertação de Mestrado.

Dados de bilhetagem permitem compreender e analisar a mobilidade em um nível de detalhe excepcional, porém podem ser considerados restritos para analisar as motivações de viagens dos usuários. A identificação de padrões de viagens pode dar complementariedade semântica aos dados de bilhetagem. Mais especificamente, esta análise de padrões de viagens, aplicada a usuários de transporte público residentes em áreas de assentamentos precários, auxilia uma melhor compreensão das características de mobilidade de uma parcela da população que, historicamente, tem acesso restrito e desigual aos recursos financeiros e às oportunidades no contexto urbano da cidade. O objetivo deste trabalho é compreender padrões temporais e espaciais dos deslocamentos urbanos por transporte público de residentes de assentamentos precários no município de São Paulo, através da mineração de dados de bilhetagem. Para tal, são aplicados três algoritmos de clusterização distintos: *K-means, TwoStep* e *Self Organizing Maps* (SOM). Também são incluídos residentes de áreas de classe média da cidade para analisar as diferenças de comportamento nos deslocamentos urbanos nas áreas estudadas em função da renda domiciliar de seus moradores. Os agrupamentos formados pelos três procedimentos apresentam resultados semelhantes. Grupos com passageiros com evidências de fluxo pendular de trabalho, compostos em sua maioria por moradores de assentamentos precários, sugerem uma associação desses moradores com empregos de baixa remuneração, com suas bilhetagens de atividade principalmente registradas em usos do solo residenciais de média / alta renda e residenciais de baixa renda.

Palavras-chave: Planejamento de transportes; Padrões de viagem; Dados de bilhetagem; Assentamentos precários, algoritmos de clusterização.

ABSTRACT

Smart Card Data (SCD) allow us to understand and to analyze mobility at an exceptional level of detail. However, they can be considered restricted when analyzing users' trip purposes. Identifying travel patterns may provide better context to smart card data. More specifically, this identification may allow the understanding of travel patterns of transit users from precarious settlement areas, a portion of the population that historically has limited and unequal access to financial resources and opportunities. This work aims to understand the temporal and spatial patterns of urban transit movements of residents of precarious settlements in the city of São Paulo, through smart card data mining. For this, we apply three distinct clustering algorithms: K-means, TwoStep, and Self Organizing Maps (SOM). Residents of middle-class areas of the city are also included to compare the behavioral differences in urban displacements in the studied areas as a function of their residents' household income. The results showed that the clusters formed by the three methods show similar results, and clusters with high number of commuters mostly composed by precarious settlement residents suggest an association of this residents with low-paid employment, with their smart card transactions, mainly registered in residential medium / high-income and residential low-income land use areas.

Keywords: Transportation planning; Travel patterns; Smart card data; Precarious settlements, clustering algorithms.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

The use of large volumes of data as a support for decision making has been widely used in several fields of knowledge. Areas such as health and education are examples of fields that benefit from this opportunity to better understand people´s behavior patterns aiming to provide better services. In transportation planning, this trend is not different. It is important to understand travel behavior patterns to meet users' mobility expectations as citizens and also to optimize the transportation infrastructure based on reliable information (JUN; DONGYUAN, 2013).

Various sources of information are usually explored to understand these spatiotemporal patterns of population displacements. The most widely used source of information in transportation planning is conventional data such as household surveys and population census. They provide detailed information on individual and household mobility patterns and also on travel modes and purposes (ANDA; FOURIE; ERATH, 2016).

However, there are some limitations on the use of this type of data. The coverage of household surveys is one of the negative factors. Due to limitations of costs and time, this type of research comprises only a small fraction of the studied universe. Therefore, only a small sample is used to synthesize transport displacements and to represent an entire population (ANDA; FOURIE; ERATH, 2016). In addition, household surveys are limited by their static feature, usually being updated every five to ten years, while the geography of transport patterns changes rapidly. Therefore, for transport studies, dynamic data sources are needed to access the nature of these changes faster. Household surveys usually register only one day of people´s activities, missing the variability, heterogeneity and richness of multiday travel patterns. Recently, many studies have been investing on the development of methods to analyze the demand for travel at a relatively low cost by processing large volumes of data from the transit system – smart card data – in a more dynamic and continuous approach (JUN; DONGYUAN, 2013; LONG; THILL, 2015; PELLETIER; TRÉPANIER; MORENCY, 2011; ZHOU; MURPHY; LONG, 2014).

Over the past two decades, smart card has gradually become the most popular transaction mode in urban transit systems, allowing to analyze and understand mobility at an exceptional level of detail (YU and HE, 2016; ANDA; FOURIE; ERATH, 2016). In addition to its significant increase in sample size, smart card data allow observing and interpreting continuous patterns over time, as opposed to static information from conventional surveys (DEVILLAINE; MUNIZAGA; TRÉPANIER, 2012). Another positive feature is that smart card data do not require additional infrastructure or investments to be obtained, since individual travel information (transaction date and time / route code / card type) is already generated from using the smart card in the ticket gate of a transit system.

In order to reach this level of detail in the available information - by transforming this large volume of data into valuable information –, it is necessary to understand the best way to process it, with adequate data mining techniques. The treatment of this large volume of data has been a challenge, as there are also some drawbacks in using large amounts of data such as smart card data. The first one is the intrinsic characteristic of smart card data that regards only information of transit users, leaving aside important modes such as automobiles, motorcycles and bicycles. Also, the dependence of computers with large processing capacity and storage space of these data are important requirements for their effective use (DEVILLAINE; MUNIZAGA; TRÉPANIER, 2012). In addition, smart card data do not have the users' socioeconomic and demographic information, because they do not have information on the users holding the cards (which is the tool to record the trip). In many cities – as is the case of São Paulo – the payment system does not require validation when alighting, thus not recording this type of information. There is also no information about the characteristics of the trip made by the user, such as the trip purpose, transfers, or the local type of origin or destination. The information obtained lacks the characteristics and motivations of the users themselves.

In this sense, exploring the travel patterns from smart card data to overcome the aforementioned restrictions, providing greater robustness by adding better context to characterize the user of the transit system, is a huge challenge. We intend to contribute by clustering similar travel patterns from users of specific areas of the city of São Paulo, in Brazil, through data mining techniques of smart card data.

Identifying homogeneous travel behavior groups can be useful in a variety of applications. It has the potential to help public transport planners understand their customer behaviors, being able to provide a more suitable service according to the demand (AGARD; PARTOVI-NIA; TRÉPANIER, 2013; ZHAO et al., 2017), and help predict user trips (ORTEGA-TONG, 2013). Also, it can be used to assess the performance of the transit network and to detect irregularities, such as frauds or defective equipment (EL MAHRSI et al., 2014). From a social point of view, evaluating the different travel behavior groups may suggest what the reasons for their similarity of patterns could be – regarding an income issue, for example.

Specifically evaluating the importance of low-income population in urban cities led us to concentrate the research on areas defined as precarious settlements. The São Paulo Metropolitan Area (SPMA) contains about 600 thousand private households occupied in sectors of precarious settlements, with more than 2 million inhabitants (IBGE, 2010). It represents about 19% of the national total in terms of households and 17% in terms of inhabitants, which shows the relevance of this population. The distribution of the precarious settlements is predominantly peripheral in the SPMA. Exploring travel patterns in these areas may bring a better understanding about the mobility characteristics of population segment that historically has limited and unequal access to financial resources and opportunities. Proposing to investigate travel patterns of transit users in these regions as to provide better information to support urban and transportation planning alternatives, intending to change the unfavorable conditions of mobility this population is currently facing.

Regarding this context, special attention will be given in this work in order to better investigate low-income formal and informal workers travel patterns. About 22% of the total wage employment of the six major metropolitan areas in Brazil (Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo and Porto Alegre) was low-paid in 2009 (Fontes et al., 2012). The authors defined low-paid employment in relative terms as hourly wages that are equal to or less than two-thirds of the median hourly wage. This definition is extensively used in international comparative studies (FONTES et al., 2012). During recent economic crisis and political instability, as Brazil has gone through in the last few years, the proportion of unemployment, poverty, inequality and job informality tends to increase (MARQUES; SARAIVA, 2017), especially in regions with a predominance of lower income population, as in precarious settlements. Therefore, being able to identify travel pattern from low-paid employees using the transit system in these regions could be important for understanding local mobility, their economic feature and allowing transit authorities to evaluate their current services for this vulnerable population.

The regions of precarious settlements, defined for the research development, consider their urban insertion throughout the city:

- Paraisópolis: belonging to the district of Vila Andrade, in the southwestern zone of São Paulo.
- Cantinho do Céu: belonging to the district of Grajaú, in the southern zone of São Paulo.
- Parque Taipas: belonging to the district of Jaraguá, in the northwestern zone of São Paulo.
- São Francisco Global: belonging to the district of São Mateus, in the eastern zone of São Paulo.

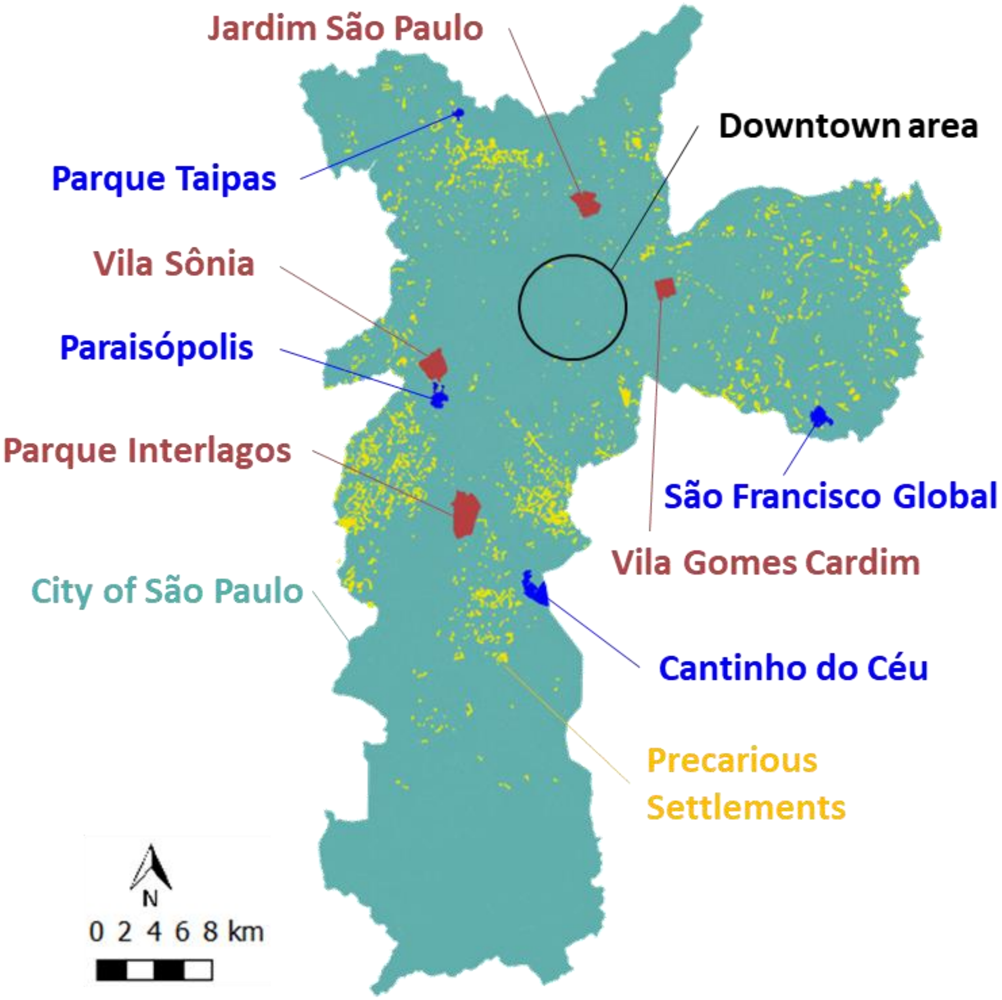
It is sometimes difficult to evaluate results of a given process without proper reference values for comparison. Here, having information about the average distance of displacement, or average gap hours between ticketing in the transit system of a given group are interesting indicators, but for someone who does not have the spatial knowledge of São Paulo, or the notion of how long people spend working or studying in the city, the values themselves could be difficult to interpret. Also, comparing areas of precarious settlements with areas with distinct characteristics could confirm (or refute) the assumption that residents of precarious settlements have different travel behaviors and, besides all the difficulties from other spheres, are also impaired regarding the transport system due to their economic conditions.

Therefore, four additional areas of the city of São Paulo, this time from more privileged neighborhoods, were selected to be also clustered together with the precarious settlement areas in order to allow evaluating the differences between them. Each new area belongs to the similar zone as one of the precarious settlement areas, and are presented as follows:

- Vila Sônia: district in the West Zone of São Paulo;
- Parque Interlagos: belonging to the district of Socorro, in the south zone of São Paulo;
- Jardim São Paulo: belonging to the district of Santana, in the north zone of São Paulo;
- Vila Gomes Cardim: belonging to the district of Tatuapé, in the east zone of São Paulo.

Figure 1 presents the geographic location of the selected areas both of precarious settlements and the additional privileged areas.

Figure 1 – Middle-class and precarious settlement areas studied here



Source: The authors' own elaboration

The choice of geographically far apart areas aims to compare eventual differences between the travel patterns from distinct localities of São Paulo.

1.1 MOTIVATION

Our motivation is described as follows. Firstly, understanding the travel pattern behavior can contribute to the city transportation planning, either as inputs for models of estimating the demand for public transport systems, pursuing demand forecasts with greater accuracy, or even as inputs for load factor evaluation in transportation systems. Information such as travel patterns groups and their variability over time can serve as a decision support tool for stakeholders in both the operation and strategic level of transportation planning. This investigation of smart card data can help to better understand the dynamics low-income areas, and the methods and results here described will hopefully contribute to future studies in this area.

A methodological approach developed for structuring smart card data (big data) into a database that has all the information needed for this analysis at a disaggregated (individual) level and much easier to manipulate (small data) may be replicated in smart card data from other localities.

By analyzing smart card data from precarious settlements' residents, the travel behavior patterns investigation is expected to provide a better understanding about transportation demand from this vulnerable population, specifically considering the low-paid employment segment, contributing to a path to better consider the socially excluded population in transportation models.

1.2 OBJECTIVE

Our main goal is to investigate spatiotemporal multiday travel patterns of transit users residing in precarious settlements in the city of São Paulo, Brazil, through smart card data mining techniques.

This work also has specific objectives as follows:

(i) Providing better context to smart card data by analyzing spatiotemporal patterns, besides describing and evaluating similarities and differences between the travel patterns of the different homogeneous groups;

(ii) Evaluating the feasibility of identifying low-paid employees' travel patterns using the proposed methodology;

(iii) Evaluating the three different types of clustering algorithms applied to the database: K-means, TwoStep and Self Organizing Map (SOM), along with their similarities and dissimilarities in the results.

1.3 RESEARCH STRUCTURE

This research is organized in five more chapters, described as follows.

After this introduction, Chapter 2 presents a literature review of the uses of smart card data in public transport researches, of travel pattern analysis and the clustering algorithms. Chapter 3 describes the dataset used and the selected regions of analysis. Chapter 4 explains the methodological approach of data treatment for mining large volumes of smart card data into small and structured database and the three different clustering methods used for classifying the transit users. Chapter 5 presents the results obtained. Finally, Chapter 6 presents the conclusion and the potential perspectives for future research.

2 LITERATURE REVIEW

This chapter firstly presents an overview to the main literature related to the different studies purposes while using smart card data, in Section 2.1, indicating the wide range of studies that this type of data can provide. Section 2.2 focuses on the researches that have addressed the problem of analyzing spatiotemporal travel patterns or travel behavior of public transport users through smart card data, and Section 2.3 presents previous works that have utilized the proposed clustering algorithms. The general term smart card data here presented comprises three different types of data: the automated fare collection system (AFC); the automatic vehicle location systems (AVL), collected by GPS; and the general transit feed system (GTFS).

2.1 SMART CARD DATA IN PUBLIC TRANSPORT PLANNING

The potential of smart card data has been used for strategic, tactical and operational performance of public transport systems (PELLETIER; TRÉPANIER; MORENCY, 2011; WILSON et al., 2009). It is a continuous source that provides a complete and real-time travel information of all users of the public transport system, allowing better understanding the behavior of the user and helping the improvement of the public transport system (LONG; THILL, 2015). Bagchi et al. (2005) describe the potential of smart card data for transit planning, discussing its capability of capturing information about passenger trips and pointing out the advantages and shortcomings of its use. Pelletier; Trépanier; Morency (2011) present a wide review of studies using smart card data in the public transit context. Also presenting an overview of smart card researches, Anda; Fourie; Erath (2016) give examples in which smart card data can be explored, such as organizing steps for rebuilding individual journeys (aiming to feature commuting travels or producing origin-destination transit matrices, for example), and even including them in an agent-based transport model.

However, Long; Thill (2015) also point out the need to take some precautions regarding smart card data limitations. Pelletier; Trépanier; Morency (2011) also mention some disadvantages and precautions in analyzing smart card data, such as privacy and security measures in data handling, and the lack of socioeconomic or demographic attributes information as the information is strictly about the journey undertaken, not the user itself. In this sense, depending on the study object, conventional information collection is necessary to complement smart card data.

Some studies have been based on both smart card and household travel survey data for exploring commuting journeys. Zhou, Murphy and Long (2014) investigate the efficiency of home-work commuting journeys undertaken in Beijing, China, through the 2008 smart card data for buses and the 2010 Beijing household survey for cars, verifying if there are excessive work trips and highlighting the potential of smart card data to trace the efficiency of commuting patterns in public transport. Long and Thill (2015) also analyze Beijing home-work commuting movements from smart card and household survey data. They suggest the feasibility of analyzing urban structures using smart card data as an alternative or to complement conventional travel behavior surveys.

Many transit systems round the world do not have alighting validation records, with passengers only swiping their cards when boarding public transport. Therefore, different methods have been developed using smart card data to estimate the alighting point for individual trips. The most widespread method is proposed by Barry et al. (2002), assuming that users of transit systems begin their next trip close to the destination of their previous trip; and that transit systems end their last trip of the day at the same origin where they began their first trip. Munizaga; Palma (2012), for example, implement this method for estimating the alighting location of trips for the city of Santiago, Chile, obtaining an origin-destination matrix of public transport through smart card data. Zhou; Murphy; Long (2014) infer the alighting location based on a consecutive 5-day-week ticket data, assuming that the maximum walking distance traveled would be 500 meters. Trepanier et al. (2007) use the same approach to estimate bus alighting in Gatineau, and Zhao et al. (2007) for boarding locations in the Chicago CTA system.

Boarding and alighting location and time data are used in the literature to evaluate travel patterns from smart card data. Tap-in and tap-out systems generate boarding-alighting structured data (Zhong et al. (2015), Zhao et al. (2017), Yu and He (2016)), while tap-in only system require inference of alighting locations (Agard et al. (2006), Morency et al. (2007), Agard et al. (2013), El Mahrsi et al. (2014)).

Concerning the use of smart card data developed in the São Paulo area, Farzin (2008) compared the 2007 household survey matrix from the São Paulo Metro with an origin-destination matrix from smart card data. Although Farzin (2008) does not consider treating modal transfers, the study indicates that the larger sample size, available from the smart card data, brings more detailed matrix results compared to the matrix obtained by the household survey – which may hide smaller flow patterns not captured in the samples. However, by that time São Paulo bus system fleet had only a very low percentage of GPS equipped vehicles, therefore limiting the analysis to a limited geographical sample.

Smart card data are widely known to lack socio-demographic information, and enriching them with semantic meaning would be a step further on its proper use (PELLETIER, TRÉPANIER; MORENCY, 2011; ANDA; FOURIE; ERATH, 2016). An example of enrichment is looking at trip purpose inferences, such as Lee; Hickman (2013). They develop an inference process of trip purpose from smart card based on heuristic and behavioral rules and build a classification through a decision tree of the results from the previous step, together with a set of tests to verify the performance of the model. Devillaine; Munizaga; Trépanier (2012) developed a method to differentiate an activity from a transfer in trips and another to assign them in trip purposes, applying it in Santiago, Chile and Gatineau, Canada. The criterion for differentiating an activity from a transfer is the interval between transactions of a particular card, and the criteria used to infer the purposes are the type of card, the time of the activity, the transaction position on the day and land use of the destination zone.

2.2 TRAVEL PATTERN ANALYSIS

Research in the area of urban travel behavior attempt to “describe, analyze, and model urban travel-activity patterns as complex entities” (PAS, 1988). This section presents some general examples of these researches for urban travel patterns besides some researches focusing specifically on characterizing public transport users' travel behavior using smart card data.

Continuous travel data from users were traditionally obtained from travel diaries surveys. However, this data is very time consuming to generate, and respondents might voluntarily avoid to register some trips. The first work to access long observation periods for travel pattern analysis is the data collected in Upsalla, Sweden, in 1971. Hanson; Huff (1986) use these detailed activity diaries kept by a sample of individuals for 35 consecutive days to classify individuals in homogeneous travel behavior groups. The travel behavior measures used to classify individuals were: the proportion of out-of-home time spent on different activity purposes, the proportion of single-stop trips, the number of trips per day, and the proportion of walking trips. Later, Pendyala et al. (2000) describe several studies analyzing travel variability, examining and comparing measures of travel behavior variability, and showing only a small percentage of individuals who repeated their behavior on all days.

Smart card data collection systems were introduced in public transport to replace the traditional paper tickets, allowing passengers to retain their cards for longer periods (BLYTHE, 2004). While the original objective with smart card systems implementation was to improve fare revenue management, this data allows planners to delve into continuous travel information collection for all card holders in the system. With the widespread implementation of smart card in public transport fare systems throughout the 2000s, a large volume of data started being generated each day in the existing systems. This revolution in data collection enabled temporal profiles and travel pattern analysis. Agard et al. (2006) and Morency et al. (2006) are among the first to address the issue regarding travel pattern and smart card data. Agard et al. (2006) study mining public transport user behavior and Morency et al. (2006) study the variability of transit users' behavior. Both aggregate trips into transactions, each representing the daily profile of a given smart card on a given date. Morency et al. (2006) apply k-means clustering to identify (separately for each user) clusters of similar days regarding boarding times. Agard et al. (2006) apply Hierarchical Agglomerative Clustering (HAC) and k-means to the transactions to study group behavior, classifying users into four groups according to the repetition of the starting period of each journey: two groups of users starting at peak hours and during the first part of the day, and two groups with low travel frequency and no clear travel pattern. Both suggest how analyzing the composition of clusters over the studied weeks helps identify groups such as “typical workers" and atypical behaviors. Utsunomiya et al. (2006) are also among the precursors to analyzing the frequency and the consistency of travel patterns of passengers using smart card transactions.

Ma et al. (2013) extract individual transit riders' travel patterns and assess their regularity from Beijing smart card data. The study focuses on characterizing the spatial and temporal travel patterns of transit riders on an individual basis through DBSCAN, and determining the regularity of these patterns through k-means++ (and enhanced k-means) and the rough-set theory. Zhong et al. (2015) approached the issue by measuring variability at individual and also at aggregated levels using one-week smart card data from Singapore. For the individual measurement, Zhong et al. (2015) constructed a profile of trip starting-time for one week and a correlation matrix of the temporal patterns of each day. Similarly, for the aggregated mobility patterns, a correlation matrix of the temporal patterns is made, now for bus stops and trains stations. Lastly, a spatial network is constructed from an OD matrix of daily trips, and community detection and PageRank centrality are applied to the data, concluding that, although variability of mobility patterns exists at an individual and spatial aggregated scale throughout the analyzed week, the overall spatial structure of the urban movement remains practically the same.

Yu and He (2016) also propose to extract individual transit riders’ travel patterns from 2014 data provided by the Guangzhou transit agency. Firstly, an estimation method is utilized for data pre-processing to obtain the individual trip information and the trip chain of each rider. Afterward, the DBSCAN clustering algorithm is used to mine the travel pattern from each transit rider and to identify the regular OD and time that the rider usually travels. A travel pattern clustering is conducted: spatial-temporal regular, spatial regular and temporal regular. Yu and He (2016) conclude that the majority of transit riders have less than 5 kinds of travel patterns; the number of spatial regular of temporal regular travel patterns is generally more than spatial-temporal regular, due to the looser criterion in this scale, and the distribution of the number of temporal regular patterns is evener and broader than the distribution of spatial regular.

Lathia et al. (2011) discuss how smart card data can be used to reveal hidden individuals' behaviors by comparing an online survey results (perceived) and real London smart card data (actual), studying various aspects of this comparison such as trips per day frequency and regularity, atypicality of travel modality and origin and destination stations, besides cash-fare purchasing habits.

In a more recent study, Zhao et al. (2017) propose to understand the spatiotemporal travel patterns of individual passengers of bus and metro systems from a smart card transaction dataset from Shenzhen, China, conducting two types of travel pattern analysis: statistical-based and clustering-based. On the statistical-based, a regularity analysis is performed through the three types of patterns separately (spatial, temporal and spatiotemporal). On the clustering-based, only spatial and temporal patterns are separately evaluated using the K-means algorithm and city-block distance. Their results are then correlated through a conditional probability matrix. Zhao et al. (2017) conclude that if a passenger is temporally regular, the passenger is very likely also spatially regular.

El Mahrsi et al. (2014) propose a different approach to understanding travel patterns using smart card data conjointly with socioeconomic data in Rennes, France. Firstly, temporal passenger profiles based on boarding information are constructed and a generative model-based clustering approach is applied. After the clustering, based on boarding information, passengers were assigned to residential areas. Afterward, socioeconomic data of the city are clustered by the Hidden Random Markov Field (HRMF) model, and 7 socioeconomic clusters are formed, distributed throughout the city. Finally, the passengers are assigned to a specific type of socioeconomic cluster. From the results of the latter step, El Mahrsi et al. (2014) conclude that it is possible to identify different groups of workers engaging in home-work commutes at different times of the day, students traveling to and from school and others, and that some socioeconomic classes are more susceptible to using public transportation than others.

Lathia et al. (2013) also structure passenger's trips into a weekday profile describing their temporal habits to discuss how smart card data can personalize transport information services. Four-time bins are used, early morning, morning rush, day time, and evening rush. Hierarchical agglomerative clustering is performed to discover groups of passengers with different habits, concluding that the use of public transportation can vary considerably between individuals. Briand et al. (2017) proposes to regroup passengers based on their temporal activities in their public transportation usage by applying a Gaussian mixture generative based model, also evaluating the cards' cluster memberships from 2005 to 2009.

Alsger et al. (2018) propose a methodology to infer passengers' trip purpose using different data sources, such as smart card data, a land use database, a transport model of Queensland, GTFS and O-D survey data. Amaya et al. (2018) propose a method to estimate the residence zone of card users, by calculating the center of gravity of the coordinates of the first morning transactions of passengers. The distance from the positions of the first morning transactions to the center of gravity were calculated, and only centroids with the largest distance lower than a pre-defined walking distance were considered. Ma et al. (2017) present data mining methods to identify transit commuting patterns based on one-month transit smart card data by both mining spatiotemporal travel regularities and extracting individual-level residence and workplace. Langlois et al. (2016) investigate clusters of users with similar activity sequence structure through principal component analysis (PCA) based on the longitudinal activity sequence of each user derived from smart card data.

2.3 CLUSTERING METHODS

For the present research the K-means method was chosen as it is one of the most cited for smart card travel pattern investigation. To sum-up a comparison was made with two other methods, not so frequently used in smart-card analysis but increasingly being adopted in transport planning and spatial analysis field, TwoStep and Self Organizing Maps, respectively. This subsection briefly introduces the theoretical grounds from the clustering methods.

The k-means clustering algorithm is one of the well-known clustering algorithms and has been widely used in analyses with segmentation in the area of public transport due to its ability to compute relatively large data sets and requiring few parameters to be fixed (MA et al., 2013; AGARD et al., 2013). As aforementioned, Agard et al. (2006), Morency et al. (2006) and Zhao et al. (2017) also used k-means clustering to explore travel pattern through smart card data. McNally; Kulkarni (1997) also used k-means to identify a range of land use transportation systems by clustering network and land use inputs, aiming to develop an empirical assessment of the interaction between the land use transportation system and travel behavior. Bouman et al. (2013) use enhanced k-means clustering to derive important activity time intervals from smart card data, identifying activity patterns that differ from the typical time windows associated with home-work activities. Kieu; Bhaskar; Chung (2013) mine travel regularity through spatial regular origins-destinations and temporal usual traveling time from transit users, with the passengers being clustered into frequent and infrequent users according to the number of trips taken using the K-means clustering algorithm. DBSCAN is also performed to explore the regularity of each cardholder.

The SPSS TwoStep clustering method is also used for travel pattern and passenger behavior analysis. Fonzone et al. (2013) perform a web-based survey to better understand the behavior of public transport passengers using networks with high-frequency services and use the TwoStep clustering to understand the choice of different routes in transit networks. Pitombo; Kawamoto; Souza (2011), using TwoStep clustering and decision tree, relate the socioeconomic characteristics, activity participation, land use patterns and travel behavior of the São Paulo Metropolitan Area residents through household travel surveys. The same type of source was used by Cerin et al. (2007), who use TwoStep clustering to associate access to destinations with walking for transport also through household travel survey. Respondents reported perceived proximity of destinations, transport-related walking, reasons for neighborhood selection, and socio-demographic characteristics. Measures of access to destinations were associated with transport-related walking, being workplace proximity the most significant contributor to transport-related walking. Zoltan; McKercher (2015) use

TwoStep clustering of smart card data to examine tourist movements in the Canton of Ticino in southern Switzerland, seeking spatially or activity-based clusters. Results showed there was only limited evidence of activity-based segmentation, with movement patterns defined largely on a spatial dimension.

In the geospatial sciences, the self-organizing maps (SOM) clustering have extensively been applied as an unsupervised classifier of remotely sensed multispectral data and has also been adopted as a tool for geographic feature identification (YAN; THILL, 2009). Himanen et al. (1998) are the first to explore the applicability of SOMs to identifying archetypical daily travel patterns. Some other studies use SOM to organize demographic and housing data gathered from censuses and surveys to investigate patterns of change in urban and regional systems (Hatzichristos, 2004; Koua; Kraak, 2004; Skupin; Hagelman, 2005).

All the aforementioned researches stress a general evaluation of travel patterns from smart card data, mainly looking for commuting patterns, and do not focus on the study of low-income areas. Here, however, we aim at linking smart card data and travel pattern, identifying low-income residents with low-paid employments by clustering methods and comparing this low-income residents' behavior with residents from other parts of the city, evaluating their differences and similarities of displacements throughout the city. One of the closest studies to this issue, identified at the literature review, is Lathia et al. (2012), which infers London's community well-being from smart card data by examining the correlation between London urban flow of public transport and census-based indices of the well-being of London's census areas. The results suggest that well-off areas attract people from communities of varying social deprivation, but deprived areas do suffer from social segregation and tend to attract people only from other deprived areas.

Also, previous works involving clustering smart card data usually select one method of clustering and perform their analysis based on the results. Here, three different clustering algorithms are performed. The common clusters formed between all the three methods can give even more robustness to the analysis, and an evaluation of advantages and shortcomings of the use of each method can be a reference for future work in this regard.

Finally, some existing studies consider either clustering the trips of each passenger individually or disregard the passengers' identity. We aim at exploring travel patterns per passenger, therefore clustering each passenger as a single observation with a profile for each user. El Mahrsi et al. (2014) and Lathia et al. (2013) use a similar definition for clustering.

Analyzing the existing literature regarding smart card data travel pattern evaluation through clustering methods, the contributions of this work can be listed as stated below:

- Enrichment of semantic information to smart card data specifically aiming to explore behavior of precarious settlements residents, who had not been much investigated in terms of travel pattern through smart card data;
- Identification of differences in travel patterns when comparing precarious settlements with other middle-class areas, especially regarding low income employment;
- Proposition of a method for inferring residence of transit users, through the DBSCAN algorithm, that considers the main cluster formed for each passenger to calculate the residence centroid;
- Application of three different clustering algorithms to evaluate the consistency of the results.

Table 1 presents a summary review of some important researches regarding travel pattern studies using smart card data.

Table 1 – Review of studies on travel pattern and travel behavior using smart card data

Author	Objective	Clustering Method	Number of Clusters Determination Method	Clustering evaluation	Atributes for Clustering	Data Period
Agard et al. (2006)	Travel behaviours, regularity, daily patterns, variability	K-means, HAC	A first grouping is computed with a k-mean method to provide 20 groups. The result of the k-mean clustering becomes the input of the HAC	Temporal	20 binary variables, representing 5 weekdays X 4 periods per day	81 days - Between January 10th and April 1st, 2005
Zhong et al. (2015)	Variability of: individual and aggregated mobility patterns and of spatial networks	Correlation matrix	Number of bus stops	Temporal	Temporal profile patterns across one day or one week (trip starting time)	One-week smart card data collected in April 2014
		Community detection	Not explained	Spatial	A spatial network is then constructed from an O–D matrix of daily trips	
Morency et al. (2007)	Regularity indicators by spatial and temporal variability	Frequency of use of the bus stops is studied, in order to express a level of regularity	-	Spatial Variability	-	276 days - Between January 1th and October 4th 2005
		K-means	Researcher experience	Temporal Variability	24 binary variables, representing the hours of a day	
Ma et al. (2013)	Individual travel pattern recognition	DBSCAN	Not necessary - Ɛ and minPts defined arbitrarily	Spatial and temporal	Transit riders' recurring boarding/alighting locations and times	The week of Monday July 5th to Friday July 9th, 2010
	Regularity clustering	K-Means++	Arbitrary - Very High (VH), High (H), Medium(M), Low (L), Very Low (VL)	Spatial and temporal	Number of travel days, Number of similar First Boarding Times, Number of similar Route Sequences, Number of similar Stop ID Sequences	

Author	Objective	Clustering Method	Number of Clusters Determination Method	Clustering evaluation	Atributes Used for Clustering	Period of data
Morency et al. (2006)	Variability of transit users behaviour	K-Means	Not explained (29 for elderly, 15 for Regular Adult)	Temporal	24 binary variables, representing the hours of a day	276 days - Between January 1th and October 4th 2005
		Activity rate on the transit network, Boardings per day, Enumeration of boarding stops	-	Spatial Variability	-	
Agard et al. (2013)	Regularity of public transport behaviour	K-Means	Through dataset mapping	Temporal	Three dimensional Cartesian coordinate system (X,Y,Z), calculated through 24 binary variables, representing the hours of a day	One-year period between January 1st and December 31st, 2008
El Mahrsi et al. (2014)	Temporal behavior of the passengers in a public transportation system and how passenger travel habits relate to socio-economic characteristics	Generative model-based clustering (mixture of unigrams)	EM algorithm to estimate the mixture of unigrams models while varying K from 2 to 30. Then Data-driven technique called the“slope heuristic”	Temporal	168 variables: 24 hours x 7 days in a week	25 days - From March 31st, 2014 up to April 25th, 2014
		Hidden Random Markov Field (HRMF)	Manually	Spatial	Area, Income, Population	
Briand et al. (2017)	Regroup passengers based on their temporal habits in their public transportation usage	Gaussians mixture generative model / CEM and EM combined algorithmes	H represents the number of Gaussians and K represents the number of clusters - Integrated Completed Likelihood (ICL)	Temporal	The new user id, the day (Monday - Sunday), and the hour of validation	Between the 1st and 28th of February for the 2005–2009 period

Author	Objective	Clustering Method	Number of Clusters Determination Method	Clustering evaluation	Atributes Used for Clustering	Period of data
Ortega-Tong (2013)	Identify public transport passenger travel patterns	K-medoids	Within-cluster variation and the Davies-Bouldin index	Spatial and temporal variability	Travel Frequency; Journey Start Time; Origin Stop/Station Frequency; Travel Distance; Activity Duration; Fare Discounts; Percentage of Bus Exclusive Days; Percentage of Rail Exclusive Days	2 years of data - 2011 and 2012
Zhao et al. (2017)	Understand the spatio-temporal travel patterns of individual passengers	OPTICS clustering method for stations Tempora and Spatiall: K-means algorithm and city-block distance	Silhouette score	Spatial and Temporal	Temporal: Ft1, Ft2, Ft3, Ft4. For a passenger, Fti is the proportion of active days during the i th time slot Ti to the total Spatial: Fs1, Fs2, Fs3, Fs4. The i th feature is the proportion of the passenger's active days to access the i th OD pair.	1 month - Nov 1 to Nov 30, 2014
Yu and He (2016)	Extract individual transit riders' travel patterns from massive dataset	DBSCAN	Not necessary - Ɛ and minPts defined by sensitivity analysis	Spatial and Temporal	Transit riders' recurring boarding/alighting locations and times	One month - September, 2014
Zhou et al. (2017)	Infer the functions occurring around the metro station catchment areas according to the patterns of staying activities derived from smart card data	DBSCAN	Not necessary - Ɛ and minPts defined arbitrarily	Temporal	Start time, the end time, and the frequency	One month - August, 2016
		K-MEANS	Bayesian Information Criterion (BIC) - 6 clusters defined	Temporal	Percentage of activities labeled as each cluster in DBSCAN that happen around each station	

Source: The authors' own elaboration

3 DATASET AND STUDY AREAS

This chapter characterizes the study areas, as well as presents some clarifications about the selected areas. Firstly, it is important explicit that "precarious settlements", as adopted by the new National Housing Policy of Brazil (PNH, in Portuguese), features inadequate urban settlements areas occupied by low-income residents (BRASIL, 2010). Irregular land divisions, favelas and alike, tenements or degraded housing complexes are a few examples of precarious settlements. Still, according to Brazil (2010), some common features of precarious settlements are:

- mostly residential areas inhabited by low-income families;
- the precarious housing conditions, characterized by numerous shortcomings and inadequacies, such as land irregularity; absence of environmental sanitation infrastructure; location in poorly served areas by a transport system and social facilities; lands subject to floods and geotechnical risks; excessive density, insalubrity and constructive deficiencies of the housing unit;
- historical background, related to strategies used by the low-income population to enable a solution to their housing needs, due to the insufficiency and inadequacy of state initiatives addressed to the issue, as well as the incompatibility between the income level of the majority of workers and the price of residential units produced by the formal real estate market.

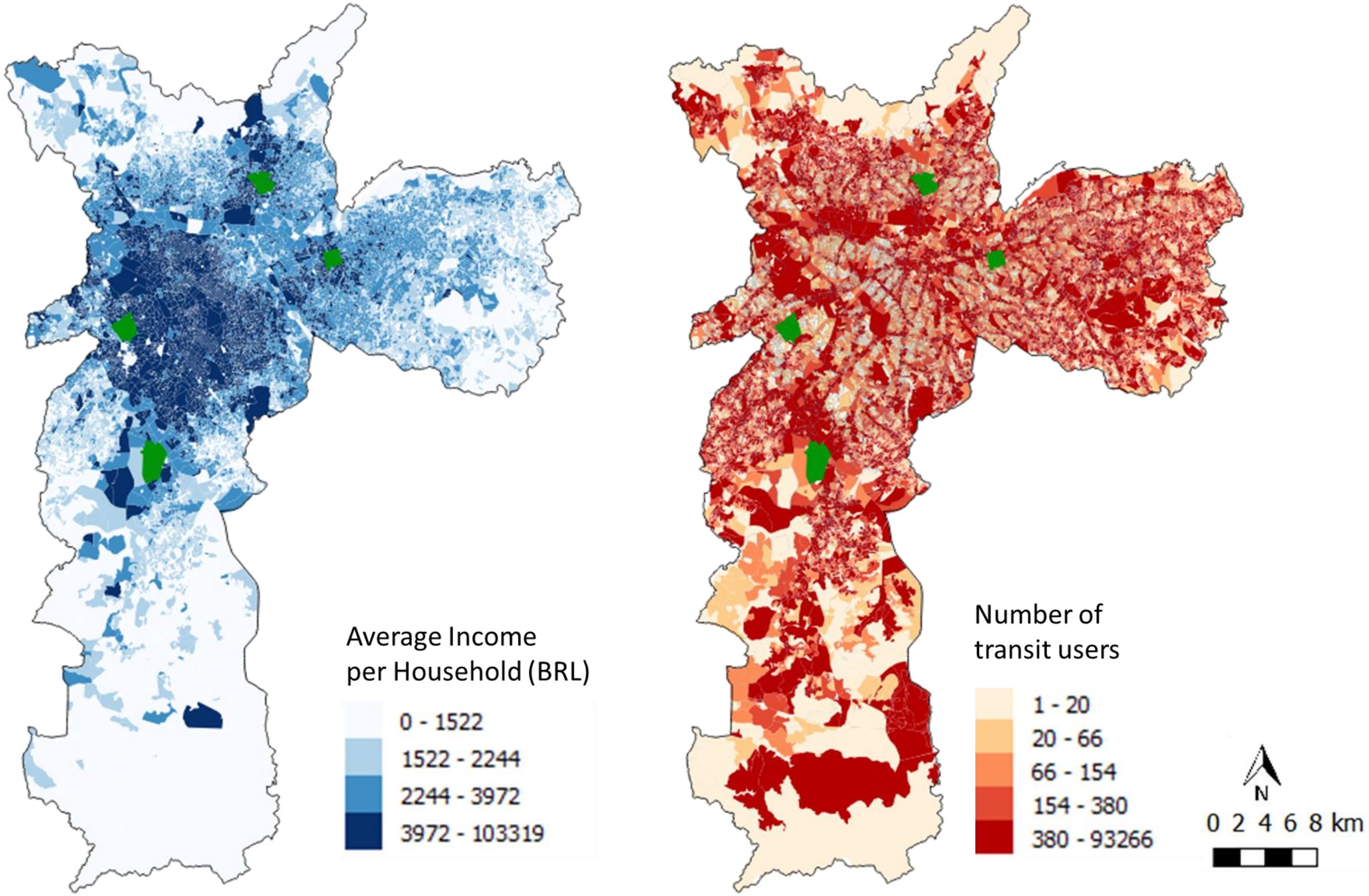
The National Bureau of Statistics (IBGE, 2010), on the other hand, uses a more specific definition in its surveys. The term "subnormal household agglomerate" is used to feature one of the precarious settlements (the favelas, in Portuguese). The term “slum” is used in the international literature to feature precarious settlements (QUEIROZ, 2015). Here, for standardization purposes, all the different terms of this housing features are simplified to precarious settlements.

The additional areas of privileged residents to compare with the precarious settlement areas are selected aiming the minimum distance from the corresponding precarious settlement, and also considering two basic criteria: the selected area must be in the first quartile of income per household distribution of all the areas in São Paulo – the highest quartile of average incomes per household; and there should be a minimum number of transit users in the given area when analyzing the smart card data. People with higher affordability tend to use less public transport systems throughout the world and, here, a significant number of transit users are necessary to make a proper comparison between areas.

The income of each area is obtained by aggregating income data from the 2010 census tracts data from The National Bureau of Statistics (IBGE) with GIS techniques. The minimum number of transit users for each area is defined by the smallest number of transit users between the areas of precarious settlements already selected – in this case, from Parque Taipas, with 1,218 transit users. Therefore, all new areas must be with at least 1,218 passengers using smart card data. The residence locality of each card holder is not provided by the original smart card data, being an inference method developed at the present work and better described in Chapter 4. Considering the aforementioned criteria, the new selected areas for comparing with the precarious settlements are Vila Sônia (West Zone); Parque Interlagos (South Zone); Jardim São Paulo (North Zone); and Vila Gomes Cardim (East Zone).

Figure 2 presents the geographical distribution of income (left) and the number of transit users (right).

Figure 2 – Geographic distribution of income (left) and number of transit users (right)



Source: The authors' own elaboration

The city of São Paulo can be considered the financial, corporate and mercantile center of South America and is one of the most densely populated cities of the Americas, with 11,253,503 inhabitants (IBGE, 2010). Likewise, São Paulo bus system is one of the largest in the world. According to São Paulo Transporte (SPTrans) – the company responsible for managing the public transport system by bus in São Paulo – the city operates with more than 1,300 bus lines and 15,000 vehicles. GPS records from this fleet are obtained every 40 seconds, resulting in approximately 26 million daily GPS records (ARBEX; ALVES; GIANNOTTI, 2016).

The smart card and GPS data used in this research were provided by São Paulo Transporte (SPTrans). The period of data was 11-week long (77 days), from 30th May to 14th August 2016. This period represented approximately 9.5 million unique cards in the system, with about 803 million transactions along the period.

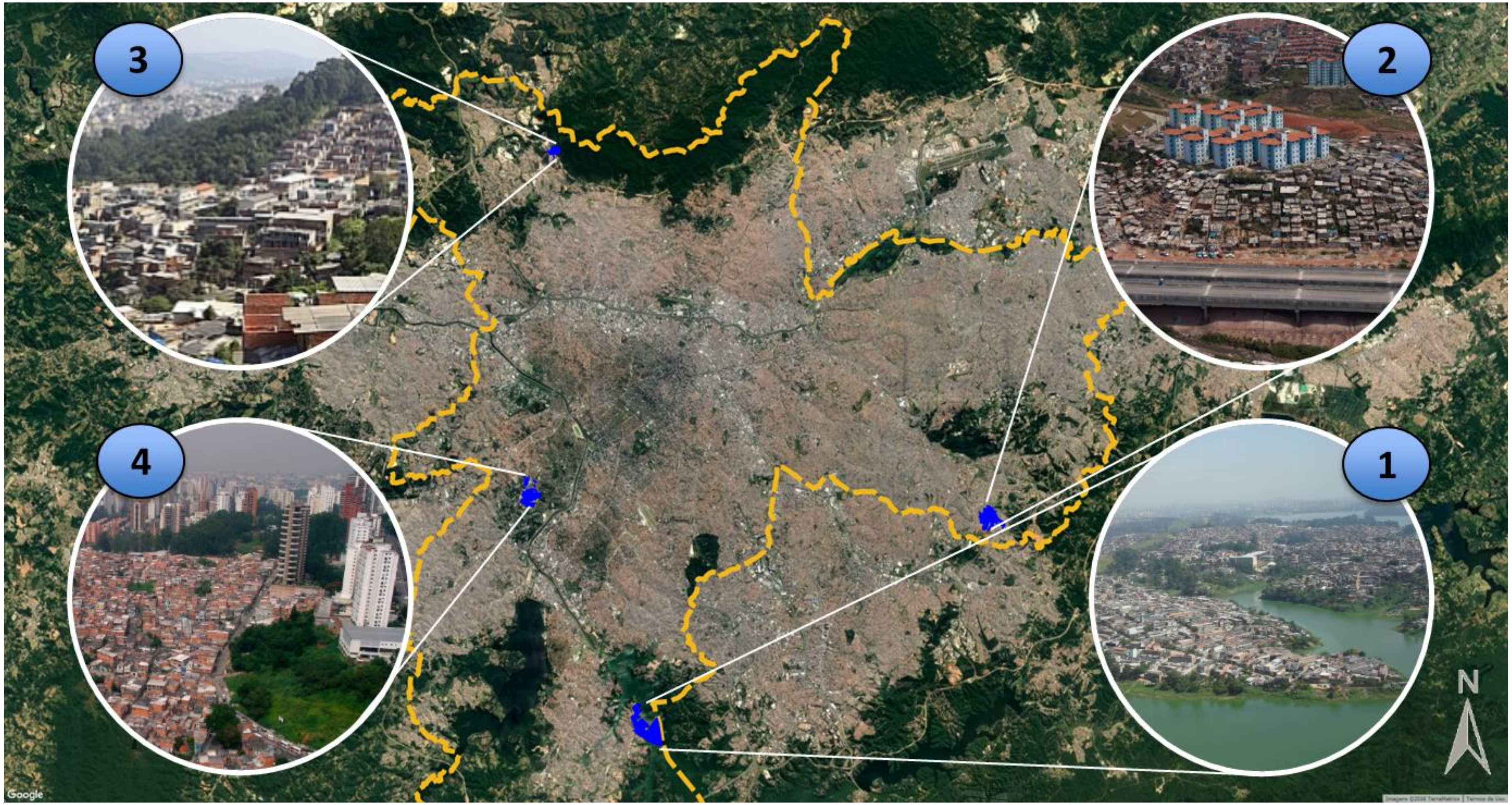
In São Paulo, the transit user swipes card only when boarding public transport. The data obtained include:

- Smart card data throughout the public transportation system, including subway, train and city bus stops;
- Global Positioning System (GPS) data containing the location of the buses;
- Public transport network structure of the region in GTFS (General Transit Feed System) format, which contains information such as bus stops, stations location, frequencies and routes of buses, trains and metro lines.

For each transaction, the main following attributes are available: date and time of the boarding transaction; anonymized card ID and type; route number and direction; and boarded vehicle ID. All the data treatment was developed as from these initial attributes.

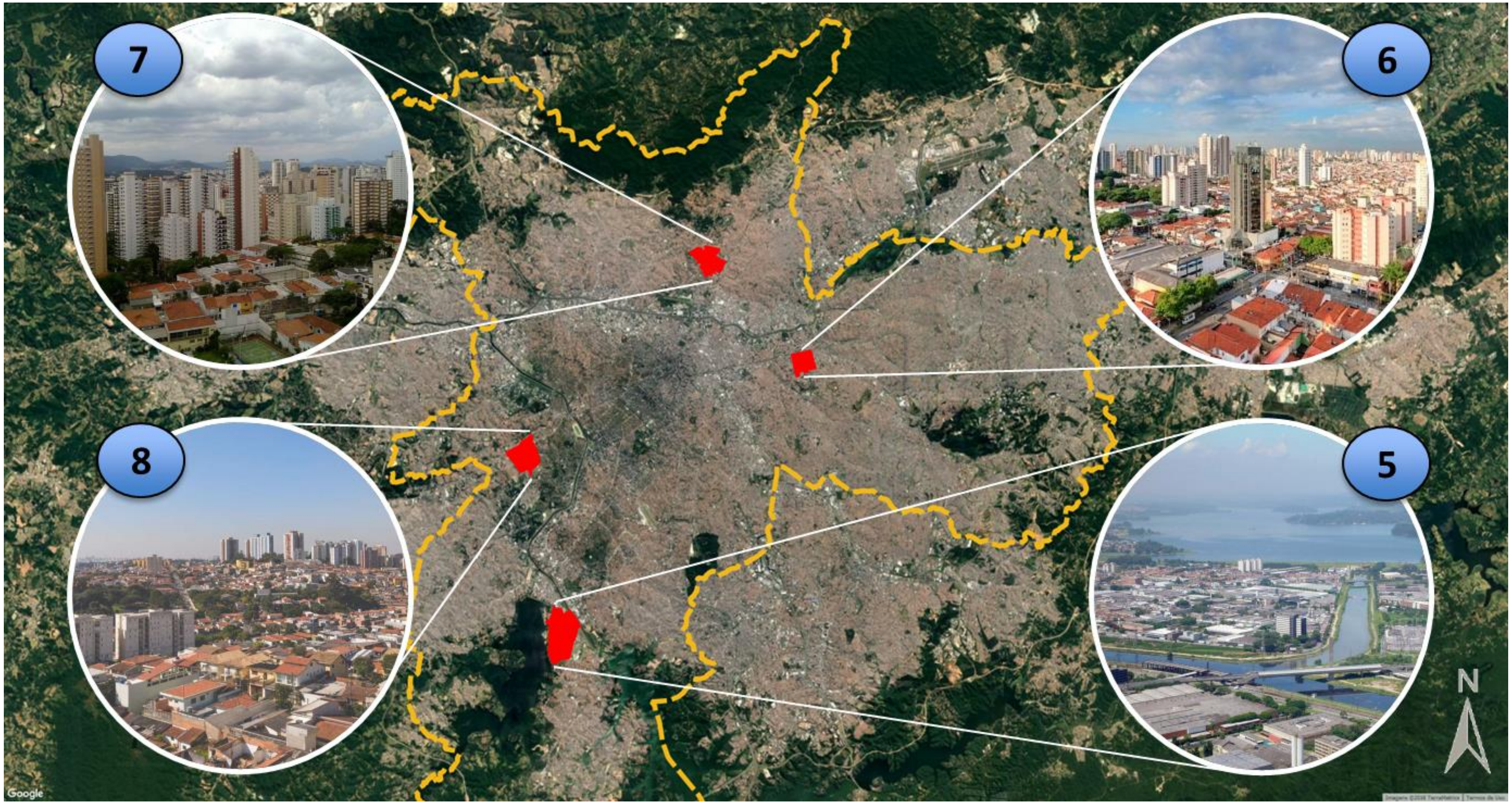
Figure 3 presents the geographic location of the precarious settlements and Figure 4 presents the geographic location of the privileged areas we will approach. Later, a brief description of each area is presented.

Figure 3 – Geographic location of the precarious settlements



Source: Google Earth

Figure 4 – Geographic location of the privileged areas



Source: Google Earth

1. The Cantinho do Céu Complex - which includes the neighborhood of Residencial dos Lagos, Cantinho do Céu and Jardim Gaivotas - is located on the southern side of São Paulo, in the district of Grajaú, on the banks of the Billings reservoir. The community, of about 10,000 houses and 35,000 inhabitants, occupies an area of about 1,500,000 m² with precarious housing and lack of basic infrastructure. The area is located close to environmentally protected areas and far from central commercial areas.

2. São Francisco Global is located on the extreme eastern side of São Paulo, in the district of São Mateus. With about 47,000 inhabitants, the total area of São Francisco Global is estimated to be 1,700,000 m², according to Sehab (Secretariat of Housing). São Francisco Global is also geographically located in an area of limited job supply.

3. Parque Taipas is a small neighborhood located in the northwestern region of São Paulo. The region is inserted in Parada de Taipas neighborhood, the northernmost quarter of the city of São Paulo, located in the Vale do Rio Juqueri and the Serra da Cantareira. According to the municipality of São Paulo, Parque Taipas has about 1,500 houses and 15,000 inhabitants.

4. Paraisópolis belongs to the district of Vila Andrade, in the southern region of the city of São Paulo. Paraisópolis is one of the largest favelas of São Paulo in terms of inhabitants, with more than 57,000 inhabitants living in about 17,000 houses in an extension of almost 1,000 km², according to the last Brazilian census, conducted in 2010 by IBGE.

5. Parque Interlagos is a middle-class neighborhood located in the southern region of São Paulo, in the district of Socorro, and is located between two reservoirs: Guarapiranga and Billings. The Autodrome of Interlagos (which holds Formula 1 races), as well as the Jurubatuba train station, in the CPTM emerald line, are located in Parque Interlagos.

6. Vila Gomes Cardim is a middle-class district of São Paulo, located in the district of Tatuapé, in the eastern part of the city of São Paulo. The area serves as a regional center, with a large concentration of leisure establishments, as well as banks, commercial establishments, a private university and a shopping mall. Carrão metro station, in the red line, is located in Vila Gomes Cardim.

7. Jardim São Paulo is one of the most valued areas of the northern region of the city, located in the district of Santana. The region features business suites, leisure facilities. There is a blue line metro station, Jardim São Paulo-Ayrton Senna, located in Jardim São Paulo.

8. Vila Sônia is a district located in the western region of the city, with about 9.9 km² and 90,000 inhabitants. The region will hold the end of the yellow line of São Paulo metro system.

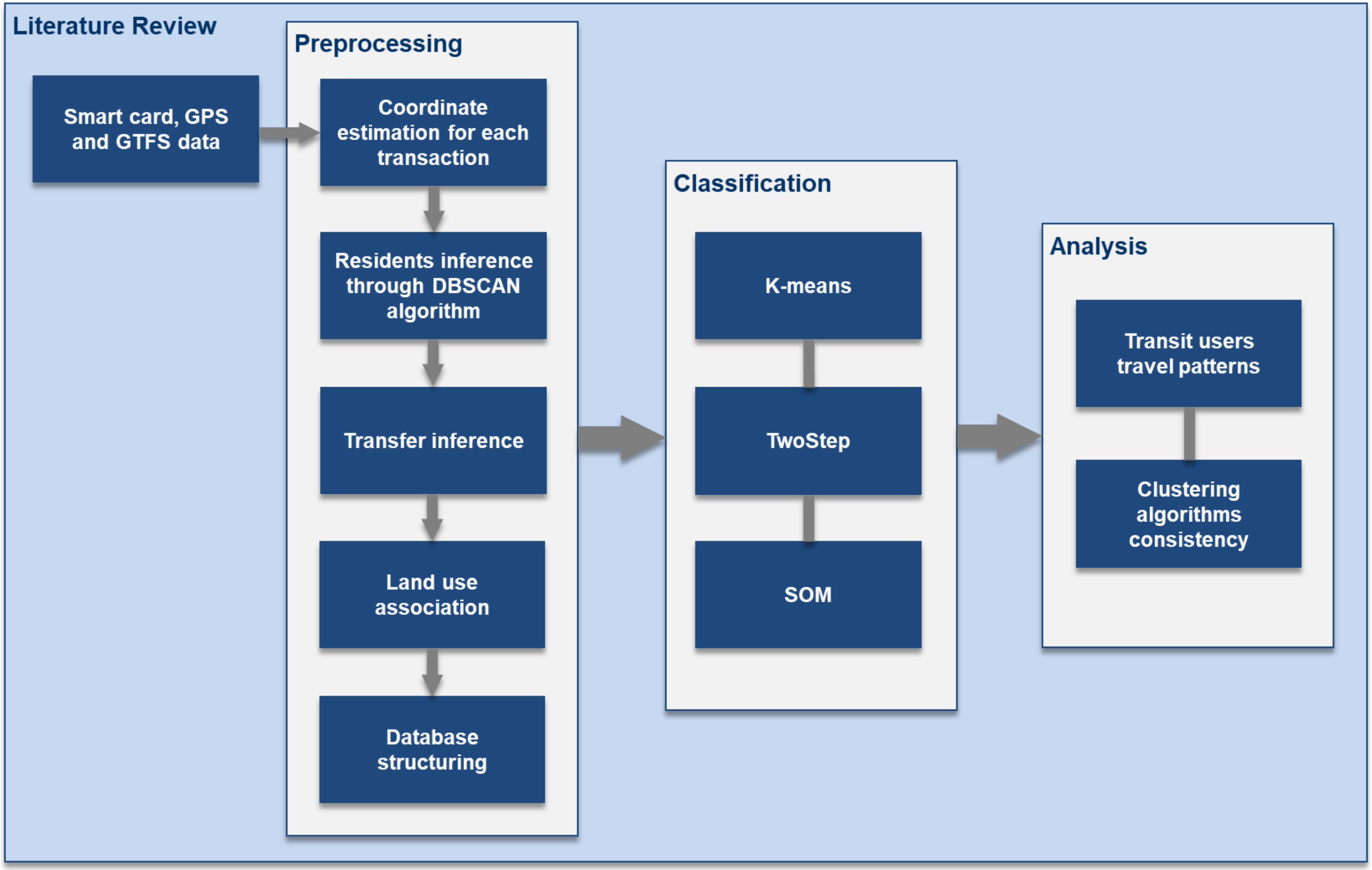
Paraisópolis has the particularity of not being located in a peripheral area of the city, as would normally be the case of precarious settlements – and the case of the other regions studied here. Despite the condition of precarious settlement, Paraisópolis is next to Morumbi, one of Sao Paulo's noblest neighborhoods, and is close to also privileged commercial areas of the city, such as Vila Olímpia, Itaim Bibi and Brooklin neighborhoods.

4 METHODOLOGY

This chapter presents the methods used here to evaluate spatiotemporal travel patterns of urban public transport users from precarious settlements, and compare them with those of other areas in the city of São Paulo. This sequence of methods was based on the literature reviewed in Chapter 2 and on the data available.

The methods for identifying different passenger groups with similar profiles comprises the following steps: preprocessing; application of different classification methods; and analysis. Each step will be better detailed in the following sections; Figure 5 presents a diagram of the sequence of processes developed. All the data treatment and developments were made using *QGIS 2.18.10* software, scripts in *R* and *Python* codes and *IBM SPSS Statistics* software. Each will be mentioned when used during the methods description.

Figure 5 – Method's sequence


Literature Review
Smart card, GPS and GTFS data
Preprocessing
Coordinate estimation for each transaction
Residents inference through DBSCAN algorithm
Transfer inference
Land use association
Database structuring
Classification
K-means
TwoStep
SOM
Analysis
Transit users travel patterns
Clustering algorithms consistency

Source: The authors' own elaboration

4.1 PREPROCESSING METHOD

This section presents the previous manipulation of the smart card data prior to structuring it for the clustering purpose, aiming both to enrich information to the data and to filter potential unwanted information or observations that could underperform clustering techniques.

4.1.1 Transactions Spatial Estimation

The smart card dataset obtained does not contain any location information. Therefore, the first stage of the preprocessing consists of estimating the position of each recorded smart card transaction using GPS data. The estimation criteria were made based on the work by Arbex and Cunha (2017). The method used to locate boarding at the rail system is different from the one for boarding buses.

For the rail system (metro and trains), an association of all boarding registered on ticket gates to their respective stations is previously made through an auxiliary table. Afterward, the procedure for estimating the spatial location of boarding is made by associating the station code with its corresponding location in the GTFS *stops.txt* file, which contains all the station locations.

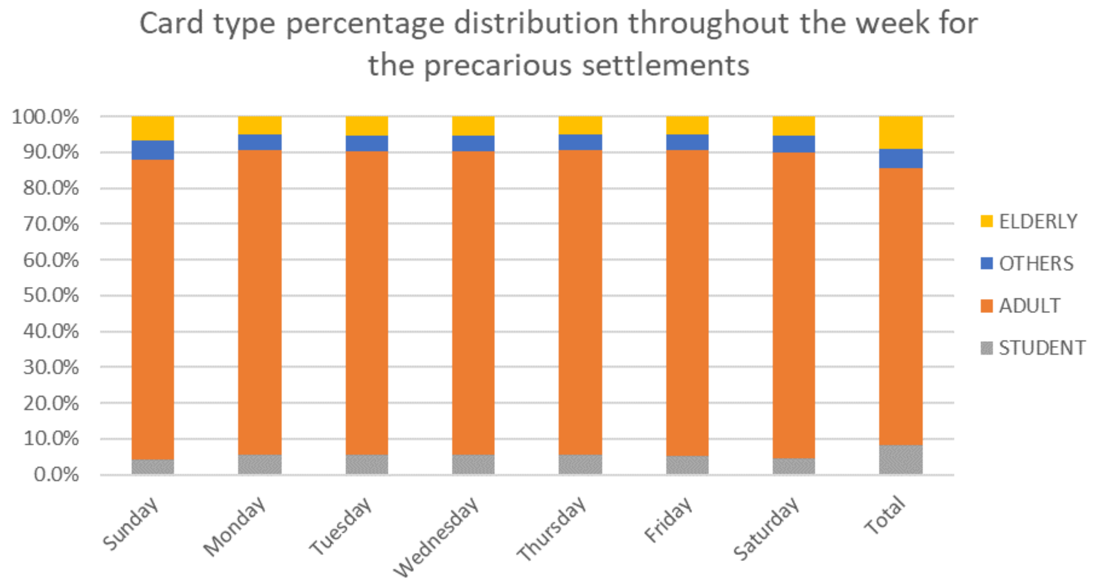
For buses, this procedure is made by filtering all smart card transactions registered for each vehicle on the database, for each day. This assures that only transactions made in a specific vehicle on a specific day of analysis are being considered for the estimation. The exact time of these smart card transactions is then crossed with the registers of all the GPS records of the given vehicle on the given day, obtaining, for each smart card transaction, the corresponding shortest time gap GPS record of that vehicle. This process is performed for all the vehicles of the database and for all the 77 days that SPTrans made available to the Paraisópolis Project financed by the World Bank, which this work is related.

4.1.2 Frequent Passenger Selection

For recognizing users' travel patterns throughout time, it is necessary to filter out occasional users. This step means to enhance clustering results by removing passengers for which the number of observed trips is insufficient to reveal travel patterns (EL MAHRSI et al., 2014). Zhao et al. (2017) filtered out passengers with less than eight days taking metro over one month of analysis. El Mahrsi et al. (2014) filtered based on the condition of ten days of smart card usage over the 25-day period. We here used the criteria of at least one validation made in each of the 11 weeks of data (frequency of once a week). Also, only business days were considered for the evaluation.

In order to facilitate the ongoing analysis, the 50 existing types of smart cards originally in the database were grouped into only 4 aggregated types: “Adult”, “Student”, “Elderly” and “Others”. As the elderly cards' holders are not obliged to swipe their cards on the ticket gate in São Paulo (the elderly can stay in the front part of the bus, without the need to go through the ticket gate), it is difficult to make any statements about their travel patterns. The same difficulty can be found in “Others”, which includes a wide variety of types of cards. Nonetheless, these users represent 17% of the total number of cardholders and their validations represent less than 10% of the total validations on business days (Figure 6). Therefore, these two types were filtered out of the database.

Figure 6 – Card type distribution throughout the week for the precarious settlements



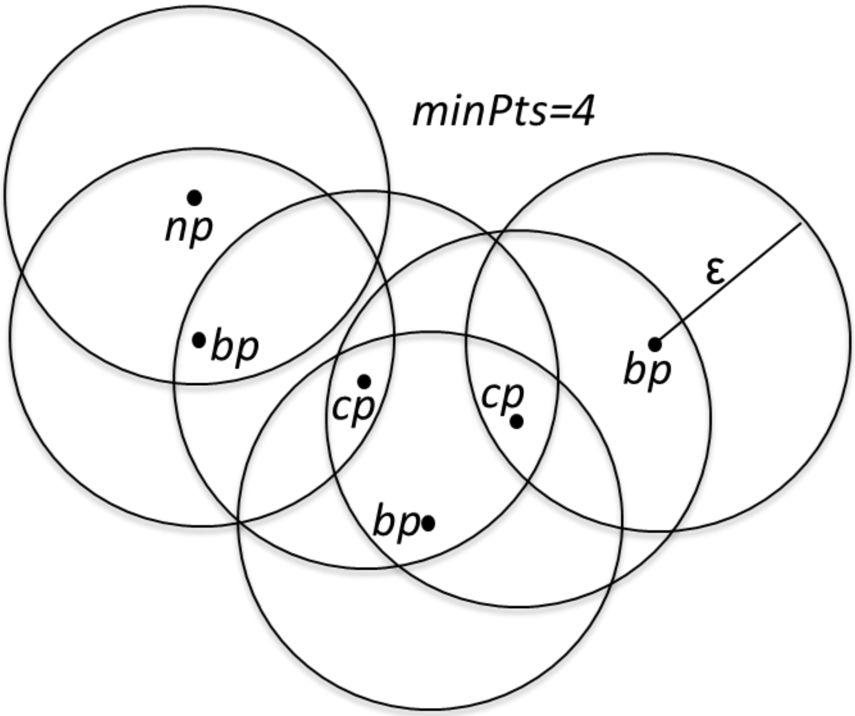
Source: The authors' own elaboration

4.1.3 Residents' Inference through the DBSCAN algorithm

After filtering out occasional users and card types, it is necessary to select, from the total smart card database, only validations made by inhabitants of the regions of study – Cantinho do Céu, São Francisco Global, Parque Taipas and Paraisópolis for the precarious settlements; and Parque Interlagos, Vila Gomes Cardim, Jardim São Paulo and Vila Sônia for the additional areas. To infer the residents of these areas, a DBSCAN clustering algorithm is applied.

The DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise) is based on the identification of clusters by the density of points in space, and was originally proposed by Ester et al. (1996). Basically, points are clustered by high-density regions, while points located in low-density regions are characterized as outliers. Therefore, for identifying regions of high and low density, the DBSCAN algorithm requires two parameters as input data for its processing: distance ε and the minimum number of points for defining a cluster, *minPts*. The Euclidean distance is the most common distance metric used. A given point is classified as a core point if there are at least *minPts* points less than ε distant from it. It is classified as a border point if it is less than ε distant from a core point but does not satisfy *minPts.* A noise point is neither a core nor a border point: *cp* are core points, *bp* are border points and *np* is a noise point (Figure 7).

Figure 7 – Example of DBSCAN definition of points



Source: Kieu et al. (2015)

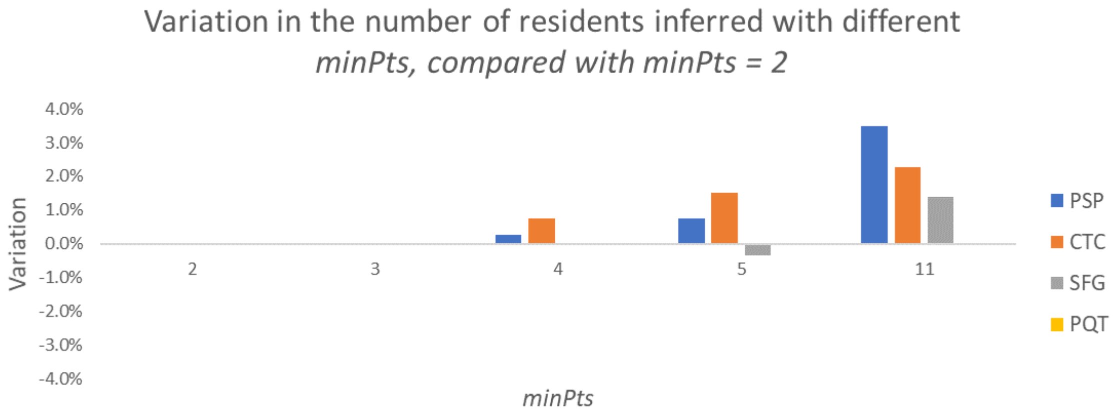
For each cardholder, all the first validations of each day are clustered by the DBSCAN algorithm. Considering both a maximum reasonable walking distance to reach a bus stop and the gap between boarding the public transport and swiping the smart card on the ticket gate, the values applied to the parameters are Ɛ = 1km and *minPts* = 2. The centroid of the largest cluster (with the highest number of validations) is then determined and, if this centroid is geographically located inside one of the study areas, the user is considered a resident of the corresponding one.

Parameter *minPts* is also tested for the precarious settlements with the criteria of 3, 4, 5 and 11 points, but no significant changes in the number of residents inferred are detected, as Figure 8 shows (maximum variation of 3.5%). From now on, for simplification purposes, some figures and tables will have the names of the study precarious settlements and middle-class areas represented by acronyms as follows:

- Paraisópolis – PSP;
- Cantinho do Céu – CTC;

- São Francisco Global – SFG;
- Parque Taipas – PQT;
- Vila Sônia – VLS;
- Parque Interlagos – PQI;
- Vila Gomes Cardim – VGC;
- Jardim São Paulo – JSP.

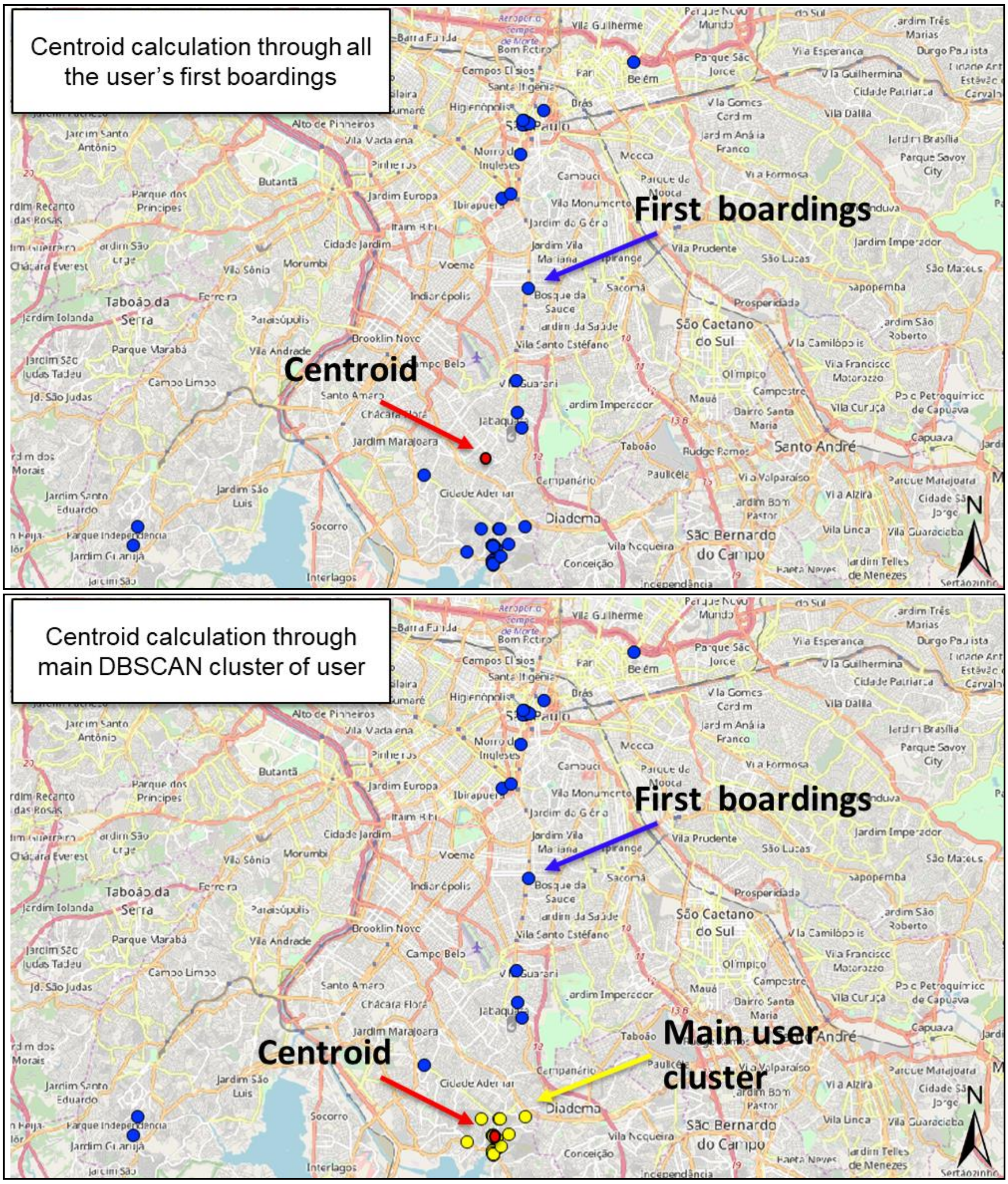
Figure 8 – Variation in the number of residents inferred with different *minPts*



Source: The authors' own elaboration

One may question why this centroid was not calculated for all the first boarding of each user, instead of only considering the main cluster of DBSCAN. To illustrate this issue, Figure 9 presents a real example of a random user. The first attempt of calculation considers all the first boarding of the user, while the second considers only the main DBSCAN cluster. This specific user had 66 days with transactions in the database. From these days, 44 first boarding transactions were classified in the main cluster formed by DBSCAN. It seems logical that the centroid should be located within this area, but that is not the case if all the first validations are considered.

Figure 9 – Centroid calculation for all first validations and only main DBSCAN cluster


Centroid calculation through all the user's first boardings
First boardings
Centroid
Centroid calculation through main DBSCAN cluster of user
First boardings
Centroid
Main user cluster
N

Source: The authors' own elaboration

DBSCAN provides a flexible solution to spatial travel pattern analysis of individual passengers (KIEU et al., 2015; MA et al., 2013). As we aim to consider only the centroid of the main cluster of validations for each user, DBSCAN meets the needs of this operation. Furthermore, DBSCAN identifies a cluster of any shape and sizes and does not require predetermining an initial number of clusters. Other clustering methods could require the number of clusters input, which could define over or undersized clusters to meet the predetermined criteria.

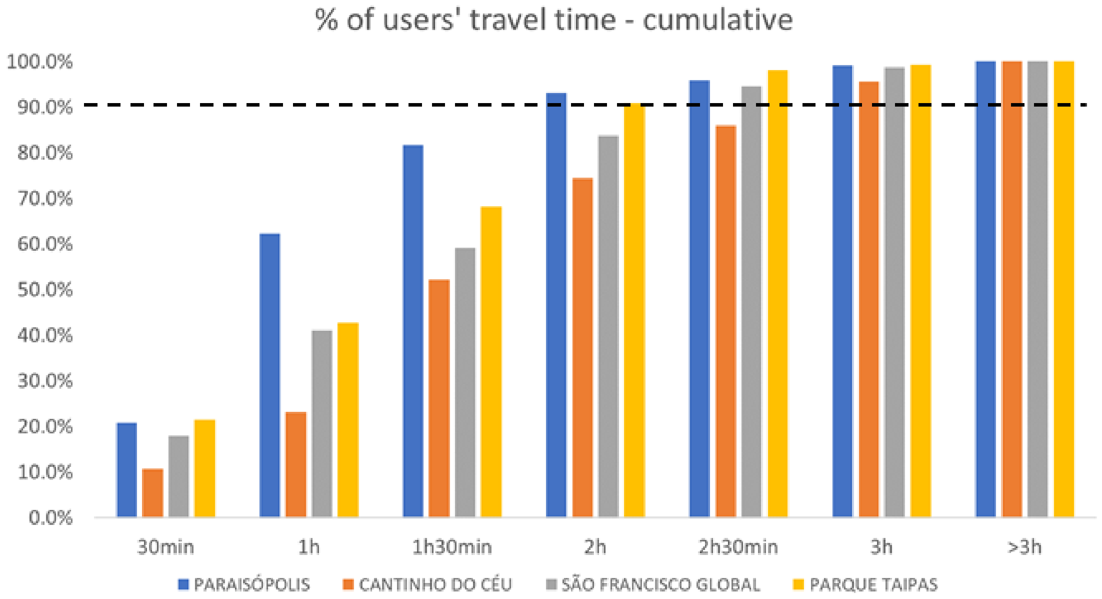
4.1.4 Identification of Activities and Transfers

The next step of the method regards activities and transfers. Considering that transfers are not activities themselves, but necessary, in some cases, for the user to reach their final destination and to access the activity that motivated their trip (DEVILLAINE; MUNIZAGA; TRÉPANIER, 2012), it is important to identify in smart card database what is considered an activity and what is a transfer, since it is not reasonable to evaluate travel patterns of users solely considering transfer validations.

Based on the criteria developed in Devillaine, Munizaga and Trépanier (2012), every time window between boarding transactions of less than a time threshold (in their case, 2 hours) of a specific user is flagged as a transfer, unless these consecutive transactions are made on two buses on the same route. For determining this threshold, the duration of trips of the study areas' residents is calculated based on the 2007 Origin-Destination Survey data from the São Paulo Metro (MA et al., 2013). Only bus, micro-bus / van, metro and train modes of the city of São Paulo are used since these are the modes captured by the database from SPTrans. Also, only the data related to the areas under study are filtered: Paraisópolis (zone 299), Cantinho do Céu (zone 276), San Francisco Global (zone 220), Parque Taipas (zone 115), Vila Sônia (zone 308), Parque Interlagos (zone 270), Vila Gomes Cardim (zone 163) and Jardim São Paulo (zone 133).

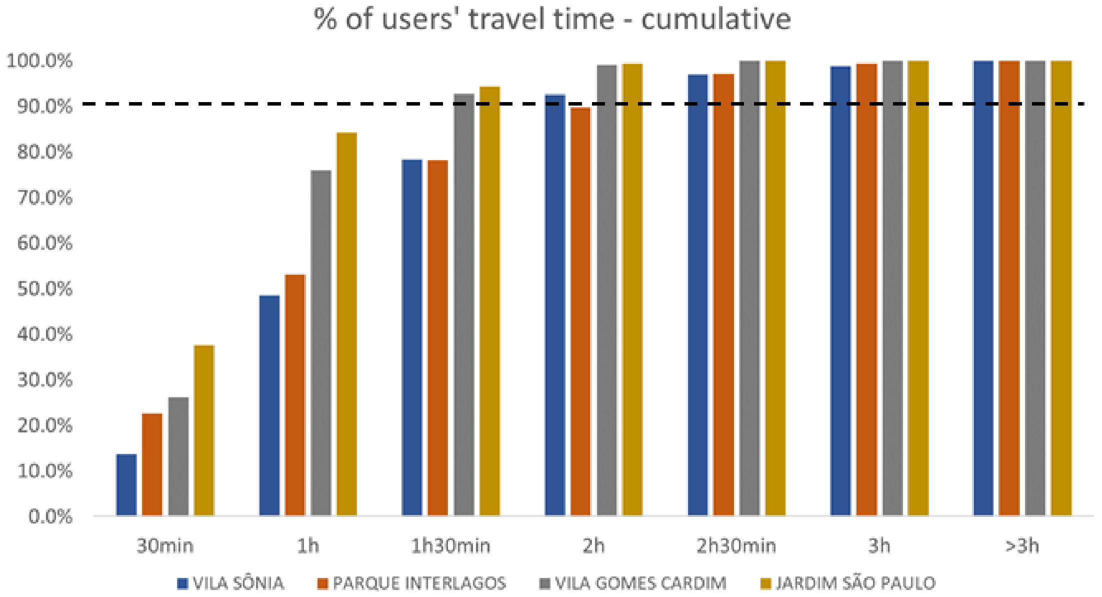
The time threshold is determined when the cumulative percentage of residents reached 90% (i.e. 90% of the residents of a given area took less than the time threshold for the whole trip to their activity). The results for the precarious settlements are 2h for Paraisópolis and Parque Taipas; 2.5h for São Francisco Global; and 3h for Cantinho do Céu (Figure 10). For the middle-class areas, the thresholds are slightly lower, which seems reasonable since these regions are provided with more transit infrastructure, with 2h for Vila Sônia, 2.5h for Parque Interlagos, 1.5h for Vila Gomes Cardim and Jardim São Paulo (Figure 11). Only trips inferred as activities are considered in the following stages.

Figure 10 – Duration of a trip to activity – 2007 Origin-Destination Survey – São Paulo Metro – Precarious Settlements



Source: The authors' own elaboration

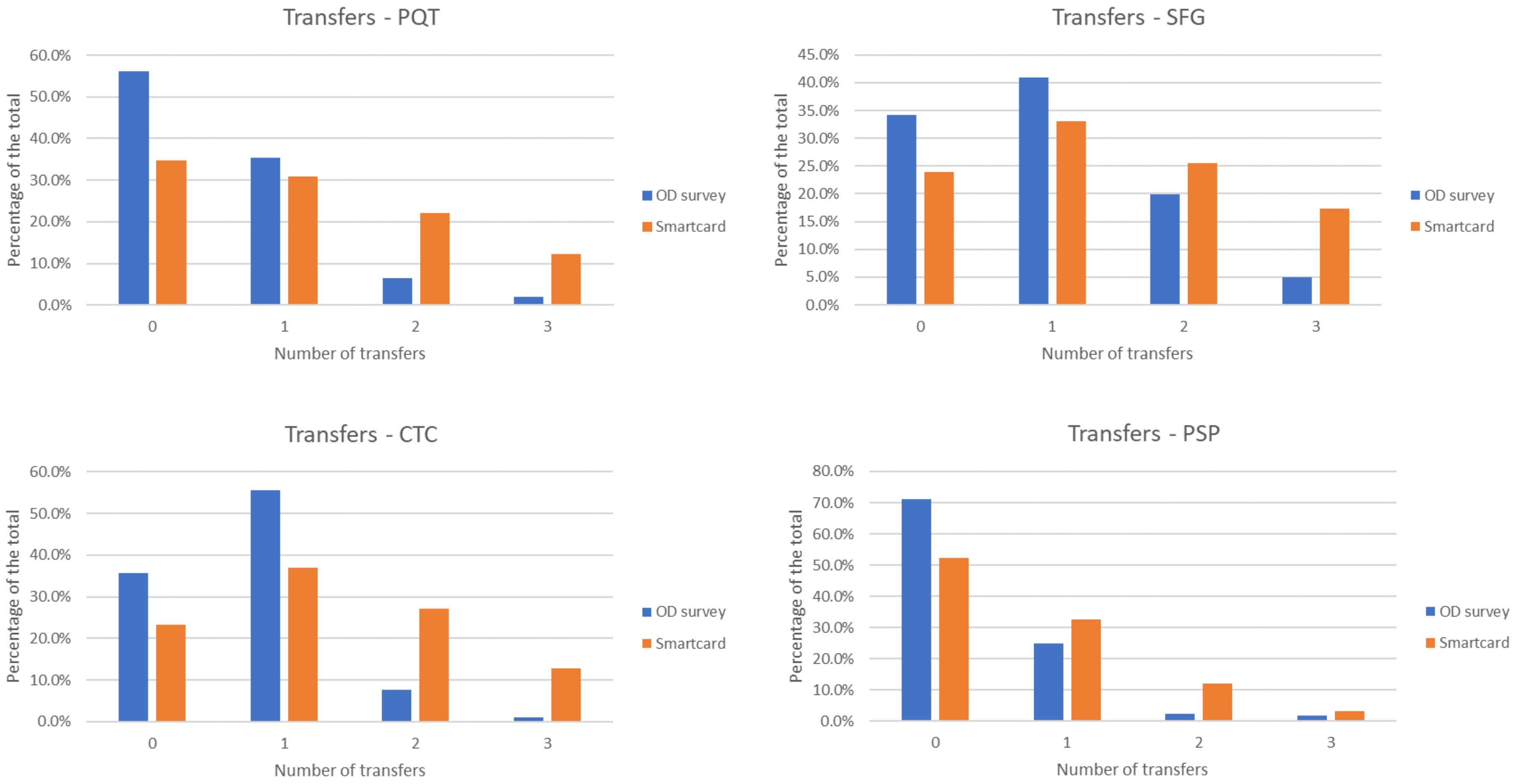
Figure 11 – Duration of a trip to activity – 2007 Origin-Destination Survey – São Paulo Metro – Middle-class Areas



Source: The authors' own elaboration

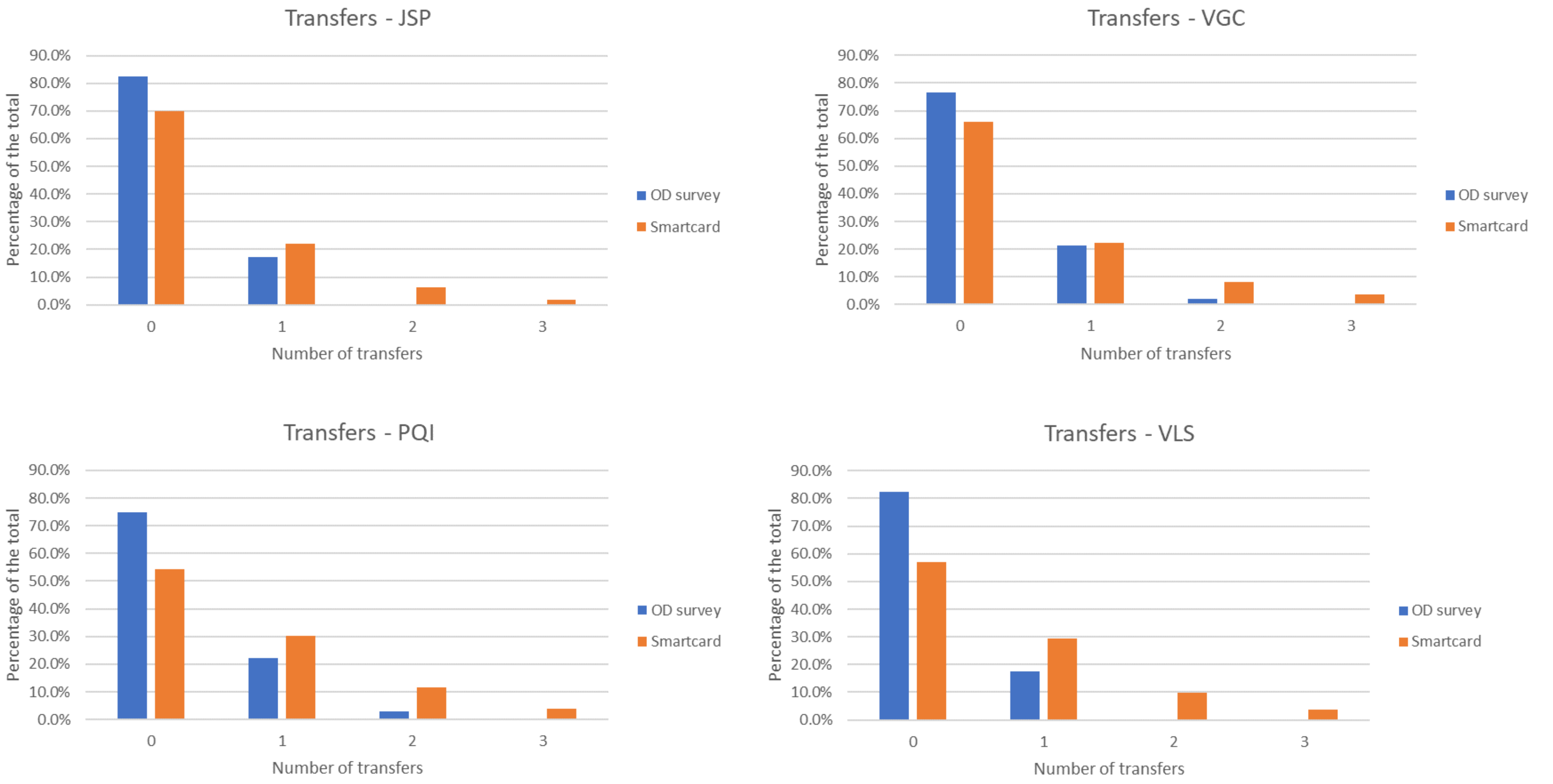
Aiming to validate the method of inference of trips and transfers, the proportion of residents' number of transfers per region obtained from the smart card data is compared with the ones from the OD survey. Figure 12 and Figure 13 show this comparison, and a reasonable similarity can be observed.

Figure 12 – Comparison of proportion of residents' number of transfers with the 2007 OD Survey – São Paulo Metro – Precarious Settlements



Source: The authors' own elaboration

Figure 13 – Comparison of proportion of residents' number of transfers with the 2007 OD Survey – São Paulo Metro – Middle-class Areas



Source: The authors' own elaboration

With the processes abovementioned developed, a total of 24,310 transit-user residents were inferred from cardholders along the 11 weeks, with 2,332,639 validations excluding transfers. The number for each studied area is:

- Paraisopolis: 4,906 transit-user residents; 461,668 validations excluding transfers;
- Parque Taipas: 280 transit-user residents; 28,247 validations excluding transfers;
- São Francisco Global: 1,354 transit-user residents; 129,008 validations excluding transfers;
- Cantinho do Ceu: 2,974 transit-user residents; 288,241 validations excluding transfers.
- Vila Sônia: 1,916 transit-user residents; 178,666 validations excluding transfers;
- Jardim São Paulo: 5,069 transit-user residents; 493,489 validations excluding transfers;
- Vila Gomes Cardim: 5,713 transit-user residents; 555,440 validations excluding transfers;
- Parque Interlagos: 2,098 transit-user residents; 197,880 validations excluding transfers.

4.1.5 Database Structure for Travel Pattern Clustering

How the processed smart card database aforementioned is structured for the three different clustering methods is presented below. It is a default structure (to evaluate the difference between results) inspired by the works by Agard et al. (2006), Agard et al. (2013), El Mahrsi et al. (2014), Morency et al. (2006), Morency et al. (2007), Ortega-Tong (2013), Lathia et al. (2013) and Zhao et al. (2017) that, with some differences in the definition of variables, all structured their datasets on passengers' profiles, described by the distribution of all their validations over the period of time of analysis from the dataset and/or specific variables that together characterize each passenger's travel routines. The structure defined for each study was used as the clustering variables to determine a passenger segmentation for a specific group. Each row of the database represents a transit user, with its features throughout the 11 weeks of analysis.

The aim of this clustering structure is to discover the travel patterns of users, helping to identify travel patterns in the way passengers use transit and characterize the demand accordingly (EL MAHRSI et al., 2014). Therefore, firstly, a set of nine descriptive variables are defined and, based on Ortega-Tong (2013), divided into four categories. The first describes temporal pattern and variability, the second captures spatial pattern and variability, the third describes activity pattern variability and the last captures socioeconomic characteristics:

- Temporal Pattern and Variability
    - Start hour of travel: the median for the first hour of the user's transactions (the median between the first hour of all days of transaction, per user);
    - Start hour of travel dispersion: The standard deviation for the first hour of the user's transactions;
    - Weekly travel frequency: The median for the weekly frequency of travel per user (the median between the frequency of travels of each week, per user);
    - Weekly travel frequency dispersion: The standard deviation for the weekly frequency of travel per user;

- Spatial Pattern and Variability
    - Daily distance: The median of maximum daily Euclidean distance (the median between the maximum reached distance of each day, per user);
    - Daily distance dispersion: The standard deviation for the median of the maximum daily Euclidean distance;

- Activity Pattern and Variability
    - Daily activity duration: The median of the maximum daily activity duration (the median between the first and the last transaction of each day, per user);

    - Daily activity duration dispersion: The standard deviation for the median of maximum daily activity duration;
- Socioeconomic Characteristic
    - User’s household income: The income per users’ household (calculated by the income data from the 2010 census tracts data from The National Bureau of Statistics).

Additionally, contextual information is introduced regarding the land use where each user is boarding. The land use information by fiscal blocks is obtained by the municipality of São Paulo and available to this research. From the sixteen types of land use of the municipality original information, a grouping was made to reduce it to five types. The groups were classified as: Residential Low Income; Residential Medium/High Income; Commercial/Services/Industrial; Residential/Commercial; and Other. A visual inspection of the land use was made through satellite images in the surroundings of the study areas, and a few land uses were updated. Figure 14 presents the final land use map.

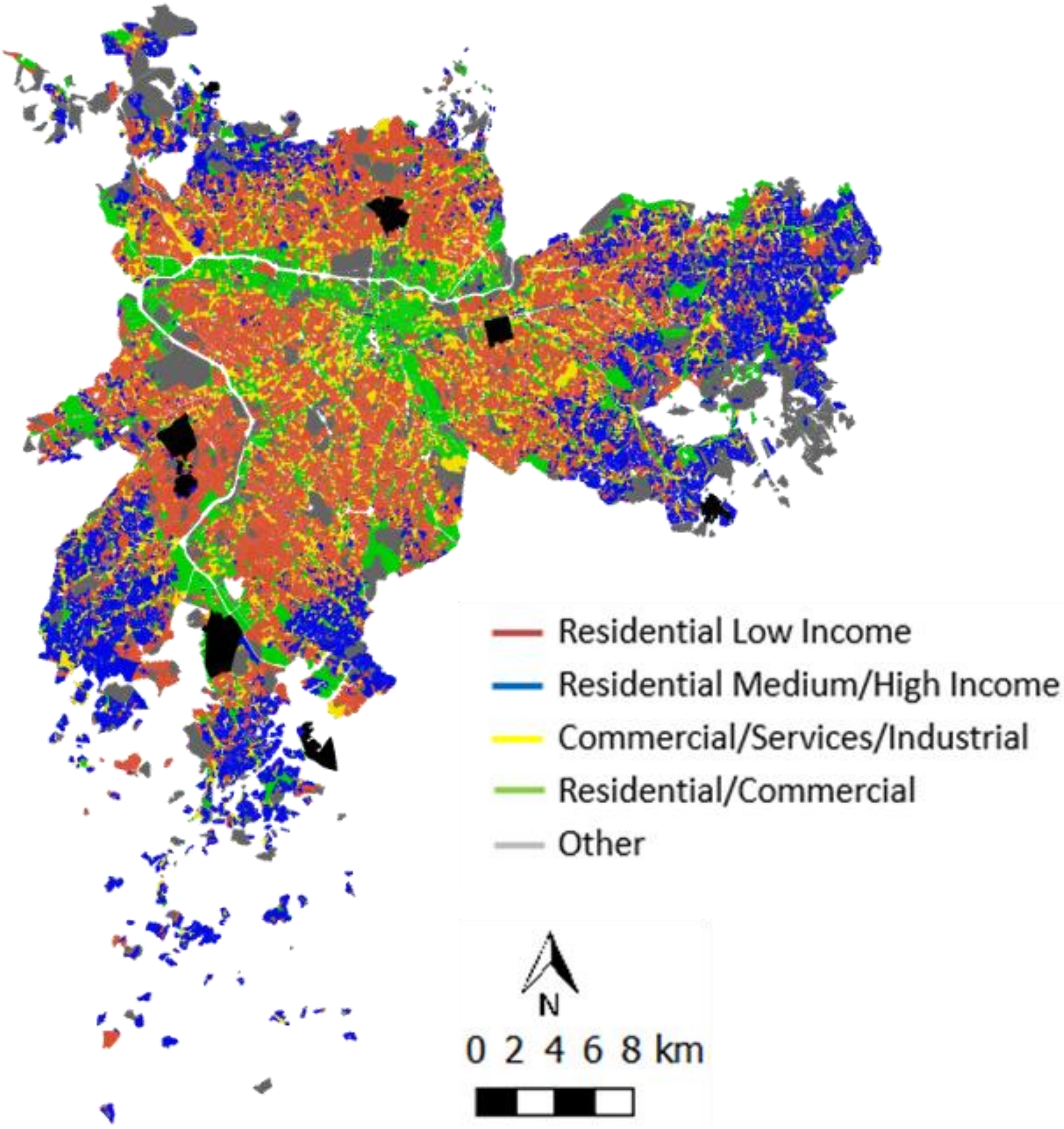
The idea is to associate each transaction in the database with a land use type, according to the land use in which the transaction is geographically located and, for each user, calculate the proportion of validations in each land use. For calculating these land use variables, the first boarding validations of the day for each user are not used to compute this proportionality, considering that the first validation is assumed to be the residence of the users, and the focus here would be analyzing their activities.

Therefore, in this structure, five additional variables are created for the database, representing the proportion of validations in each of the five types of land use considered for the study:

    - Residential Low-Income land use; the proportion of boarding validations in “residential low-income” land use over all boarding validations of each user;
    - Residential Medium/High-Income land use; the proportion of boarding validations in “residential medium/high-income” land use over all boarding validations of each user;

- Commercial/Services/Industrial land use; the proportion of boarding validations in “commercial / services / industrial” land use over all boarding validations of each user;
- Residential/Commercial land use; the proportion of boarding validations in “residential / commercial” land use over all boarding validations of each user;
- Other land uses; the proportion of boarding validations in “other” land uses over all boarding validations of each user.

Figure 14 – Grouped land uses for clustering algorithm



Source: The authors’ own elaboration

Therefore, the final number of attributes to be clustered is fourteen. As presented above, all the variables are numerical, and the structure of the database is the same for all three clustering methods, aiming to standardize the analysis. Table 2 illustrates an example of the clustering structure. For example, user 1000123 (hash_bilhete – the card ID) is from Paraisópolis ("PSP" on Area column). This user usually enters the transit system around 7 am (the median of the first transaction of the day of 7.0), has activities that last around 10 hours (the median of the maximum daily travel time of 10.2), the median distance of his/her activities from his/her residence is 8.6km (the median of the maximum daily distance of 8.6), has a frequency of transit use of five days a week (the median of the weekly frequency of 5.0) and his household income is estimated in BRL 1,435.00. Around 57% of the validations of this user (not considering the first validation of the day) were in a Comercial/Services/Industrial area (0.57 on Com_SerInd column) along the 11 weeks of data. This user has low variability throughout the weeks of analysis, suggesting a commuting pattern from these variables (low standard deviations in all attributes aforementioned).

User 1000214, on the other hand, is from Jardim São Paulo ("JSP" on Area column). This user usually enters the transit system around 6:15 am (the median of the first transaction of the day of 6.2), has activities that last around 9:30 hours (the median of the maximum daily travel time of 9.4), the median distance of his/her activities from his/her residence is 1.7km. Around 54% of the validations of this user (not considering the first validation of the day) were in a Residential High-Income area (0.54 on Res_MedHigInc column) along the 11 weeks of data.

Before performing any clustering algorithm, all the variables are standardized by removing the mean and scaling to unit variance. Centering and scaling happen independently on each variable. Standardization is essential in clustering algorithms with dataset containing diverse types of variables, as they might behave badly depending on the distribution of observations between variables (BARROSO; ARTES, 2003).

Table 2 – Random users' example of the database structure

Hash_ Bilhete	Area	Income	Med_ FirstHour	Std_ FirstHour	Med_ Duration	Std_ Duration	Med_ MaxDist	Std_ MaxDist	Med_ WeekFreq	Std_ WeekFreq	Res_ LowInc	Res_ MedHigInc	Com_ SerInd	Res_ Com	Other
1000123	PSP	1435	7.0	1.8	10.2	1.2	8.6	2.7	5.0	0.5	0.18	0.05	0.57	0.15	0.05
1000214	JSP	5562	6.2	3.6	9.4	3.2	1.7	2.6	5.0	0.9	0.02	0.54	0.23	0.07	0.14
1000216	VLS	2978	6.6	1.1	6.6	1.9	2.3	0.8	4.0	1.4	0.02	0.61	0.08	0.08	0.21
1000247	CTC	1085	9.5	2.8	10.3	2.3	18.3	6.4	5.0	0.7	0.09	0.20	0.15	0.45	0.11
1000248	PSP	1435	4.5	1.8	10.1	3.1	8.7	1.4	3.0	0.5	0.07	0.55	0.07	0.23	0.08
1001029	VLS	4179	12.5	5.4	1.8	6.3	4.2	2.1	5.0	0.0	0.00	0.28	0.42	0.04	0.26
1001052	PSP	1241	12.0	3.6	9.3	2.9	5.6	1.1	4.0	0.5	0.00	0.42	0.27	0.06	0.25
1001145	CTC	1368	14.5	5.0	3.1	5.3	8.2	4.1	4.0	0.7	0.16	0.36	0.16	0.30	0.02
1001252	JSP	8120	6.1	4.3	12.1	5.5	6.9	2.8	5.0	0.9	0.11	0.16	0.12	0.17	0.44
1001346	PQI	4063	6.9	4.7	10.8	5.3	7.9	1.8	5.0	0.3	0.02	0.02	0.06	0.02	0.88
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Source: The authors' own elaboration

4.2 CLUSTERING METHODS

The final database developed regards multiple dimensions, as verified in Section 3, and data mining techniques are applied aiming to search for groups of residents with similar travel patterns and, specifically, for a group of low-income job residents.

Data mining is “the extraction of implicit, previously unknown, and potentially useful information from data” (WITTEN; FRANK, 2005) and can be considered a broader term of an interdisciplinary field composed of statistical analyses, database systems, machine learning, pattern recognition, neural networks, fuzzy systems and other soft computing techniques (VELICKOV; SOLOMATINE, 2000).

Machine Learning, as aforementioned, is one of the various techniques for data mining in data science. It provides the technical basis of data mining, retrieving useful information from data. More specifically, it is used to extract desirable information from the raw data in databases, expressed in a comprehensible form that can be used for a variety of purposes (FRIEDMAN; HASTIE; TIBSHIRANI, 2001; JAMES et al, 2013; WITTEN; FRANK, 2005). Roughly, machine learning can be classified into two learning categories: supervised or unsupervised. This section presents a brief overview of machine learning techniques and an explanation of the clustering algorithms chosen.

Within machine learning techniques, regression and classification – for supervised learning –, as well as clustering and dimensionality reduction – for unsupervised learning –, are some of the most common (JOSEPH et al., 2016). Supervised classification and unsupervised clustering are both classification methods that include several algorithms aiming to group observations based on similar qualitative or quantitative characteristics (ORTEGA-TONG, 2013).

Classification methods involve several techniques and algorithms aiming to group similar elements based on certain features, be they qualitative or quantitative. Supervised classification is used when the issue is to label newly unlabeled pattern based on a training sample of labeled (pre-classified) patterns.

The given labeled patterns are used to learn the descriptions of classes and are used to label the new elements. An example of the supervised technique is the k-nearest-neighbor prediction rule. (JAIN et al., 1999; XU; WUNSCH II, 2005).

Unlike supervised classification, there are no previously known classes for unsupervised classification, aiming to group a given collection of unlabeled patterns into meaningful clusters based on similarities of the input data, that is, these new groups formed are data-driven (JAIN et al., 1999; XU; WUNSCH II, 2005). Unsupervised classification is known as clustering techniques.

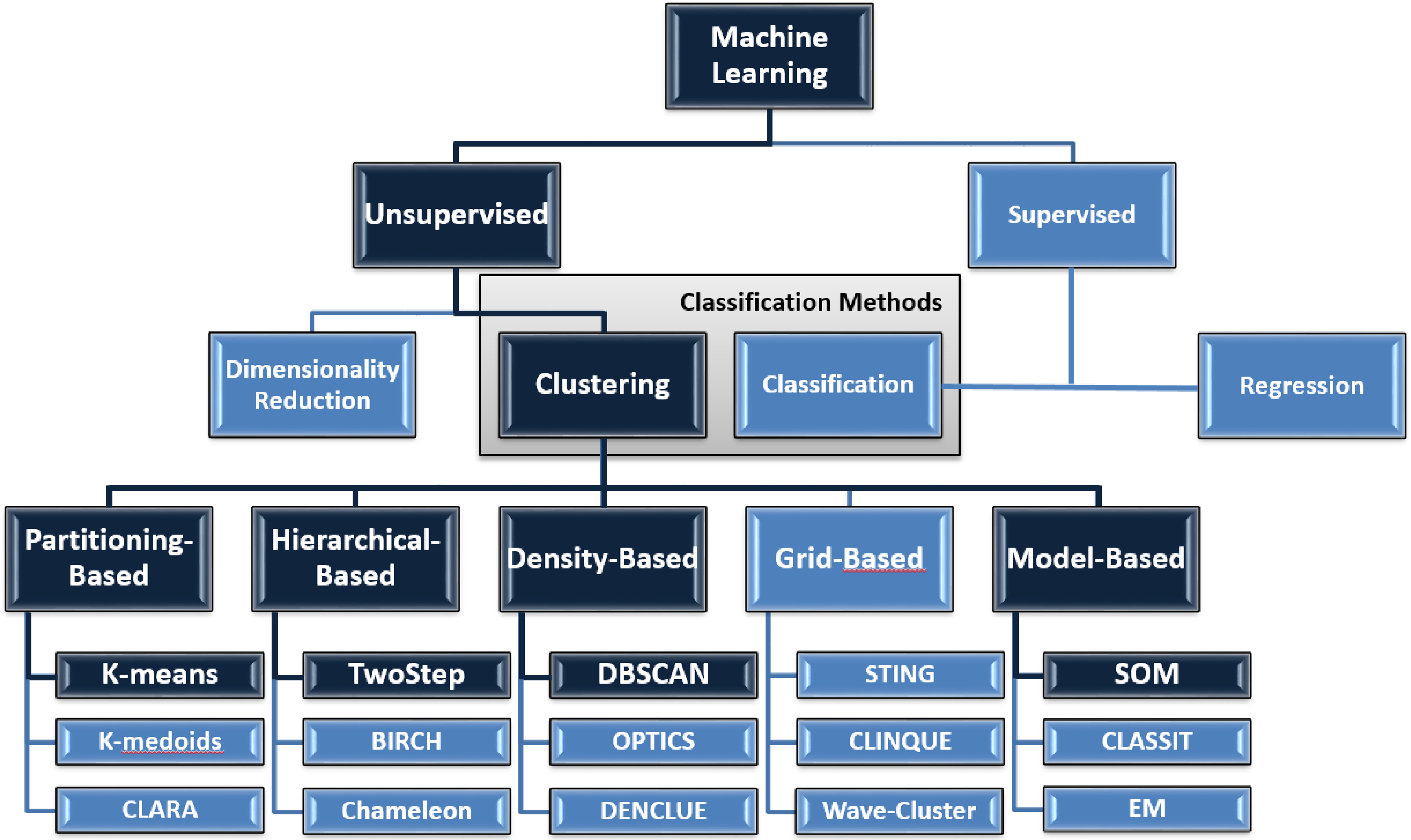
Clustering can be considered as the technique of partitioning a certain base into natural groups called clusters, so that elements within a group are very similar, while elements across clusters are dissimilar (ZAKI; MEIRA, 2013). Clustering techniques can be useful in many “exploratory pattern-analysis, grouping, decision-making, and machine-learning situations, including data mining, document retrieval, image segmentation, and pattern classification” (JAIN et al., 1999). Since smart card database has no previously known information about passenger categories based on their travel patterns, a clustering process – unsupervised classification – needs to be performed to identify different user groups with similar travel patterns of smart card data.

Within clustering techniques, there are also many different types of clustering characteristics and, therefore, many approaches to subdivisions, such as Jain and Dubes (1988), Zaki and Meira (2013) and Fahad et al. (2014). Figure 15 presents a taxonomy of these techniques, with the clustering methods structured as partitioning-based, hierarchical-based, density-based, grid-based and model-based.

Partitioning-based clustering algorithms divide a dataset into a number of predefined partitions, in which each partition represents a cluster. Hierarchically-based clustering creates a hierarchical decomposition of the given dataset, and a dendrogram represents the datasets, whereby individual data is presented by leaf nodes. There are agglomerative (bottom-up) and divisive (top-down) approaches for hierarchically-based clustering. The density-based clustering the dataset is separated based on their regions of density, connectivity, and boundary. These clustering algorithms continue to expand the given cluster as long as the density (number of objects or data points) in the neighborhood exceeds a threshold. Grid-based methods quantize the object space into a finite number of cells that form a grid structure. The model-based clustering optimizes the fit between the given data and a mathematical model. It assumes that the data is generated by a mixture of underlying probability distributions. There are statistical and neural network approaches based on the model-based method (FAHAD et al., 2014).

Four out of the five groups of clustering algorithms are used in this research – painted in dark blue in the taxonomy of Figure 15. The DBSCAN method is used to infer the residents of each area described in Section 3. Here, aiming at identifying different user groups with similar travel patterns, three clustering algorithms are selected: K-means, TwoStep, and SOM. All the three have good performance in handling very large datasets; testing different groups of clustering algorithms (with different assumptions of grouping data objects) can provide even more robustness to the common groups arising from the three methods. Also, a clustering validation will be performed comparing the three clustering algorithms to evaluate the measure performance for each method for this dataset.

Figure 15 – A taxonomy of clustering methods



Source: Adapted from Zaki; Meira (2013) and Fahad et al (2004)

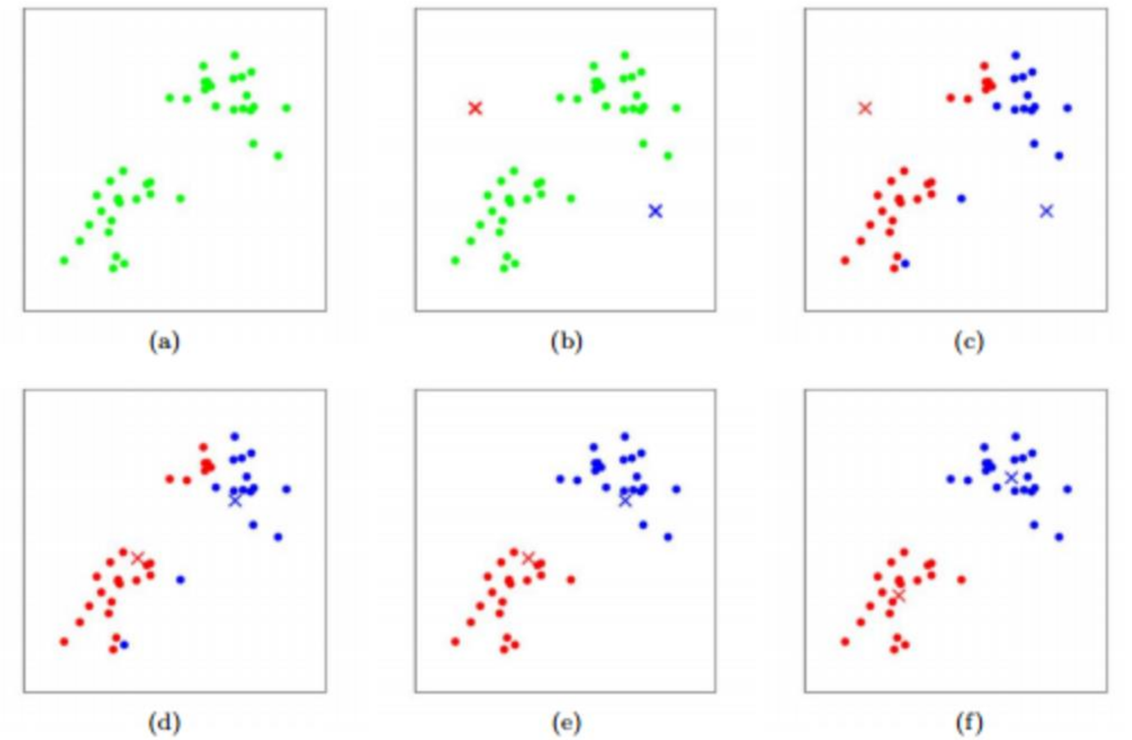
4.2.1 K-means Clustering

The k-means algorithm is very simple and can be easily implemented for solving many practical problems. It tries to partition ‘N’ records into ‘k’ clusters by minimizing the within-cluster sum of squares, i.e., minimizes the sum, over all clusters, of the distance to the centroid of each cluster (MA et al., 2013; MORENCY et al., 2007). The mean of the record values is continuously updated and each element is assigned into the cluster with the closest center until a convergence criterion is satisfied – usually no (or minimal) reassignment of elements to new cluster centers, or minimal decrease in squared error (FORGY, 1965; JAIN et al., 1999).

In the public transport area, different analyses with segmentation choose the k-means clustering method because it supports large sets of data computation and requires few parameters to be fixed. The time complexity of k-means is *O(NkId)*, being *N* the number of users, *k* the number of clusters, *I* the number of iterations and *d* the number of attributes (XU; WUNSCH II, 2005; AGARD et al., 2013). The Euclidean Distance was the metric used to calculate the distance between elements.

Figure 16 is a didactic representation of how the k-means algorithm works. Given a dataset with a specific number of attributes (a), initial cluster centroids are placed randomly, with the number of clusters formerly defined as input (b). Afterward, based on the minimum distance from these centroids, each element of the dataset is inferred to one cluster (c). With these inferences, a new centroid is calculated for each cluster (d), and steps (c) and (d) are iteratively repeated until there is no change in clusters between the elements of the dataset (e; f).

Figure 16 – Graphical sequence of the k-means algorithm



Source: Chris Piech, based on a handout by Andrew Ng (2013) [1]

One disadvantage of the K-means clustering is that it can converge to a local (and not global) optimum. A solution could be a lucky or exhaustive choice of starting points (MORENCY et al., 2007). Therefore, for all k-means processing, the k-means algorithm is run with 10 different random centroid seeds, and the final result is the best output of the 10 consecutive runs in terms of inertia. For each run, a loop of 300 iterations of the k-means algorithm is made for convergence. Also, the k-means clustering algorithm is performed using Python codes.

[1] Available in: <http://stanford.edu/~cpiech/cs221/handouts/kmeans.html>. Accessed Jun. 06, 2018.

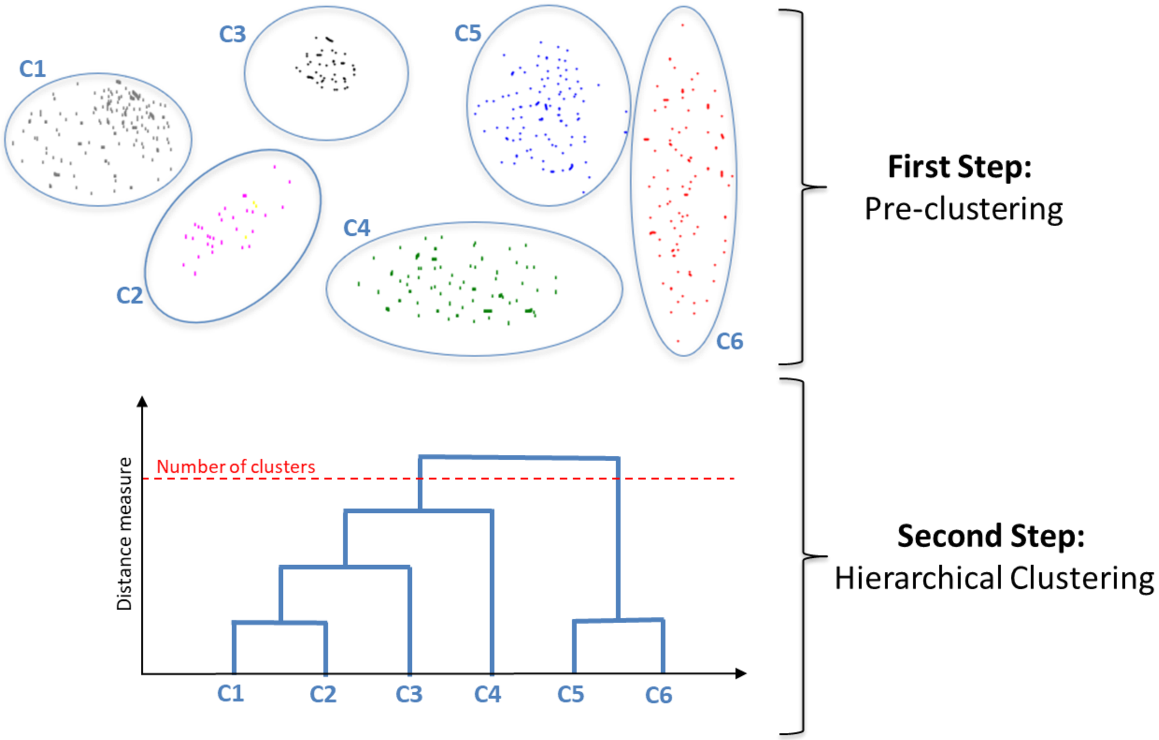
4.2.2 TwoStep Clustering

In the first step of the SPSS TwoStep algorithm, the dense regions of the records are pre-clustered into sub-clusters. The aim of this step is to compute a new data matrix with fewer cases for the second step. In the second step, the sub-clusters resulting from the first phase are again clustered by applying the classical hierarchical clustering algorithm, i.e., the clusters are merged stepwise until all clusters are in one cluster. The hierarchical clustering in the second step is very efficient in this case because the number of dense regions formed in the first phase is far less than the total number of data records in the original dataset (IBM, 2011; CHIU et al., 2001; PITOMBO et al., 2011). Therefore, the TwoStep clustering algorithm works well with large databases.

The TwoStep clustering allows determining how the similarity between two clusters is computed, and the options are Euclidean Distance and Log-Likelihood. The Euclidean distance between two points is clearly defined as a direct distance between the two cluster centers, and can only be applied if all the variables are continuous. A cluster center is defined as the vector of cluster means of each variable. The log-likelihood distance measure is a probability-based distance. The distance between two clusters is related to the decrease in log-likelihood as they are combined into one cluster. Log-likelihood can handle both continuous and categorical variables, and normal distributions are assumed for continuous variables, multinomial distributions for categorical variables and that the variables are independent of each other (IBM, 2011; BACHER et al., 2004). We chose the log-likelihood distance measure.

The time complexity of the SPSS TwoStep clustering is $O(N^2)$, being N the number of users (XU; WUNSCH II, 2005). Figure 17 illustrates the sequence of the Two-Step processes abovementioned. The TwoStep clustering algorithm is performed using the IBM SPSS Statistics Software™.

Figure 17 – Two-Step clustering sequence



Source: The authors' own elaboration

4.2.3 SOM Clustering

The last algorithm applied is the Self-Organizing Map (SOM). It is an unsupervised neural network used to visualize high-dimensional data sets in lower dimensional representations, called a map (KOHONEN, 2001; LYNN, 2014).

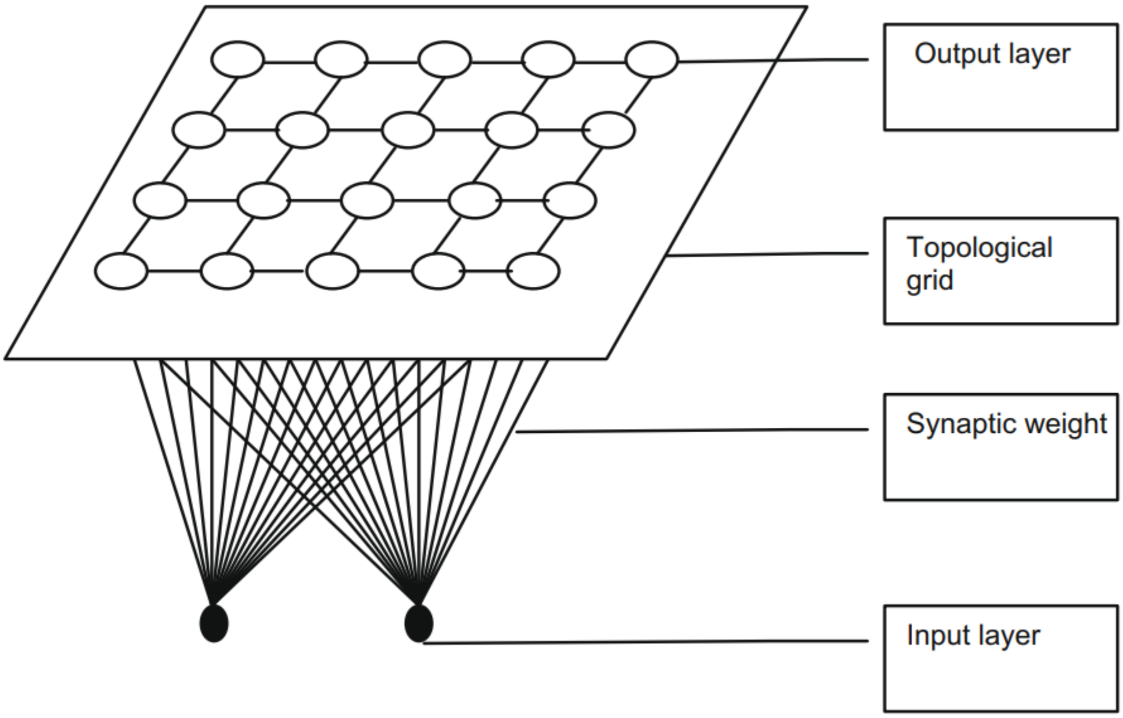
A Self-Organizing Map has the following features:

- The size of the map grid (output layer) is pre-defined.
- Each cell in the grid is assigned an initializing vector in the data space. The 14 attributes to be used here as variables for clustering will be represented in each grid cell (with a 14-dimensional vector). Initiation can either be random or following specific methods; random initiation is used here.
- Output neurons will self-organize to an ordered map and neurons with similar weights are placed together. They are connected to adjacent neurons by a neighborhood relation, dictating the topology of the map (MORENO et al., 2006).

- The network evolves until converging the output map into a representation.
- Users close in the data space are close on the SOM grid map, representing spatial clusters.

Since the SOM preserves the most important topological and metrical relationships of the primary data elements, it can also be used for pattern classification (SILVEN et al., 2003). Figure 18 shows a graphical representation of how SOM works.

Figure 18 – Illustration of SOM dynamics

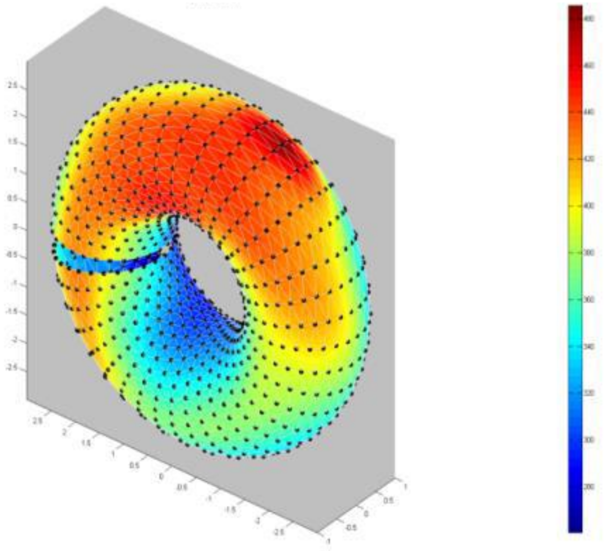


Source: Mostafa (2010)

Tian et al. (2014) determine the size of the map by calculating the number of neurons from the number of observations using the equation $M \approx 5 \times \sqrt{N}$, where M is the number of neurons and N is the number of observations; here, the number of transit users. As presented in Chapter 3, 24.310 transit users will be clustered joining all the 8 areas of study. Applying the equation above, approximately 780 neurons are necessary for the map, resulting in a 2-dimensional map of 28 x 28 neurons.

It is also possible to define the topology of the grid, being it rectangular or hexagonal. The rectangular shape has fewer neighbors (4 for interior cells) than the hexagonal shape (6 for interior cells); the hexagonal topology is therefore used, together with the toroidal topology, whereby the top-bottom and right-left edges are adjacent. Figure 19 illustrates a toroidal grid.

Figure 19 – A toroidal grid in SOM clustering



Source: Carneiro (2015)

The time complexity of the SOM algorithm is *O(N²)*, being *N* the number of users (ROUSSINOV; CHEN, 1998). After SOM is performed, a hierarchical cluster analysis with a complete linkage method is applied to clearly delineate the edges of each cluster. The SOM – Self-Organizing Map clustering algorithm is performed using R codes.

Considering all three clustering algorithms used for identifying different user groups with similar travel patterns – K-means, TwoStep and SOM – and the clustering algorithm used to infer the residents of each area – DBSCAN –, Table 3 presents advantages and limitations of each clustering method aforementioned, summarizing a comparison between the four.

Table 3 – Advantages and limitations of clustering methods

Clustering Method	Advantages	Limitations
K-Means (Partitional)	- High performance - Scalable and simple - Run time faster than hierarchical - Cluster can change for better convergence when centroids are re-computed - Can handle large datasets	- Relies on the random initialization of the cluster center (may fall into local optimum) - Reliance on the user to specify the number of clusters in advance - Quantitative variables only
TwoStep (Hierarchical)	- Easy to understand and interpret - Handle categorical and numerical data - Dendograms great for visualization - Can determine the number of clusters	- Inability to make corrections once the splitting/merging decision is made - Difficult to define levels for clusters - Poor solutions in high dimensional data may be found without proper evaluation
SOM (Model)	- Data mapping easily interpreted - Capable of organizing large, complex data sets - Natural start	- Slow training, hard to train against slowly evolving data - Requires that nearby points behave similarly
DBSCAN (Density)	- Discovery of arbitrary-shaped clusters with varying size - Resistance to noise and outliers	- High sensitivity to the setting of input parameters

Source: The authors' own elaboration, adapted from Saraswathi; Sheela (2014), Sisodia et al. (2012), Namratha; Prajwala (2012).

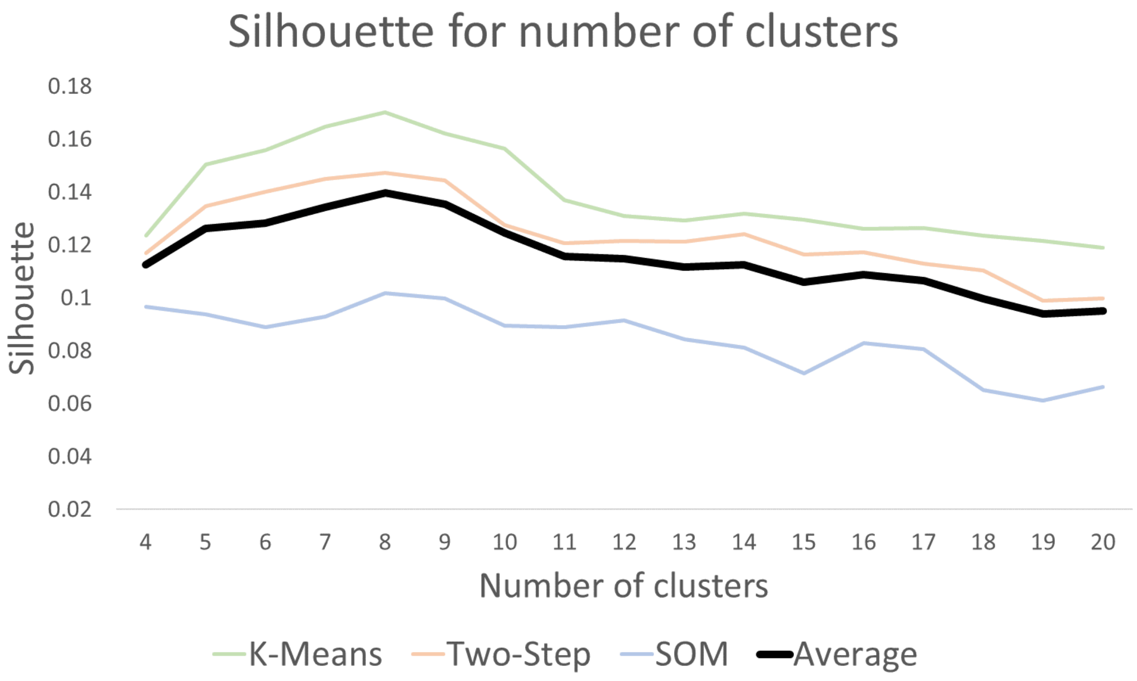
4.2.4 Number of clusters

The number of clusters is an input of the 3 methods depending on the level of granularity expected for the analysis. Morency et al. (2007) defined the number of clusters based on the researchers' experience. Zhao et al. (2017) used the average silhouette coefficient to determine the number of groups. Agard et al. (2013) chose the number of clusters by plotting the data and analyzing it spatially. Here, we use the average silhouette coefficient to find the best number of clusters.

The silhouette coefficient value is used to measure how close each passenger in one cluster is to passengers in the neighboring clusters and thus provides a way to assess the number of clusters visually. For the *i*th passenger, the silhouette coefficient is calculated by $Si = \frac{b_i - a_i}{\max(a_i, b_i)}$, where a_i is the average dissimilarity of the *i*th passenger with all the other passengers within the same cluster. One can interpret a_i as how well *i* is assigned to its cluster; and b_i is the lowest average dissimilarity of the *i*th passenger to any other cluster of which *i* is not a member (ROUSSEEUW, 1987; ZHAO et al., 2017).

Silhouette coefficients have a range of [-1, +1]. Values closer to +1 indicate that the element is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster. Figure 20 shows that, for k = 8, the silhouette coefficient is optimal for the 3 methods.

Figure 20 – Silhouette coefficient for the 3 methods



Source: The authors' own elaboration

5 TRAVEL PATTERN EVALUATION AND DISCUSSIONS

This chapter presents the results and analysis regarding the clusters of travel patterns derived from the methods developed in Chapter 4. Section 5.1 presents some results describing the smart card database features as a whole, before structuring it for the clustering analysis; section 5.2 presents the results and the interpretation of the clusters formed by the three clustering methods here performed (K-means, TwoStep and SOM), section 5.3 presents a spatial distribution of the clustering methods and their analysis, and section 5.4 presents results of the clusters validation from each of the three methods applied. The chapter presents the main results, basis and discussions for the clustering results. The Appendix presents complementary results from this section.

5.1 PREPROCESSING RESULTS

Firstly, the study presents smart card data results before structuring it to perform the clustering methods. These results aim at visualizing some basic features of the areas studied and the behavior of their residents. The charts presented in the preprocessing results are calculated after the inference of transfers and activities – removing any bias from the transfer effects could impact the results – but before filtering weekends and frequent passengers, so that we can have a general overview of the residents from the precarious settlements and middle-class areas, before focusing on our niche of interest.

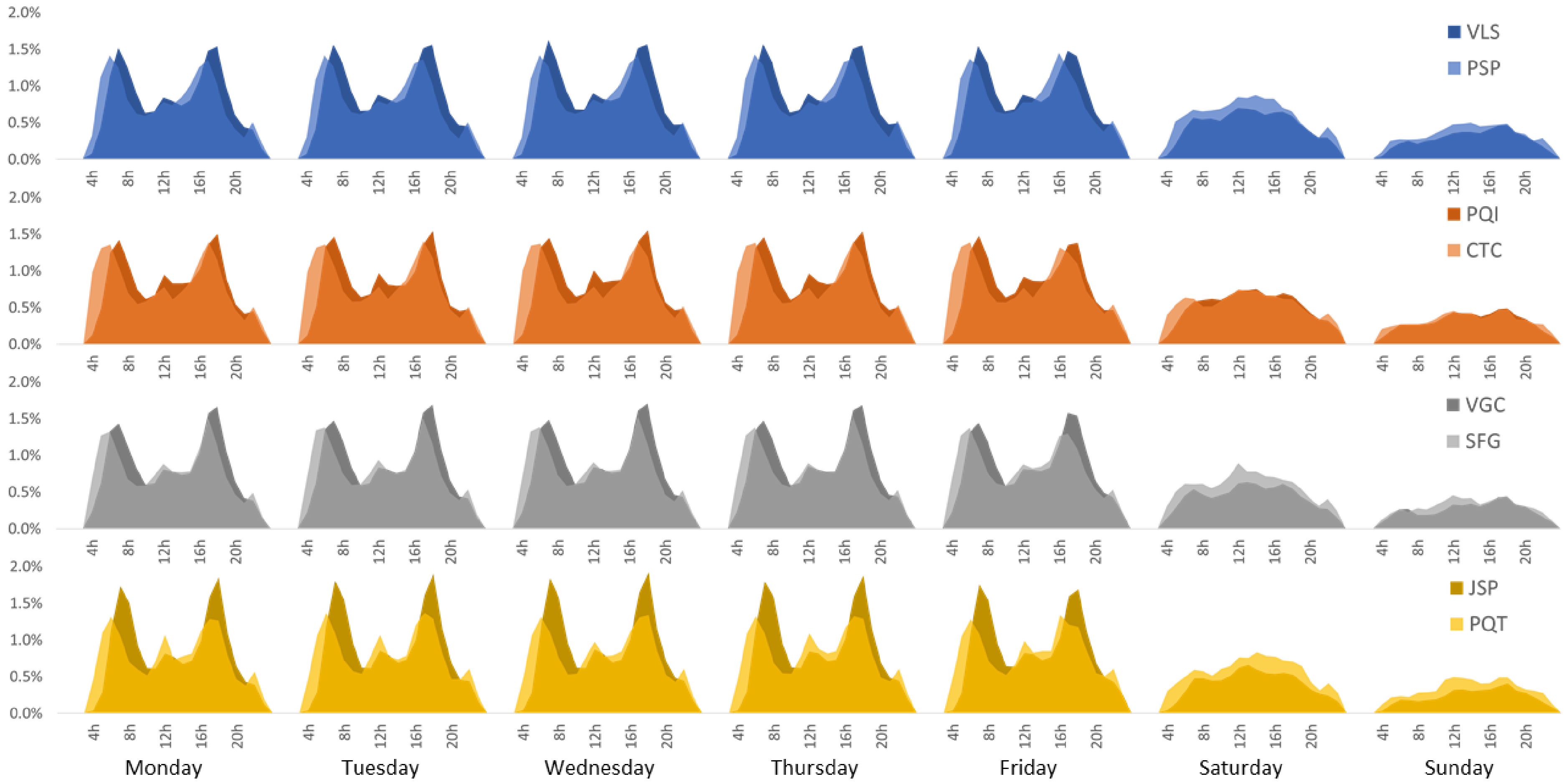
Figure 21 shows a comparison of time profiles between precarious settlements and middle-class areas, according to their location in São Paulo. Paraisópolis is compared to Vila Sônia (both from the western zone), Cantinho do Céu is compared to Parque Interlagos (both from the southern zone), São Francisco Global is compared to Vila Gomes Cardim (both from the eastern zone) and Parque Taipas is compared to Jardim São Paulo (both from the northern zone). The time profile is based on ZHONG et al. (2015). The y-axis indicates the percentage of the validations of the given hour compared to the total number of validations of the day of the week, and the x-axis is an hourly timeline.

Peak hours can be clearly identified in the time profiles, and the difference between weekdays and weekend is significant. The morning and evening peak disappear on weekends. On Saturdays, a small peak can be observed especially in the beginning of the morning, around lunchtime and at the end of the afternoon, suggesting a half day of work for the validations on the peak around lunchtime.

Comparing the precarious settlements with the middle-class areas, we can state that users from precarious settlements validate earlier than users from middle-class areas, both in the morning and in the afternoon peak, with a slight translation of the peak to the left for the precarious settlements. This behavior suggests that the farther located areas of precarious settlements from the job supply force residents to enter the transit system earlier to reach the destination on time.

This earlier validation of precarious settlements residents comparing to the ones of middle-class areas is also present when we evaluate the temporal distribution of the transaction pair between the first and the last transaction hour of the day from these residents by a heat matrix, based on Ma et al. (2013), aiming at demonstrating their temporal travel patterns and their pattern regularity (Figure 22). As shown in the red cells, most of the residents from precarious areas begin their first trip between 5 AM and 7 AM, and end their travel between 4 PM and 6 PM, while most of the residents from middle-class areas begin their first trip between 7 AM and 9 AM, and end their travel between 5 PM and 7 PM. These red cells from the figure are likely to represent a typical commuting trip chain with temporal travel pattern and regularity, whereby the users enter the transit system in the morning heading to their place of work and in the evening return home.

Figure 21 – Comparison of time profiles from the studied areas



Source: The authors' own elaboration

Figure 22 – Temporal distribution for passengers from middle-class and precarious areas throughout the weeks of analysis

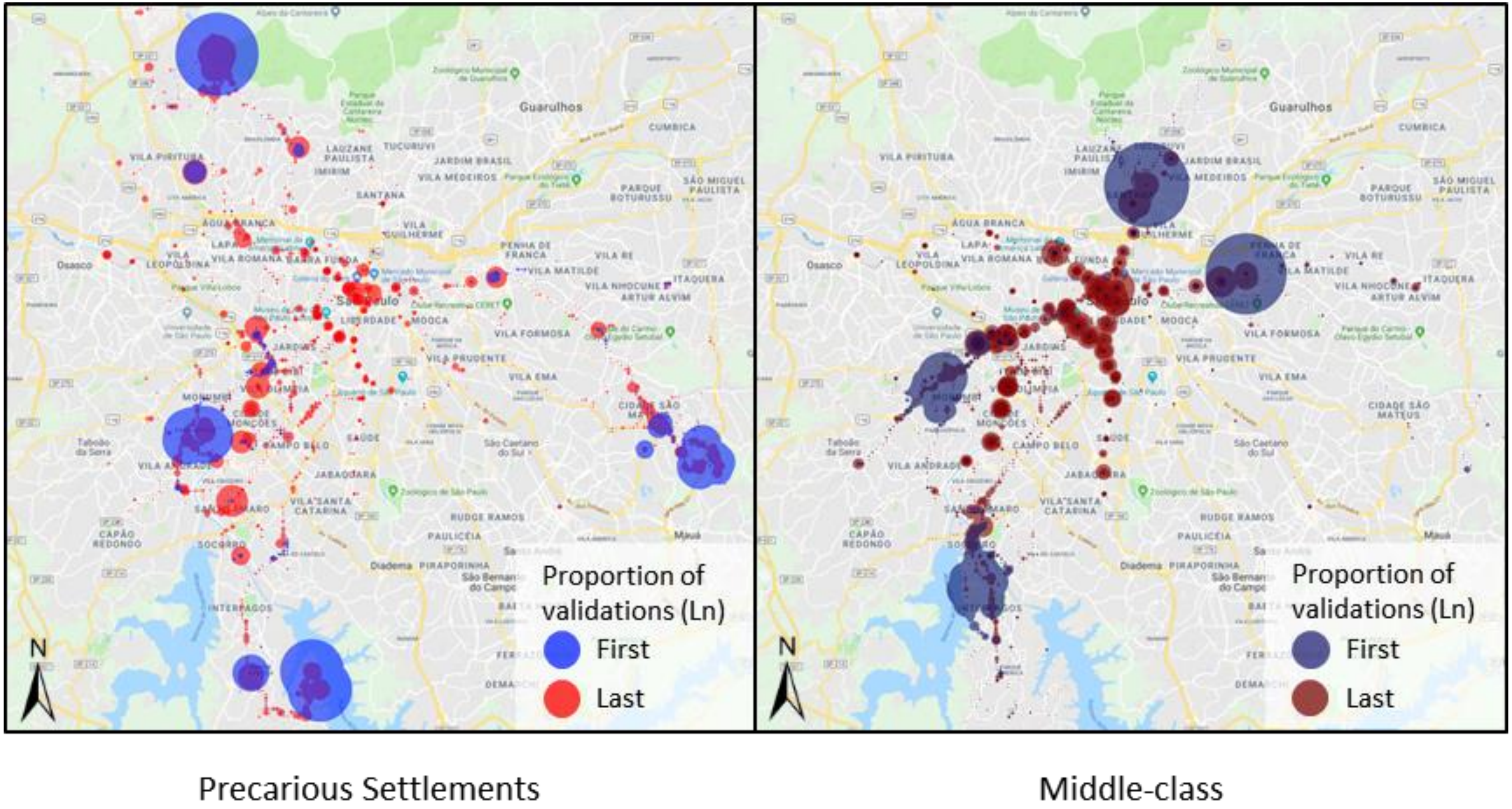
			Last validation hour of the day																			
			4h	5h	6h	7h	8h	9h	10h	11h	12h	13h	14h	15h	16h	17h	18h	19h	20h	21h	22h	23h
First validation hour of the day	Precarious settlements	4h	794	902	127	56	100	161	172	271	718	1,315	7,340	5,430	5,772	7,353	4,517	2,439	1,101	488	364	146
		5h		2,855	1,444	275	252	291	359	561	1,101	2,071	6,831	10,910	15,057	16,743	9,128	5,103	2,000	956	905	345
		6h			4,003	1,496	512	473	555	989	1,919	1,864	3,788	8,371	15,399	16,744	11,074	5,231	2,507	1,711	1,702	636
		7h				3,296	1,322	819	832	886	1,208	1,320	2,407	4,623	10,211	12,421	10,593	5,130	2,457	1,713	1,736	742
		8h					2,362	1,227	804	780	775	741	1,128	1,962	3,788	5,643	6,950	5,346	3,014	1,458	1,610	634
		9h						1,717	1,147	940	781	672	740	1,032	1,755	1,926	2,391	3,459	3,019	1,395	1,759	825
		10h							1,458	1,296	956	741	631	677	803	926	1,077	1,751	3,078	2,043	2,211	1,197
		11h								1,740	1,231	950	719	649	686	782	933	1,252	1,819	2,675	4,840	1,431
		12h									1,780	1,366	1,021	830	716	853	940	952	1,388	1,841	7,551	2,047
		13h										2,011	1,443	1,176	935	716	609	388	698	994	4,149	2,824
		14h											2,206	1,689	1,039	688	444	329	299	378	1,093	1,751
		15h												2,896	1,716	755	460	238	192	216	604	667
		16h													3,312	1,516	722	292	243	312	253	269
		17h														2,890	1,276	591	322	401	395	247
		18h															2,440	1,079	450	502	726	266
		19h																1,543	746	400	304	79
		20h																	1,268	459	137	70
		21h																		728	226	59
		22h																			901	193
		23h																				482
	Middle-income class areas	4h	521	226	21	13	25	36	49	170	423	607	3,143	1,920	2,066	1,529	1,480	1,056	536	204	164	65
		5h		2,664	1,102	195	120	130	173	347	1,119	1,947	3,627	4,505	6,908	8,274	5,334	4,111	2,194	1,250	793	271
		6h			7,025	2,348	520	543	615	1,147	2,959	3,897	4,346	6,074	14,230	24,058	17,294	8,972	6,752	4,563	2,646	693
		7h				10,913	2,969	1,045	887	1,321	2,324	2,062	4,732	6,018	11,125	29,085	35,448	14,398	7,712	5,406	4,418	1,455
		8h					8,554	2,085	1,159	1,118	1,399	1,581	1,739	3,989	7,715	14,410	30,233	18,786	8,674	5,172	4,293	1,614
		9h						5,430	1,798	1,488	1,159	1,042	1,291	1,759	4,844	6,690	9,285	12,720	6,394	3,309	2,573	977
		10h							3,320	1,704	1,302	999	947	1,177	1,433	2,681	3,611	4,903	4,102	2,263	1,813	719
		11h								2,959	1,541	1,148	966	970	1,331	1,650	2,634	3,155	2,464	1,854	2,213	651
		12h									4,407	1,910	1,472	1,352	1,545	2,405	2,437	3,015	2,421	1,953	3,936	1,197
		13h										4,276	2,086	1,955	1,540	1,498	1,633	1,400	2,915	1,842	3,480	1,768
		14h											3,706	2,355	1,820	1,209	859	857	741	1,383	1,446	974
		15h												4,375	2,557	1,451	739	624	559	569	1,114	553
		16h													6,216	3,272	1,262	571	545	604	640	616
		17h														9,692	3,552	1,046	836	957	1,168	417
		18h															8,828	2,429	1,101	1,214	1,723	488
		19h																5,495	1,470	843	650	271
		20h																	3,474	1,098	219	145
		21h																		2,105	463	70
		22h																			1,508	286
		23h																				785

Source: The authors’ own elaboration, based on Ma et al. (2013)

Still evaluating the first and the last transactions, we now look at the spatial distribution of the transaction pair through a proportional circles map, inspired by Briand et al. (2017), for a better idea of how validations distribute throughout the city (Figure 23). The size of the circle is proportional to the number of validations in the given location. The distribution between precarious and middle-class areas are also different in a spatial evaluation, especially regarding the last transactions. The first transactions are mostly located at the users' residences for both the precarious and middle-class areas. For the middle-class areas, the last validations are more densely concentrated in commercial and central areas, such as Paulista Ave., Faria Lima Ave., Vila Olímpia, Pinheiros and Santo Amaro. These locations are also highlighted in the last validations of precarious settlements, but the overall distribution of validations is sparser.

Figure 24 shows the histograms of the frequency of validations along the period of analysis, also comparing the precarious settlements with the middle-class areas. The number of occasional passengers, with a low number of validations during the analyzed period, is high for both areas, and higher for middle-class areas. As days go by, the number of users strongly decreases, stabilizing around 25 days. Afterward, from approximately 54 to 66 days, one can state an increase in the percentage of users, suggesting that these are regular passengers using the transit system on business days comprehended between 55 and 66 days (5 to 6 days a week along the 11 weeks of data). The higher proportion of occasional passengers in middle-class areas also suggests that precarious settlements residents rely more on the transit system than middle-class area residents, which are able to use public transport more sparsely, according to their convenience. However, precarious settlement users have fewer modal choices and are therefore more frequent on the transit system.

Figure 23 – Map representing the proportion of validations per study area

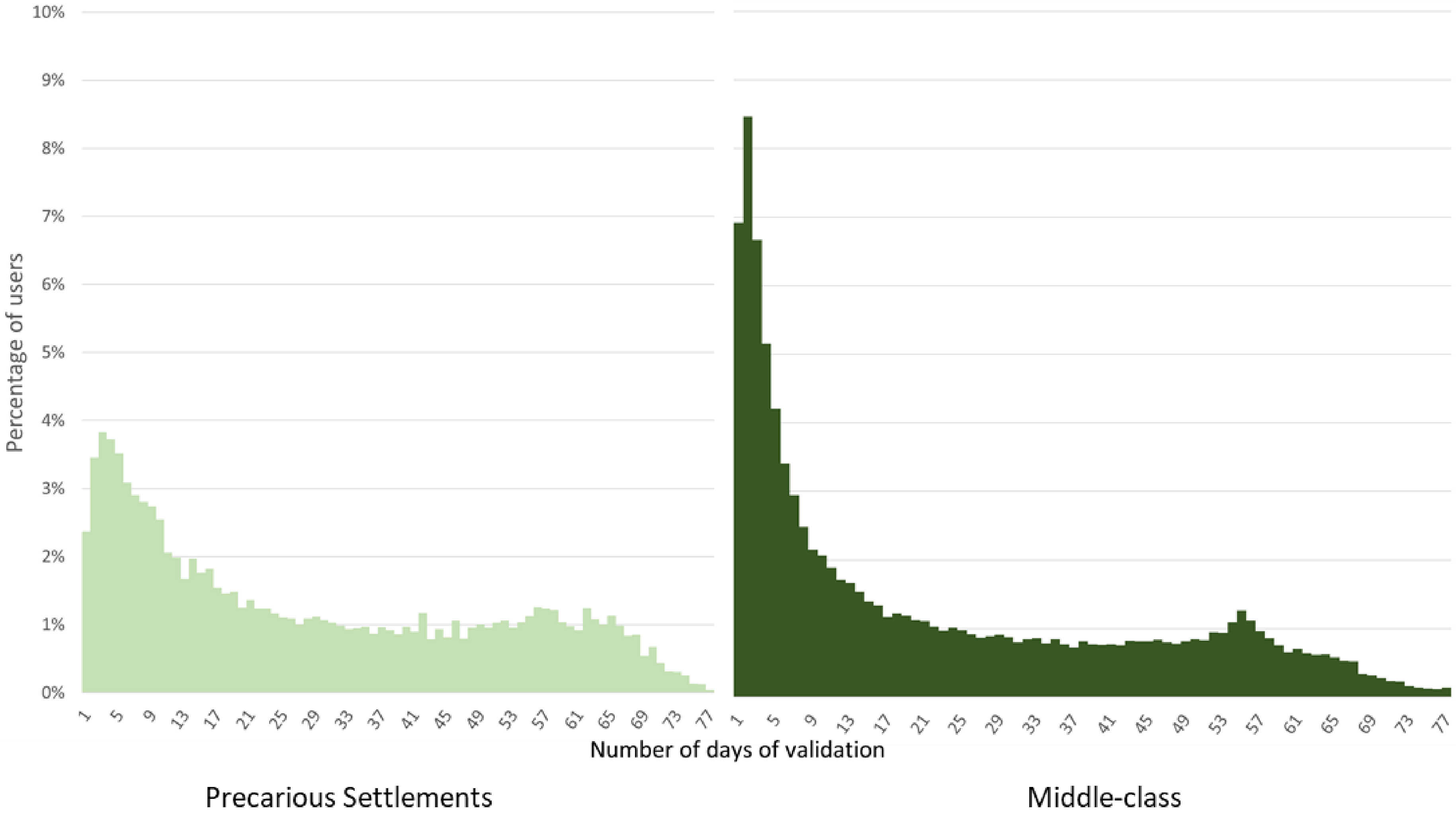

Proportion of validations (Ln)
First
Last
N
Proportion of validations (Ln)
First
Last
N

Precarious Settlements

Middle-class

Source: The authors' own elaboration

Figure 24 – Comparison of histograms between areas



Source: The authors' own elaboration

5.2 CLUSTER ANALISYS

This section presents different aspects of the clustering results and analyzes each of them in terms of their features from the clustered attributes, their similarities and differences within clusters and the variation of the clustering results between the three different algorithms used. The assignment of a specific user to a cluster means that the behavior of this user is closer to the behavior of other passengers assigned to the same cluster than the behavior of any people in other clusters (AGARD et al., 2013).

Figure 25 to Figure 32 present the features of each cluster derived from k-means clustering for simplification purposes, although TwoStep and SOM clustering also present similar results and are presented in the Appendix. The first group of charts, on the left, presents scales with respect to the clustered variables of the Start hour of travel; Daily distance; Daily activity duration; their respective dispersions and Income, with the median of each variable in orange for reference. The top right corner depicts the proportion of activity validations in each of the five types of land use considered for the study in the given cluster. Below this chart, in blue, the size of the cluster is presented and, in green, the proportion of middle-class and precarious settlement residents in the given cluster.

Passengers classified in Cluster 1, 3 and Cluster 5 are naturally associated with regular travel pattern commuting passengers. They all have low variability in their validations, with an early first validation of day – around 7:30 AM –, a duration of intra-validations (associated to activities) of about 10 hours – which suggests full time working hours –, and a high frequency of transit use – 5 days per week. They also have similar traveled distances, of about 9 km. Their differences start to appear when analyzing their households median income. Cluster 3 and 5 have exactly the same median of BRL 3,869.00, while the one from Cluster 1 is lower, of BRL 2,979.00. This difference helps to explain their area proportions, with Cluster 3 and 5 being mainly from middle-class areas (~70%) and Cluster 1 being almost equally divided by middle-class and precarious settlement areas (52% middle-class, 48% precarious settlement). Another difference lies on the land use activity validations. Cluster 1 validates mainly the Residential/Commercial land use (46%), Cluster 3 the Other areas

(74%) and Cluster 5, the Commercial/Services/Industrial areas (73%). The greater similarities of Cluster 3 and 5 suggest that Other land use, despite its general term, is intrinsically related to job supply areas. Together these three clusters represent about 34% of all users – 10% from Cluster 1, 10% from Cluster 3 and 14% from Cluster 5.

While Cluster 1, 3 and 5 have evidence of commuting passengers working in areas of regular labor supply concentration, Cluster 2 has the particularity of being formed by users that suggest an association with low-paid employment. This association is due to the largest proportion of users from precarious settlement areas (61%) and the largest proportion of activity validations in Residential Medium/High-Income land use areas (51%), whereby the job supply is mainly composed by workers with functions such as caretakers, doormen, gardeners, janitors, maids and similar jobs which usually provide this type of service to residents of houses and buildings of these areas. This evidence is reinforced by the other variable results from Cluster 2, with its users being almost the lowest household incomes between all clusters (median of BRL 1,488.00 – a low value in Brazilian standards), second only to Cluster 4, presented next. The distances traveled by users from Cluster 2 are lower than the average of all clusters (including Cluster 1, 3 and 5), suggesting that the Residential Medium/High-Income land use working areas are closer than the typical commercial/industrial/service areas, with users of Cluster 2 working in middle-class areas relatively near the precarious settlements. Also, these passengers leave home early in the morning (around 7:30 AM), spend about 10h in their activity and have a median frequency of 5 days per week, and a low variability in validations, again rectifying a commuting pattern. Cluster 2 represents more than 16% of the studied passengers.

Cluster 4 also suggests a particular commuting pattern, besides the possibility of low-paid employment. It has the lowest household income of all clusters (BRL 1,321.00), a vast majority of users from precarious settlement areas (around 75%) and an activity validation mainly in Residential Low-Income land use areas, suggesting a passenger with local job, in small commercial and service spots on their own neighborhood, reinforced by their low displacement in the city, of only 4.1 km in an overall median of 7.1 km. Its first hour of validation (around 7:30 AM) and duration of activity (9 hours) suggests a commuting pattern, and a median weekly frequency of 4 days suggests that these users do not work on all working days, maybe working from three to five times a week (as their dispersion shows). Cluster 4 also has a significant proportion of Student card type (Table 4), suggesting children going to their full-time school, which is a reasonable assumption when analyzing the patterns since children usually study near home.

With a high variability of first hour, duration, distance and weekly frequency validations, Cluster 6 users have low frequent afternoon part-time activities or less (median if 3 times a week, median of first hour validations of 12:30 PM and duration median of 2.8 hours), using the transit system to reach nearby facilities (median of traveled distance of 4.5 km) and being users with pattern of non-travelers. Examples of users from Cluster 6 may be housekeepers leaving home for leisure activities or their children going to extracurricular activities – cluster 6 is the cluster with the highest proportion of Student card type (shown in Table 4). This cluster is mainly formed by middle-class area residents (69%) – with less need of regular formal work for all households residents –, has an equally distributed activity validation between land use areas and is formed by 20% of all users.

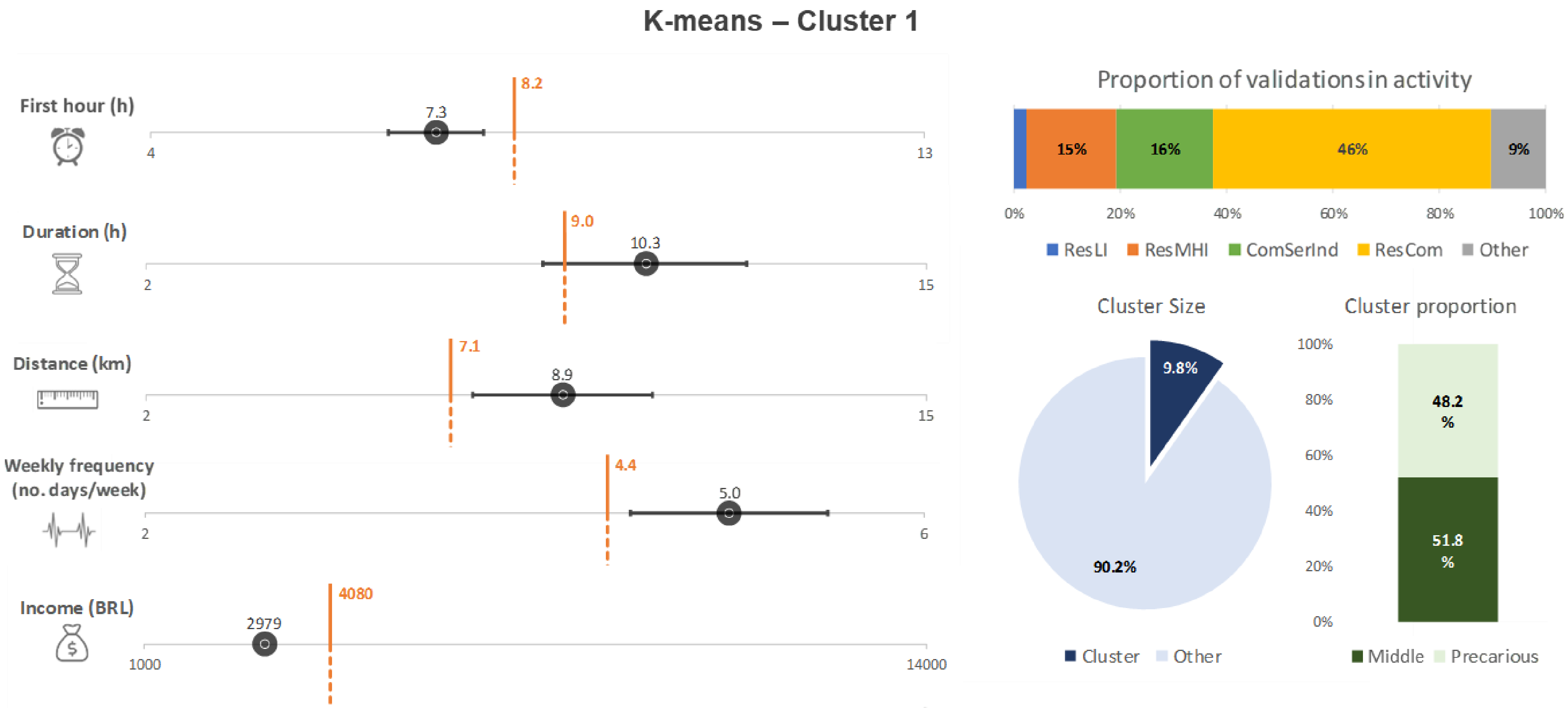
Cluster 7 has the least defined pattern of all clusters due to the high dispersion of all its variables. In fact, Cluster 7 has the highest variability as compared to the other clusters. It comprises passengers that suggest both full and part-time employments – median of first hour validations of 7:30 AM, but with a dispersion of 3.4 hours, also encompassing users possibly validating for the afternoon work shift; and duration median of 10 hours, but also with a high dispersion of 4.3 hours which may reach a 6-hour part-time shift. It has a dispersed activity validation land use and an equally distributed proportion of middle-class and precarious settlement residents (54.6% of precarious settlement). Cluster 7 represents 20% of the users.

Lastly, Cluster 8 is clearly formed based on its users' household income. With BRL 13,740.50, it is far the highest household income of all clusters, confirmed by the proportion of more than 97% from middle-class areas. Its activity validation land use is fairly distributed, the median of first hour validation is later than the other clusters (around 8:30 AM), the median duration between validations is of 9 hours and it has a low distance to reach activity, compared to other clusters (5.4 km). It represents only 3.6% of all users.

Table 4 presents a compilation of all the attribute values of each cluster for the three clustering algorithms. In this compilation of results, it is possible to compare their results in a single table, and the color scale shows that all the three methods have similar results of clusters, with almost the same highlighted cells for each clustering method. A color-free table is presented in the Appendix for each clustering method.

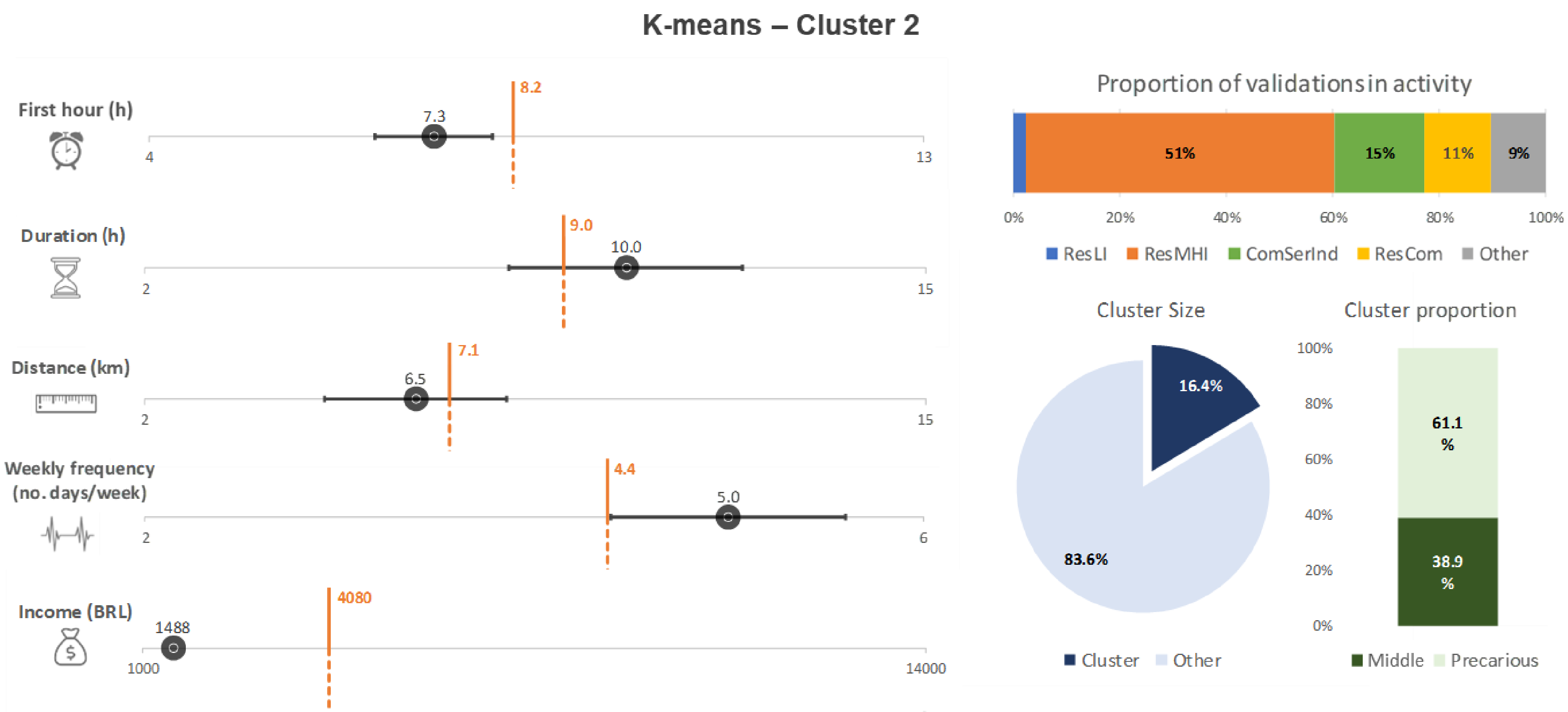
Table 5 presents the proportion of areas in each cluster and the proportion of clusters in each area. Evaluating the clusters (the tables on the left) Paraisópolis is noted to have a strong participation in Cluster 2 and Vila Gomes Cardim in Cluster 3. Also, Cluster 4 (the poorest) has users mainly from Cantinho do Céu. We can also observe that cluster 8 from K-means and SOM is almost entirely formed by Vila Sônia and Parque Interlagos, respectively, except for the TwoStep clustering, which did not form Cluster 8 based on its users' household income. Evaluating the areas (the tables on the right), we can state that Paraisópolis users are mainly classified into Cluster 2, Sao Francisco Global and Cantinho do Céu mainly into Cluster 7 and Parque Taipas into Clusters 4 and 7. The middle-class areas are more distributed, with a slight concentration in Cluster 6.

Figure 25 – Groups formed from the k-mean clustering algorithm – Cluster 1



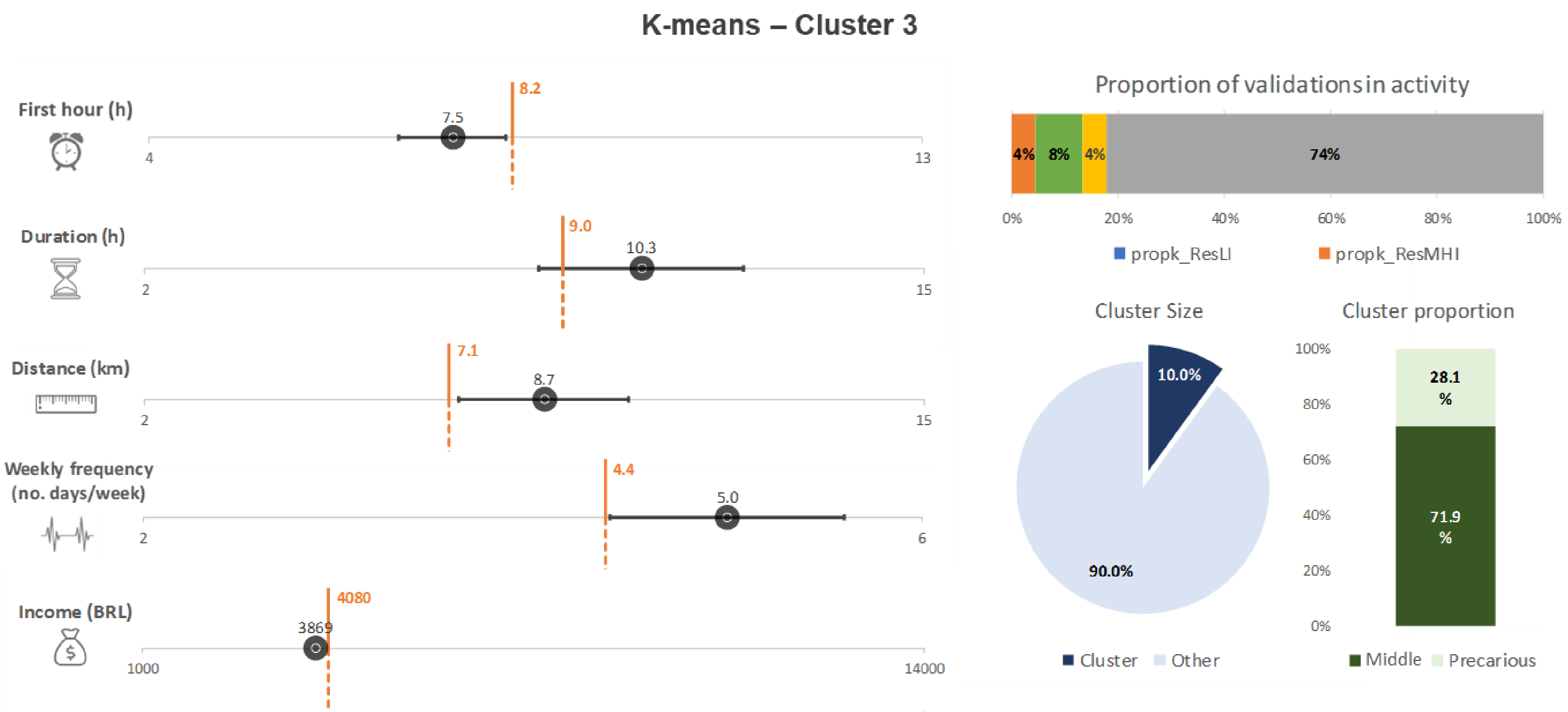
Source: The authors’ own elaboration

Figure 26 – Groups formed from the k-mean clustering algorithm – Cluster 2



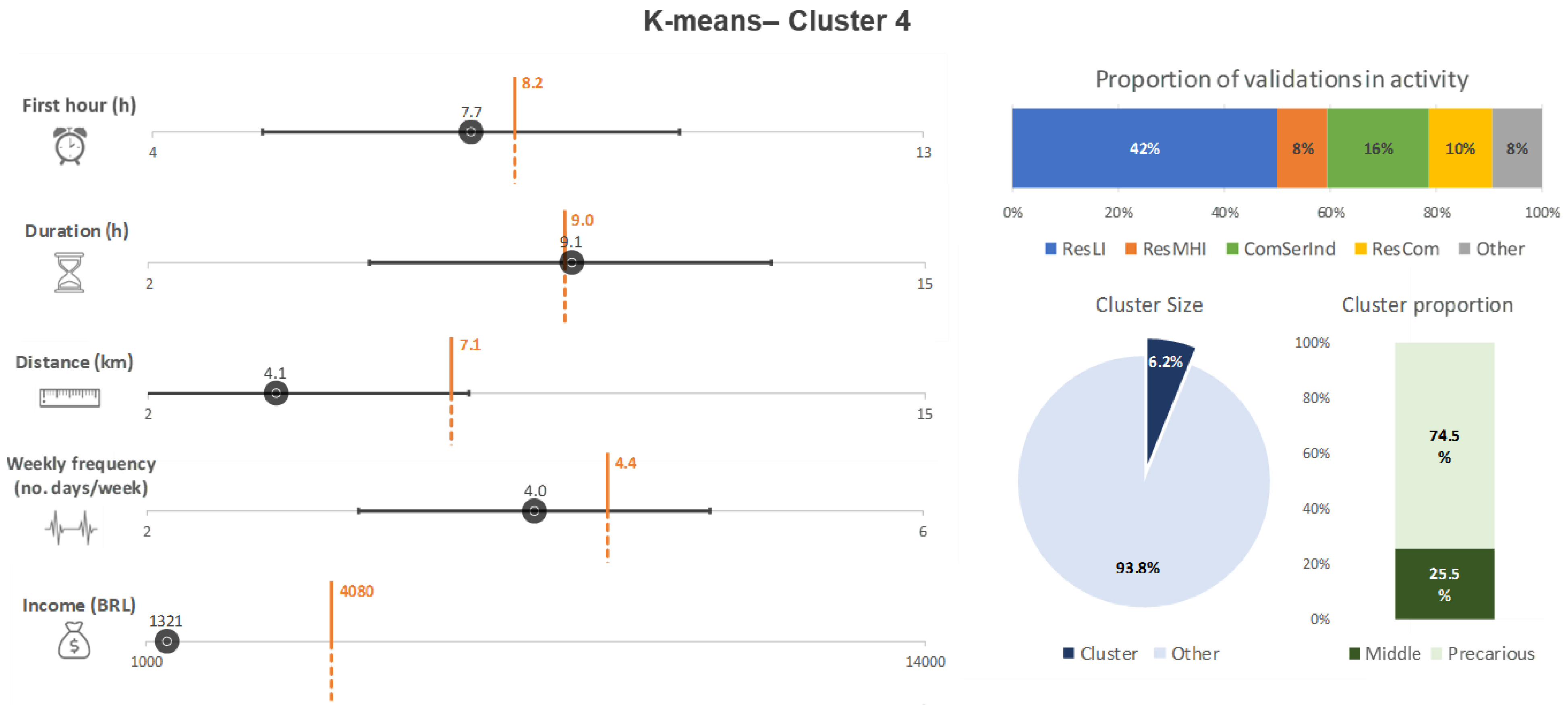
Source: The authors' own elaboration

Figure 27 – Groups formed from the k-mean clustering algorithm – Cluster 3



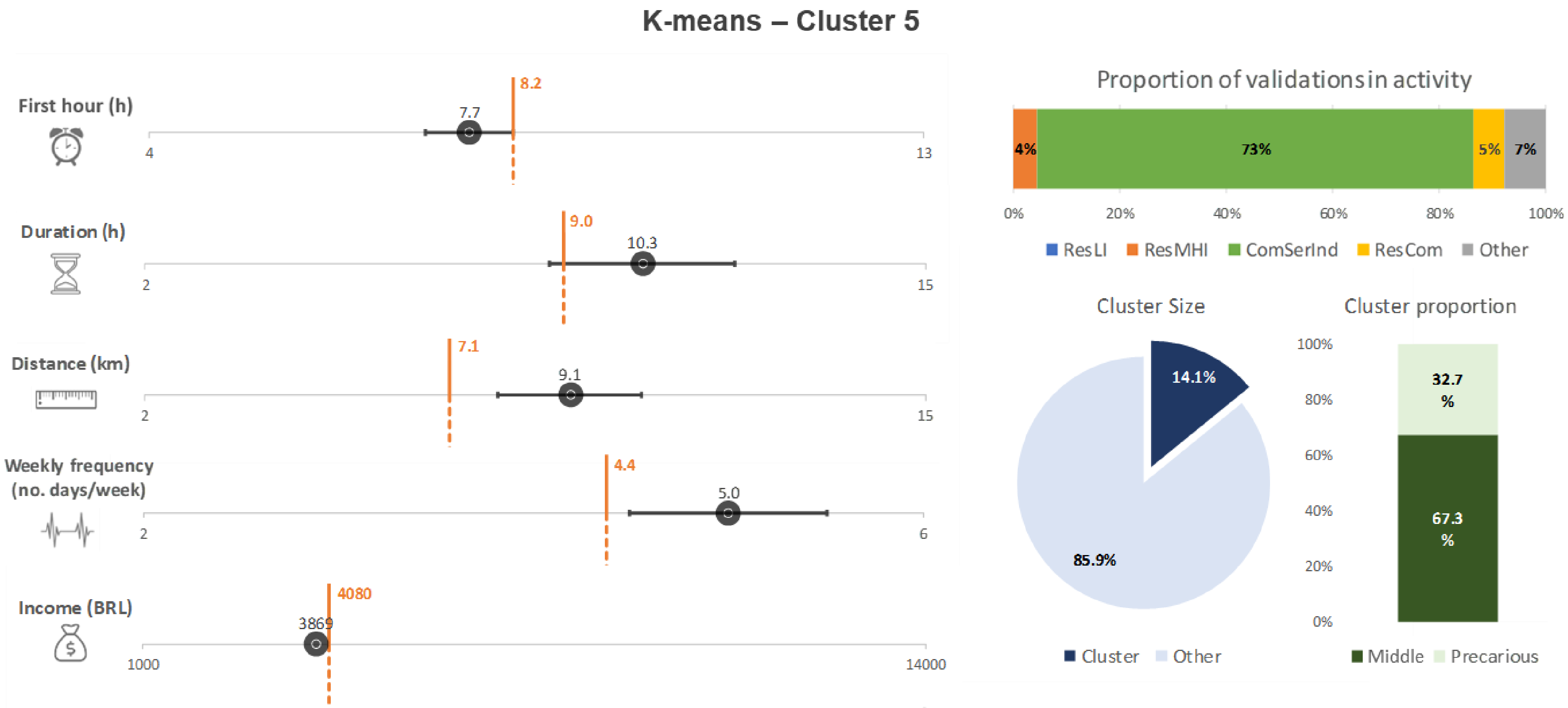
Source: The authors' own elaboration

Figure 28 – Groups formed from the k-mean clustering algorithm – Cluster 4



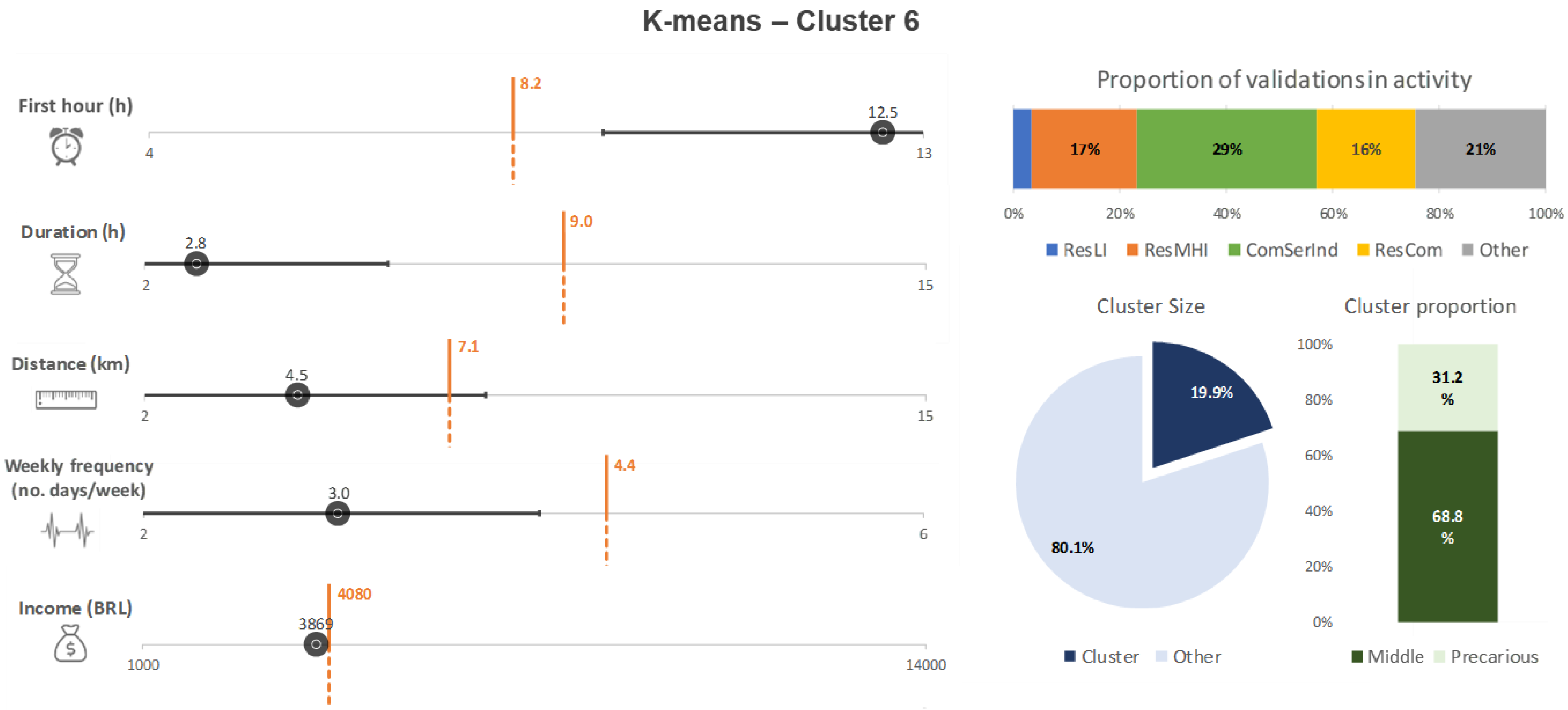
Source: The authors' own elaboration

Figure 29 – Groups formed from the k-mean clustering algorithm – Cluster 5



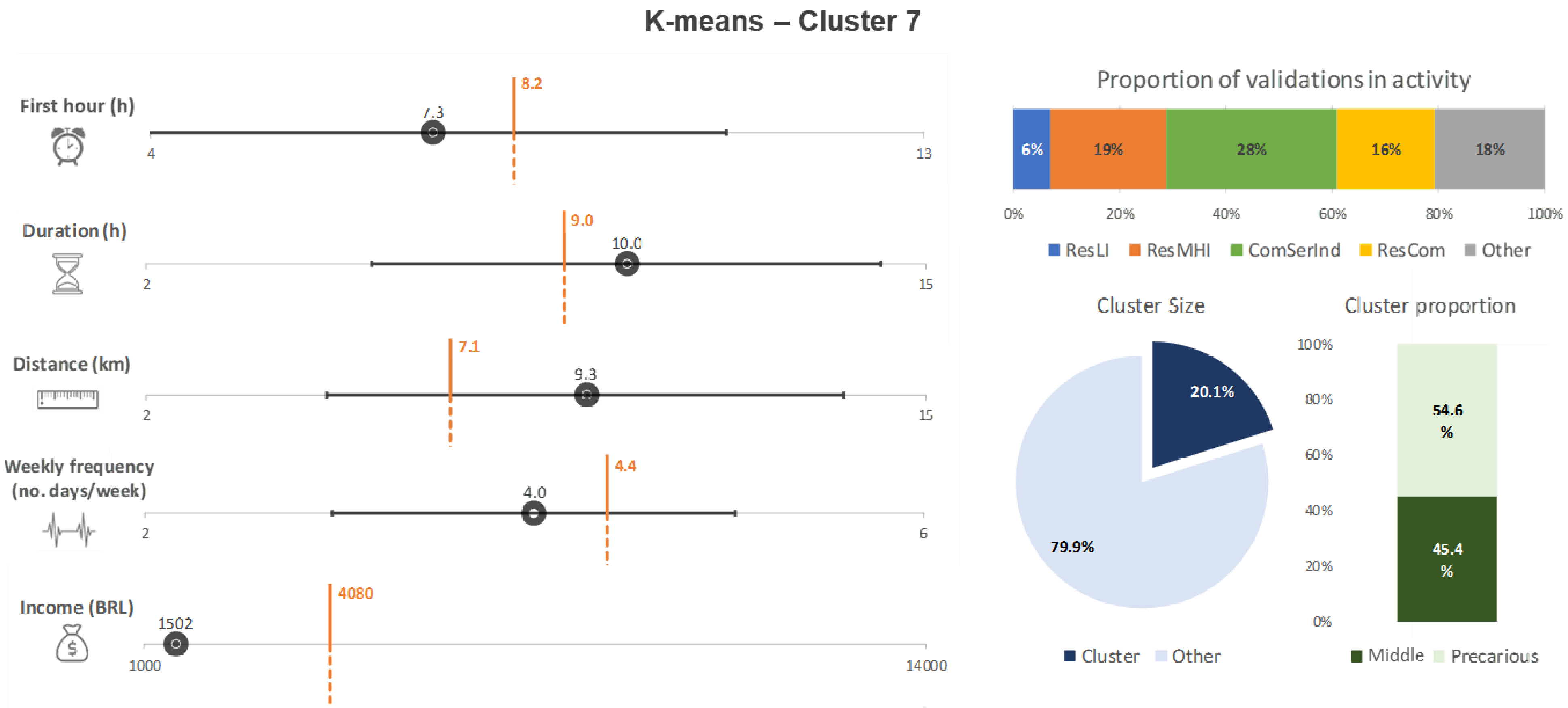
Source: The authors' own elaboration

Figure 30 – Groups formed from the k-mean clustering algorithm – Cluster 6



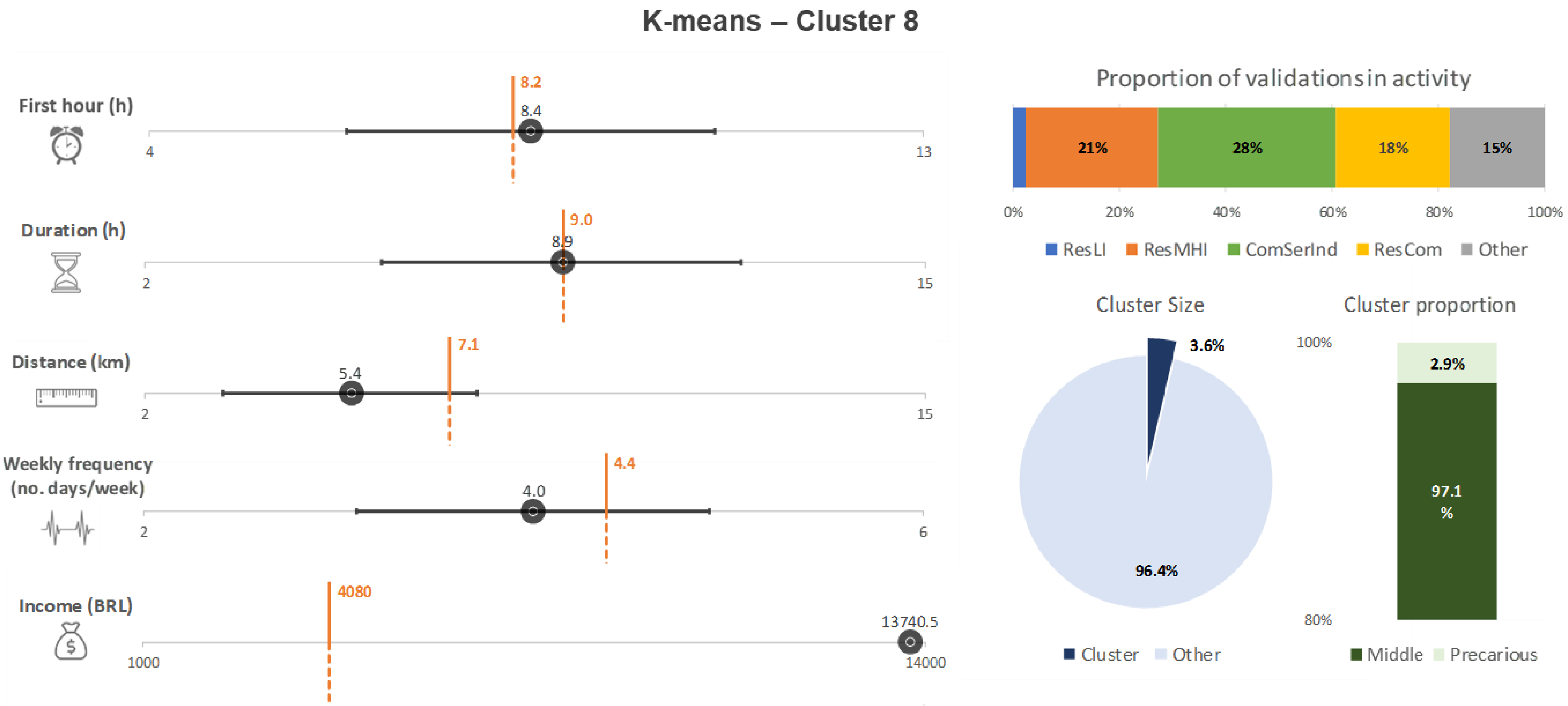
Source: The authors' own elaboration

Figure 31 – Groups formed from the k-mean clustering algorithm – Cluster 7



Source: The authors' own elaboration

Figure 32 – Groups formed from the k-mean clustering algorithm – Cluster 8



Source: The authors' own elaboration

Table 4 – Result comparison among clustering algorithms

	Cluster Number	Income (BRL)	Med_ FirstHour (h)	Std_ FirstHour (h)	Med_ Duration (h)	Std_ Duration (h)	Med_ MaxDist (km)	Std_ MaxDist (km)	Med_ WeekFreq (days/week)	Std_ WeekFreq (days/week)	Res_ LowInc	Res_ MedHigInc	Com_ SerInd	Res_ Com	Other	Area		Cluster Size
																Middle-Class	Precarious	
KMEANS CLUSTERING	1	R$ 2,979	7.3	0.6	10.3	1.7	8.9	1.5	5.0	0.5	2%	15%	16%	46%	9%	52%	48%	10%
	2	R$ 1,488	7.3	0.7	10.0	2.0	6.5	1.5	5.0	0.6	2%	51%	15%	11%	9%	39%	61%	16%
	3	R$ 3,869	7.5	0.6	10.3	1.7	8.7	1.4	5.0	0.6	0%	4%	8%	4%	74%	72%	28%	10%
	4	R$ 1,321	7.7	2.4	9.1	3.4	4.1	3.3	4.0	0.9	42%	8%	16%	10%	8%	25%	75%	6%
	5	R$ 3,869	7.7	0.5	10.3	1.5	9.1	1.2	5.0	0.5	0%	4%	73%	5%	7%	67%	33%	14%
	6	R$ 3,869	12.5	3.2	2.8	3.2	4.5	3.2	3.0	1.0	3%	17%	29%	16%	21%	69%	31%	20%
	7	R$ 1,502	7.3	3.4	10.0	4.3	9.3	4.3	4.0	1.0	6%	19%	28%	16%	18%	45%	55%	20%
	8	R$13,741	8.4	2.1	8.9	3.0	5.4	2.1	4.0	0.9	2%	21%	28%	18%	15%	97%	3%	4%
TWOSTEP CLUSTERING	1	R$ 4,816	8.0	1.8	9.7	2.9	8.3	2.1	5.0	0.8	0%	12%	17%	46%	10%	78%	22%	9%
	2	R$ 1,488	7.0	0.4	10.3	1.2	8.4	1.4	5.0	0.5	2%	43%	17%	17%	9%	40%	60%	17%
	3	R$ 3,869	7.5	0.7	10.3	1.8	8.8	1.5	5.0	0.6	0%	4%	8%	4%	74%	73%	27%	10%
	4	R$ 1,322	8.0	2.6	8.7	3.4	4.2	3.3	4.0	0.9	40%	9%	18%	11%	9%	25%	75%	7%
	5	R$ 3,869	7.7	0.4	10.3	1.3	9.4	0.9	5.0	0.5	0%	3%	77%	4%	6%	70%	30%	12%
	6	R$ 3,869	12.7	3.3	2.5	3.1	4.7	3.3	3.0	1.0	3%	16%	28%	15%	24%	73%	27%	18%
	7	R$ 1,368	6.9	3.1	10.4	4.0	14.3	6.0	5.0	0.9	8%	15%	29%	15%	17%	31%	69%	11%
	8	R$ 3,517	8.0	3.0	9.3	3.8	5.5	2.3	4.0	1.1	3%	30%	28%	15%	14%	53%	47%	16%
SOM CLUSTERING	1	R$ 3,869	7.1	1.3	10.5	2.6	10.1	2.3	5.0	0.8	0%	5%	9%	70%	5%	63%	37%	3%
	2	R$ 1,596	7.3	0.6	10.2	1.6	7.2	1.6	5.0	0.5	2%	36%	19%	17%	11%	43%	57%	24%
	3	R$ 3,869	7.5	0.5	10.3	1.3	9.9	0.9	5.0	0.5	0%	2%	6%	3%	82%	73%	27%	7%
	4	R$ 1,343	8.2	2.6	8.6	3.4	4.3	3.3	4.0	0.9	38%	10%	18%	11%	9%	26%	74%	8%
	5	R$ 3,869	7.9	0.5	10.2	1.5	8.6	1.0	5.0	0.5	0%	4%	74%	5%	7%	72%	28%	11%
	6	R$ 3,869	12.2	3.3	2.5	3.3	5.0	3.3	3.0	1.0	3%	15%	27%	16%	22%	72%	28%	17%
	7	R$ 1,596	7.6	3.1	9.8	3.9	9.0	3.7	4.0	1.0	4%	20%	27%	15%	17%	47%	53%	26%
	8	R$12,453	8.4	1.8	9.0	2.7	5.6	2.0	4.0	0.8	2%	21%	26%	17%	14%	98%	2%	5%

Source: The authors’ own elaboration

Table 5 – Proportion of areas in each cluster and proportion of clusters in each area

Kmeans	PSP	SFG	CTC	PQT	VLS	VGC	PQI	JSP	Total
1	27%	5%	14%	2%	7%	25%	9%	11%	100%
2	46%	2%	12%	1%	8%	17%	6%	8%	100%
3	11%	6%	10%	1%	5%	47%	7%	14%	100%
4	17%	17%	36%	4%	2%	11%	8%	4%	100%
5	14%	6%	12%	1%	8%	39%	9%	11%	100%
6	21%	4%	5%	1%	11%	29%	11%	17%	100%
7	20%	11%	22%	2%	5%	22%	9%	9%	100%
8	3%	0%	0%	0%	46%	3%	40%	7%	100%

Kmeans	1	2	3	4	5	6	7	8	Total
PSP	12%	33%	5%	5%	9%	18%	18%	0%	100%
SFG	8%	6%	9%	16%	13%	14%	34%	0%	100%
CTC	10%	14%	7%	16%	12%	8%	32%	0%	100%
PQT	14%	10%	6%	20%	9%	11%	29%	0%	100%
VLS	8%	14%	6%	2%	13%	25%	12%	19%	100%
VGC	9%	11%	18%	3%	21%	22%	17%	0%	100%
PQI	9%	10%	7%	5%	12%	24%	18%	15%	100%
JSP	10%	12%	12%	2%	15%	31%	16%	2%	100%

TwoStep	PSP	SFG	CTC	PQT	VLS	VGC	PQI	JSP	Total
1	13%	2%	5%	1%	21%	22%	21%	15%	100%
2	41%	4%	15%	1%	7%	19%	7%	7%	100%
3	11%	6%	10%	1%	6%	46%	7%	14%	100%
4	18%	16%	36%	4%	2%	10%	9%	4%	100%
5	13%	5%	11%	1%	9%	42%	9%	11%	100%
6	18%	5%	4%	1%	11%	32%	12%	17%	100%
7	8%	18%	40%	2%	1%	19%	6%	4%	100%
8	41%	2%	4%	1%	12%	19%	9%	12%	100%

TwoStep	1	2	3	4	5	6	7	8	Total
PSP	6%	31%	5%	6%	7%	14%	4%	29%	100%
SFG	3%	11%	9%	19%	9%	13%	32%	5%	100%
CTC	3%	18%	7%	18%	9%	6%	33%	5%	100%
PQT	6%	15%	6%	24%	7%	10%	21%	10%	100%
VLS	22%	14%	6%	2%	11%	22%	1%	21%	100%
VGC	8%	12%	17%	3%	18%	22%	8%	12%	100%
PQI	20%	12%	7%	6%	10%	22%	7%	15%	100%
JSP	12%	11%	12%	2%	12%	28%	4%	18%	100%

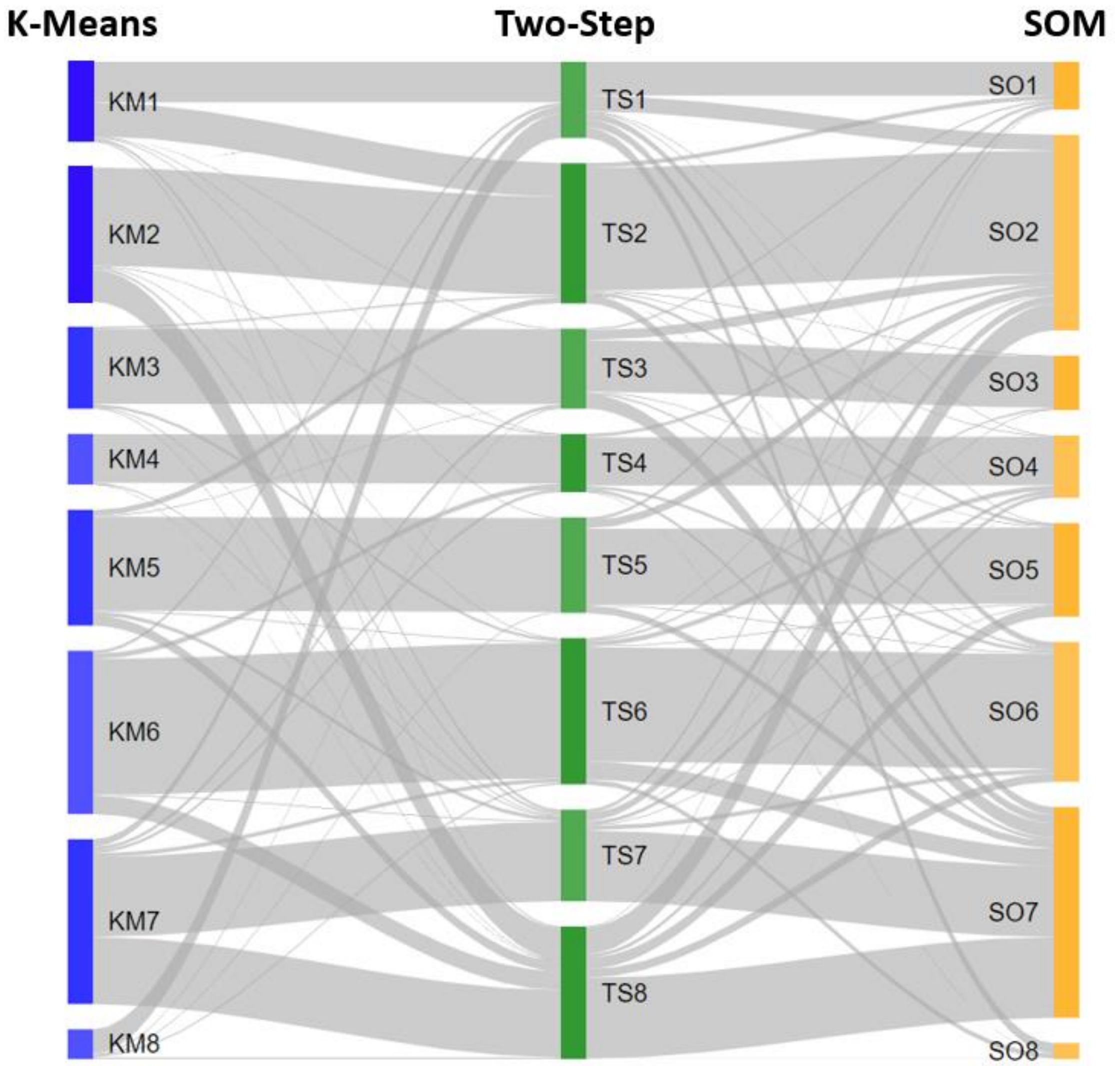
SOM	PSP	SFG	CTC	PQT	VLS	VGC	PQI	JSP	Total
1	8%	7%	19%	2%	6%	35%	9%	13%	100%
2	41%	3%	12%	1%	7%	20%	7%	8%	100%
3	9%	5%	11%	1%	5%	50%	6%	13%	100%
4	21%	16%	33%	4%	2%	11%	9%	4%	100%
5	14%	4%	8%	1%	9%	42%	9%	12%	100%
6	17%	5%	6%	1%	10%	34%	10%	17%	100%
7	22%	9%	20%	2%	7%	21%	9%	10%	100%
8	2%	0%	0%	0%	40%	5%	39%	13%	100%

SOM	1	2	3	4	5	6	7	8	Total
PSP	1%	43%	3%	7%	7%	13%	25%	0%	100%
SFG	4%	13%	6%	20%	8%	13%	37%	0%	100%
CTC	4%	21%	5%	18%	7%	7%	37%	0%	100%
PQT	5%	18%	5%	23%	5%	12%	33%	0%	100%
VLS	2%	19%	3%	2%	12%	20%	20%	21%	100%
VGC	4%	18%	13%	3%	18%	22%	21%	1%	100%
PQI	3%	17%	4%	7%	10%	17%	24%	18%	100%
JSP	4%	18%	8%	3%	12%	27%	23%	5%	100%

Source: The authors' own elaboration

Trying to evaluate the stability of the clusters, i.e. whether they keep their features independently of the clustering algorithm applied – showing that the clusters are not merely classified at random, but a strong indicative that the users classified in the same cluster indeed have similarities in travel patterns – Figure 33 presents a Sankey test of the clusters, in which users were classified throughout the clustering methods.

Figure 33 – Exchange between groups within clustering algorithms



Source: The authors' own elaboration

The stable flows in the Sankey test indicate that users keep within their groups throughout the three clustering algorithms (mainly straight flows), with some minor variations. Evaluating the exchanges of users, 55% of the users were clustered in the same group among all the clustering algorithms, 37% of the users were clustered in the same group among at least two clustering algorithms and only 8% of the users were clustered in three different groups among the three clustering algorithms. Table 6 presents the percentage of users in the same cluster by the clustering method.

Table 6 – Percentage of users in the same cluster by clustering method

	K-Means	TwoStep	SOM
K-Means	100%	70%	69%
TwoStep	70%	100%	63%
SOM	69%	63%	100%

Source: The authors' own elaboration

5.3 CLUSTER SPATIAL DISTRIBUITION

This section describes and analyzes the spatial distribution of users with different travel patterns, according to their clusters. Understanding different patterns may allow identifying passenger groups with a higher level of travel behavior predictability, which could eventually support the evaluation of potential transport planning improvements (ORTEGA-TONG, 2013).

For each cardholder, the first validations of each day are discarded and the remaining validations are clustered by the DBSCAN algorithm. The parameters applied are the same from section 4.1.3, i.e. ε = 1km and minPts = 2. Therefore, for each user, a centroid of activity is inferred. These users are then aggregated based on their clusters, and Figure 34 to Figure 41 present the maps resulting from this process using the K-means clustering results. The residence area is the beginning arrow and the activity centroid is the ending arrow, for each user. The upper map plots the precarious settlement area flows, while the lower map plots the middle-class area flows. The maps resulting from the TwoStep and SOM clustering are presented in the Appendix.

The similarities of clusters 1 and 3 (and 5, which will be mentioned next) can be reinforced when analyzing their spatial distribution, showing a direction to commercial (central) areas of the city and which can therefore still be associated to regular travel pattern commuting passengers.

While Clusters 1 and 3 are more focused on activities towards central areas, Cluster 2 shows a spreaded distribution of its activity locations, still suggesting the access to Residential Medium/High-Income areas.

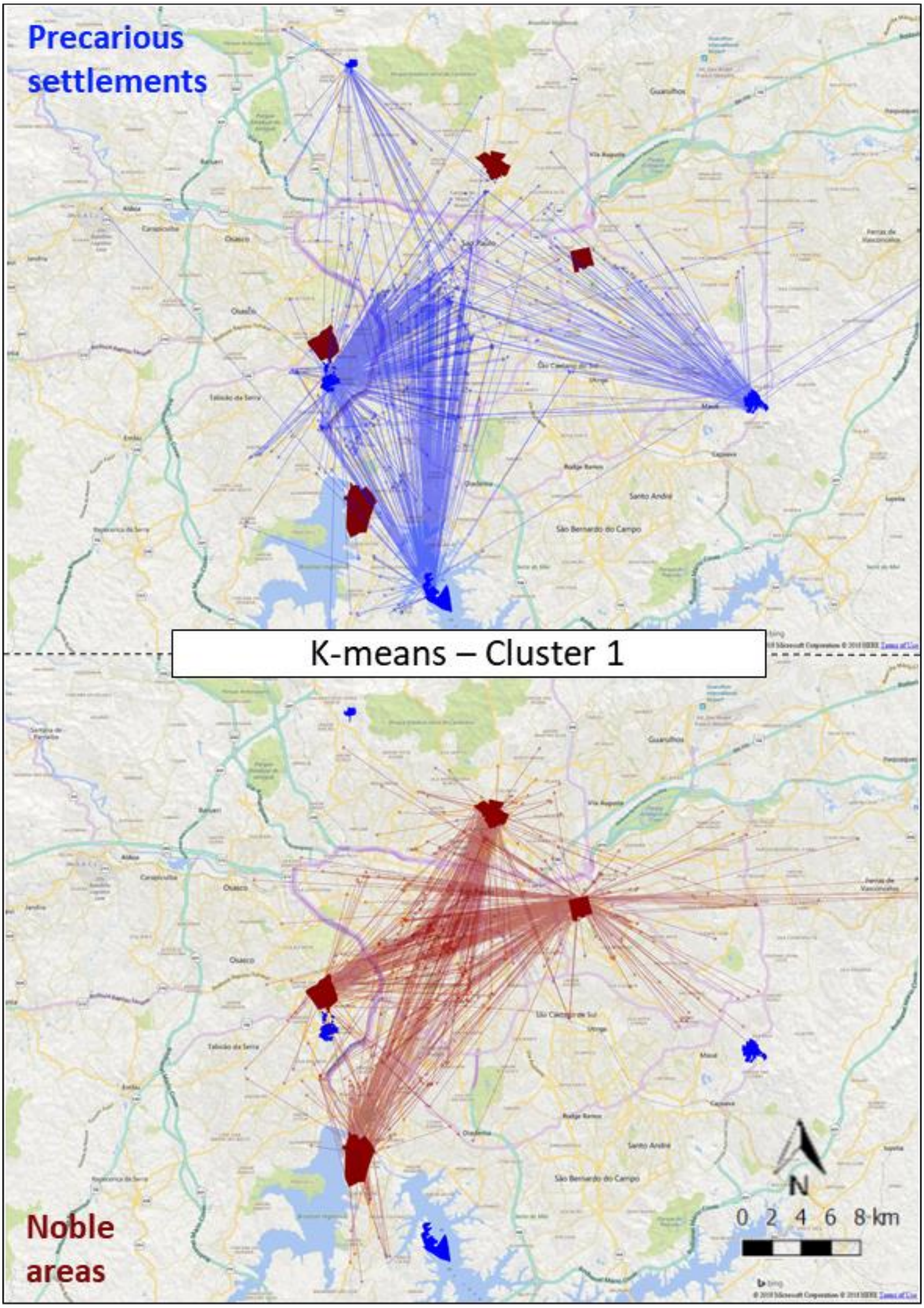
Cluster 4 concentrates its activities in peripheral areas far from the center of São Paulo, near their residence locations (especially for precarious settlement areas). This peripheral distribution is consistent with the features of Cluster 4, formed by passengers with possible local employment and low displacement in the city.

Cluster 5 shows a high concentration of activities in São Paulo downtown, which has a wide supply of commerce and services, and converges with the commuting pattern features found in section 5.2.

Cluster 7, consistent with the statements made in section 5.2, has the highest spread of activities, from short to long distances and throughout the city. Cluster 6 also has a spread distribution, but more focused on nearer locations from the passengers' residences, confirming the pattern of low displacements.

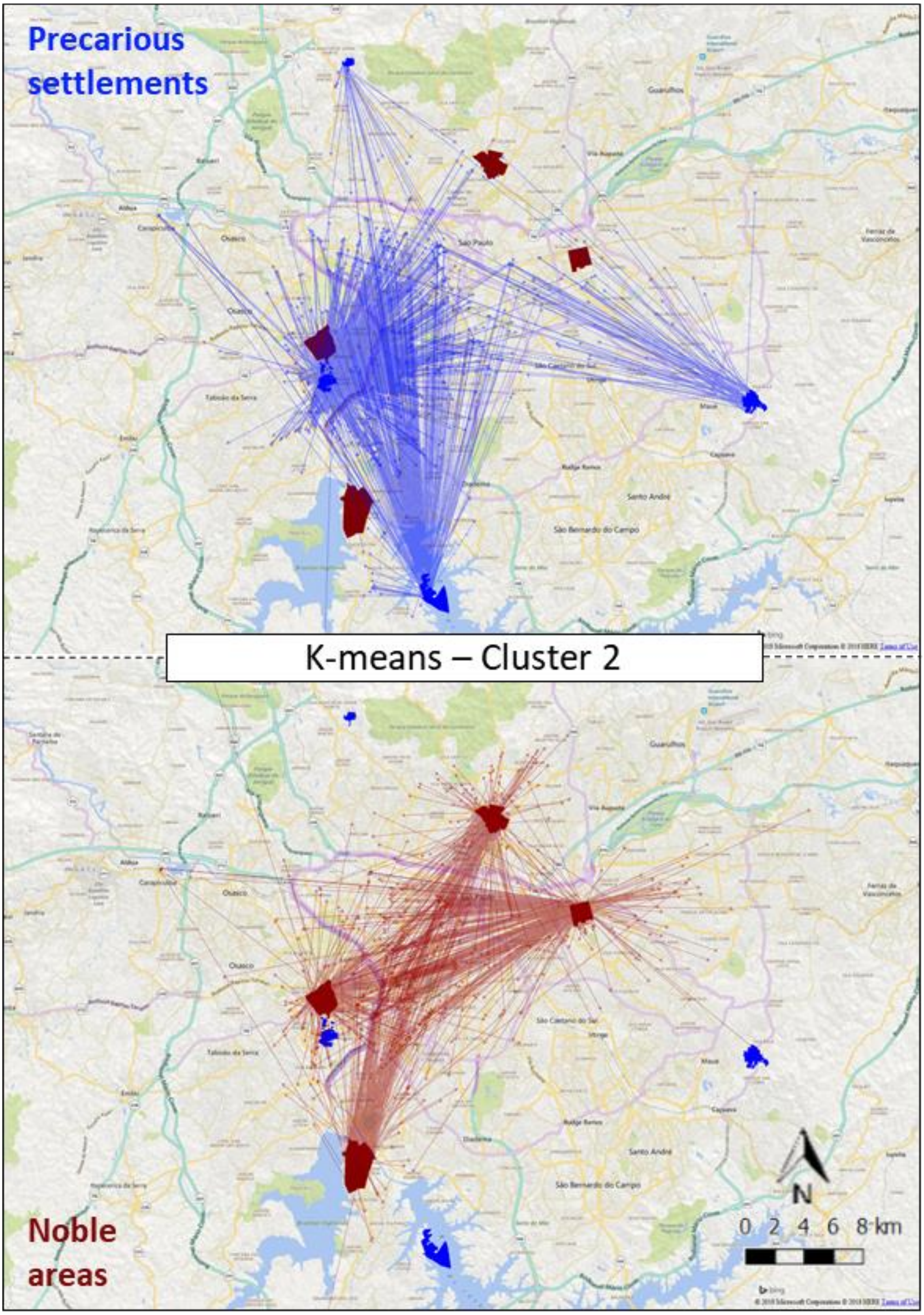
Cluster 8 is the smallest of all, and the particular users in this cluster have the activities located mainly in the center and commercial areas of the city.

Figure 34 – Activity distribution by cluster – K-means / Cluster 1



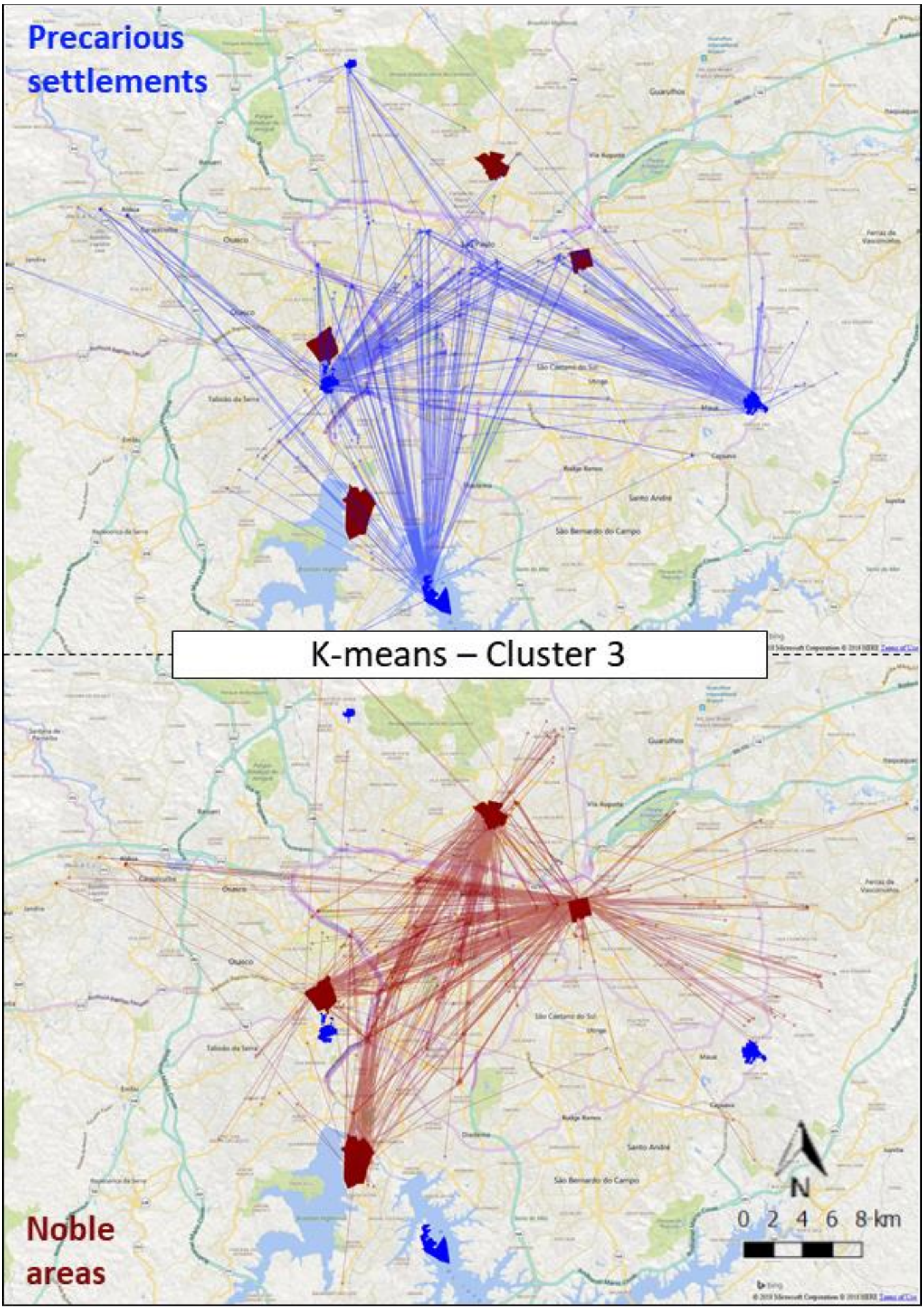
Source: The authors' own elaboration

Figure 35 – Activity distribution by cluster – K-means / Cluster 2



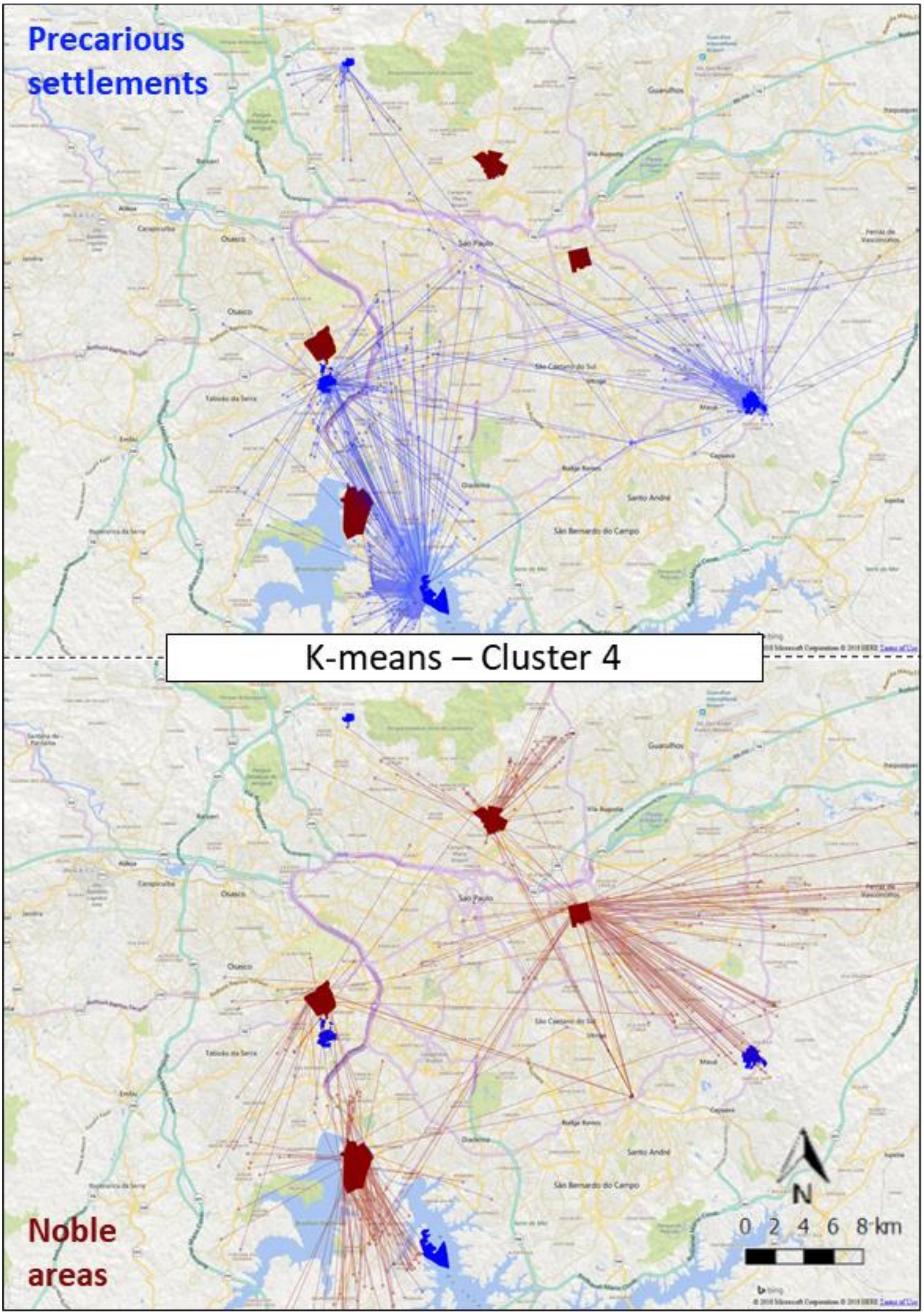
Source: The authors' own elaboration

Figure 36 – Activity distribution by cluster – K-means / Cluster 3



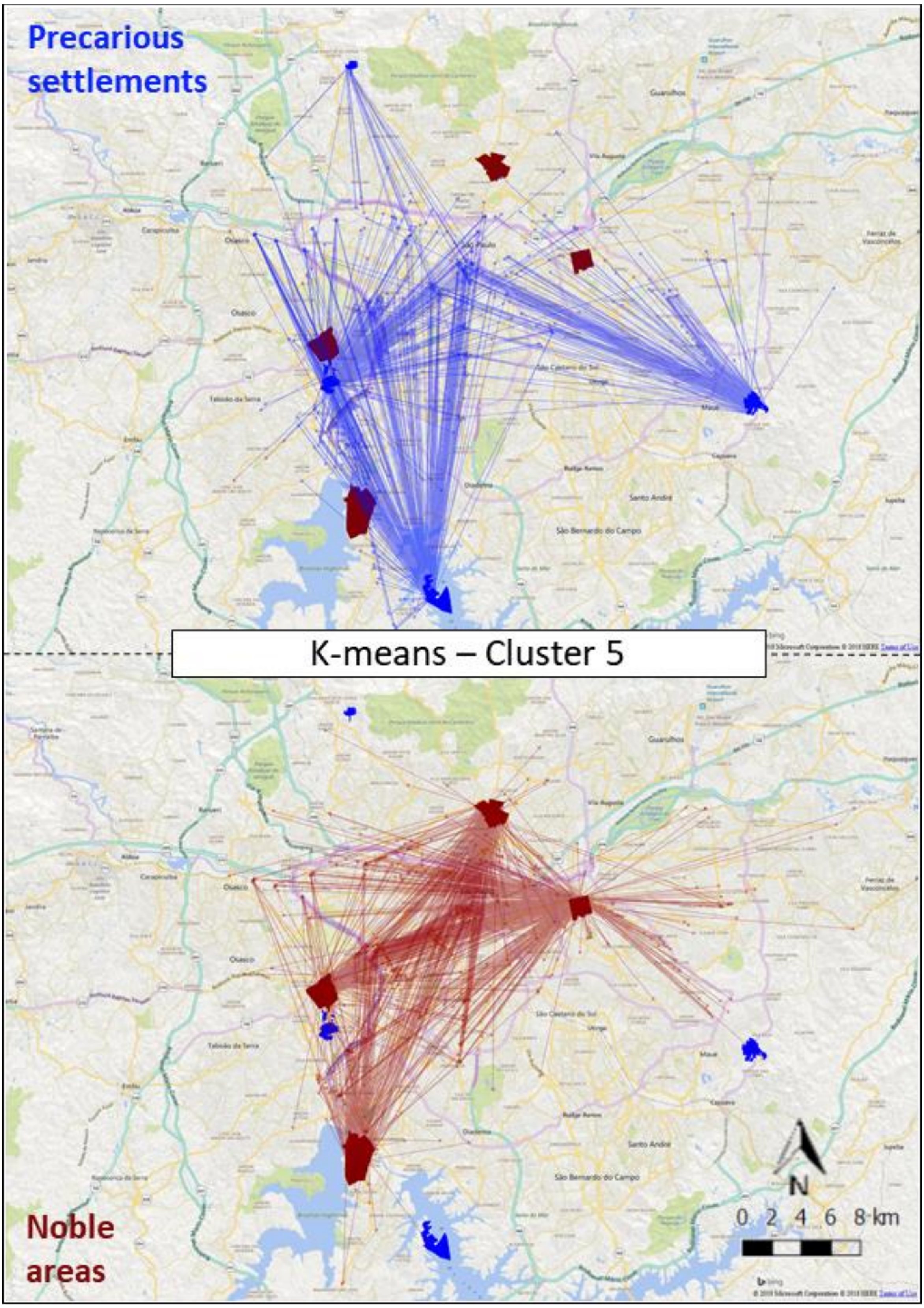
Source: The authors' own elaboration

Figure 37 – Activity distribution by cluster – K-means / Cluster 4



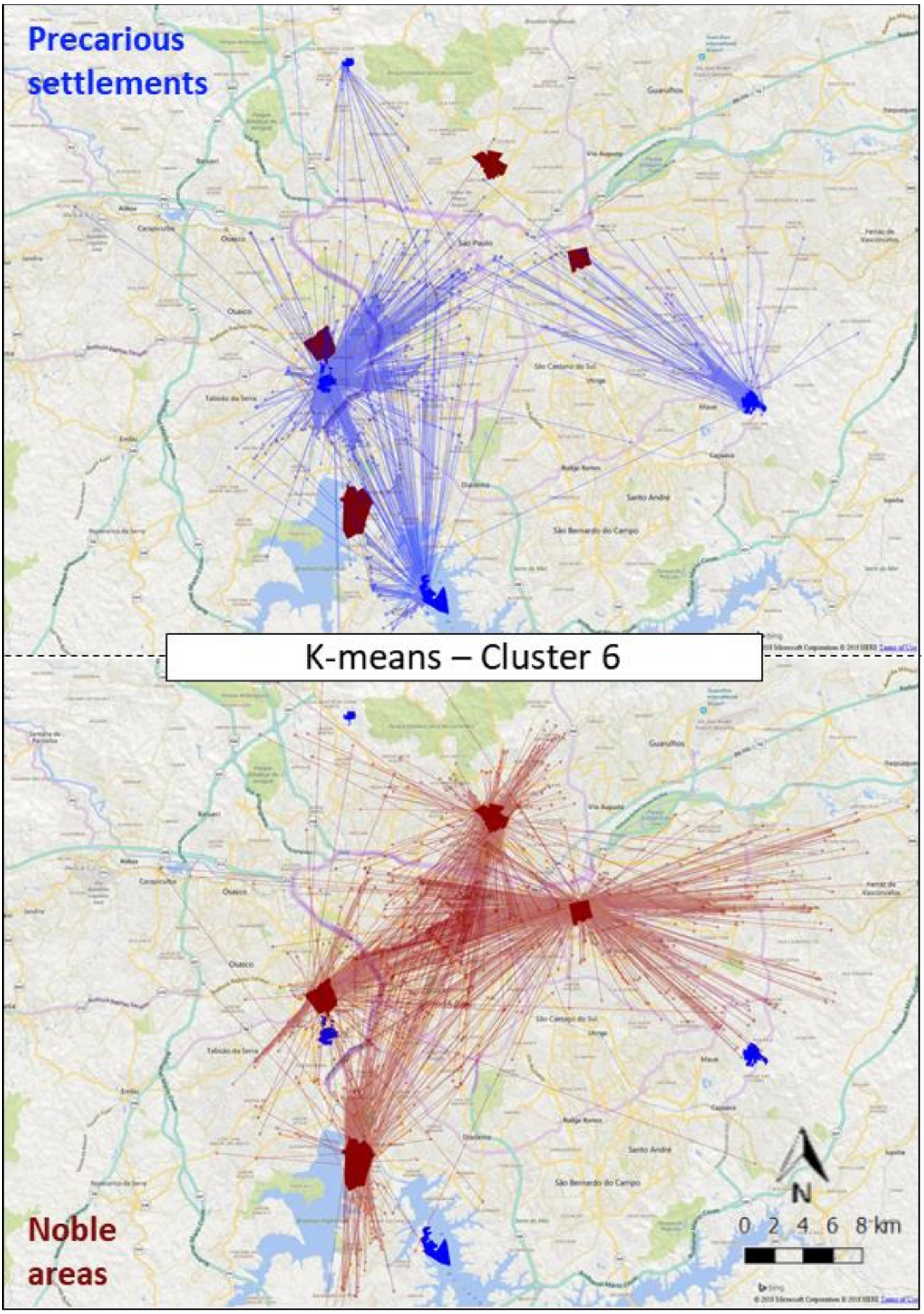
Source: The authors' own elaboration

Figure 38 – Activity distribution by cluster – K-means / Cluster 5



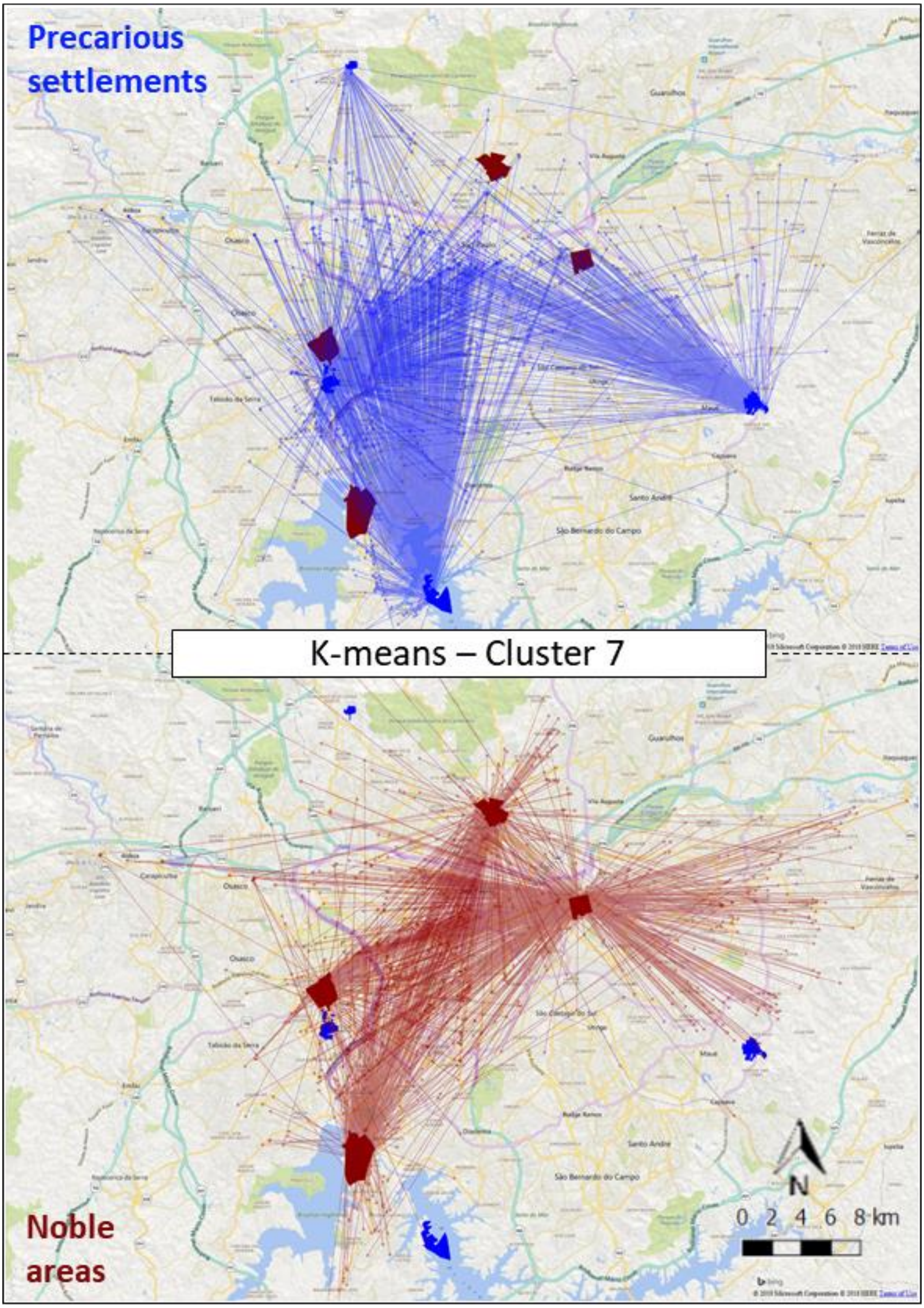
Source: The authors' own elaboration

Figure 39 – Activity distribution by cluster – K-means / Cluster 6



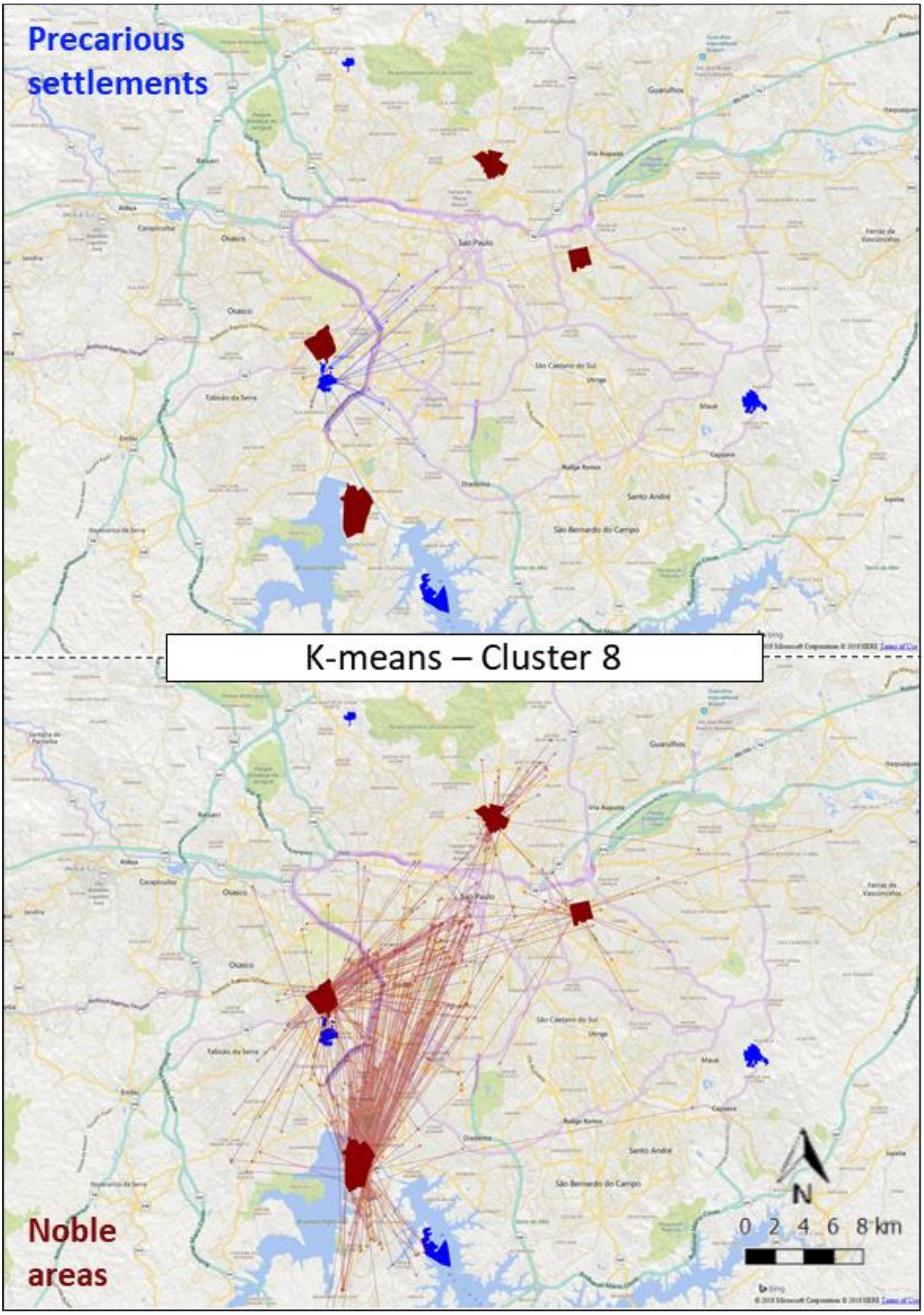
Source: The authors’ own elaboration

Figure 40 – Activity distribution by cluster – K-means / Cluster 7



Source: The authors' own elaboration

Figure 41 – Activity distribution by cluster – K-means / Cluster 8



Source: The authors' own elaboration

5.4 CLUSTER VALIDATION

After performing each clustering algorithm on smart card data, one may wonder whether their results are valid or simply artifacts of the clustering algorithm. Indeed, clustering algorithms may produce misleading results, and different clustering methods may produce different results due to their different assumptions and methodologies applied to the data (LEGENDRE; LEGENDRE, 2012). Therefore, clustering validation is performed in this section.

Clustering validation can be considered key to the success of clustering applications, and are usually classified into three approaches: internal, external and relative. External criteria evaluate the results of a clustering algorithm based on external information not present in the data, which is imposed to reflect the intuition about the pre-defined clustering structure. Internal criteria only rely on information already in the data set, evaluating a clustering structure without any external/additional information. Relative criteria compare a clustering structure to other clustering schemes yielded by the same algorithm but with different parameter values. In practice, external information is often not available, as in the case of smart card data. Therefore, internal validation measures are here selected for cluster validation (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001; LIU et al., 2010; MAULIK; BANDYOPADHYAY, 2002).

Post-hoc significance levels (such as ANOVA tests) are not recommended to be performed in clustering algorithms. This means that the same structuring of smart card data should not be used either to perform clustering or to evaluate significant differences between the observations in each cluster. As the clustering results were derived based on criteria to maximize their separation, even if there is no actual structure in the data, the clustering algorithm will impose one. By grouping nearby points, the clustering algorithm will decrease the within-group variance and increase the across-group variance, biasing the results towards false positives.

Therefore, we here present the basic concepts of four internal clustering validation measures – the Calinski-Harabasz index, the Dunn's index, the Davies-

Bouldin index, and the *S_Dbw* index – and apply them to the clustering techniques results from the former chapter to compare their performances.

The Calinski-Harabasz index (*CH*) is analogous to an F-ratio in ANOVA, evaluating the cluster validation based on the average between and within-cluster sum of squares:

$$CH = \frac{[traceB/(K-1)]}{[traceW/(n-K)]} \text{ , where}$$

n: total number of elements;

K: number of clusters;

$traceB = \sum_{k=1}^{K} n_k \|z_k - z\|^2$;

$traceW = \sum_{k=1}^{K} \sum_{i'=1}^{n_k} \|x_i - z_k\|^2$;

n_k: number of elements in cluster *k*;

z: global centroid;

z_k: centroid of cluster *k*;

x_i: data point in cluster *i*.

The Dunn's index (*D*) uses the minimum pairwise inter-cluster separation and the maximum cluster size among all clusters as the intra-cluster compactness.

$$v_D = \underbrace{min}_{1\le i\le K} \left\{ \underbrace{min}_{1\le i\le K, j\neq i} \left\{ \frac{\delta(C_i, C_j)}{\underbrace{max}_{1\le k\le K}\{\Delta(C_k)\}} \right\} \right\} \text{ , where}$$

$\Delta(C_k) = \underbrace{max}_{x,y\in S} \{d(x,y)\}$: size of the cluster *k* and
$\delta(C_i, C_j) = \underbrace{min}_{x\in S, y\in T} \{d(x,y)\}$: distance between clusters *i* and *j*.

The Davies-Bouldin index (*DB*) is similar to the Dunn's index, relating the average distance of elements of each cluster to their respective centroids to the distance of the centroids of the two clusters:

$$DB = \frac{1}{K} \sum_{i=1}^{K} R_i \text{ , where}$$

K: number of clusters;

$R_i = max_{j,j\neq i} \left\{ \frac{S_i + S_j}{d_{ij}} \right\}$;
$S_i = \frac{1}{|n_i|} \sum_{x\in C_i} \{ \|x_j - z_i\| \}$, the scatter within cluster *i*;

n_i: size of the cluster *i;*
x_j: data point in cluster *j;*
z_i: centroid of cluster *i*;
$d_{ij} = \|z_i - z_j\|$, distance between cluster C_i and C_j.

The S_Dbw index (*S_Dbw*) takes compactness to the intra-cluster variance and density to the inter-cluster separation. The index is the summation of these two terms:

$$Scat(NC) + Dens\_bw(NC) \text{ , where}$$

$$Scat(NC) = \frac{1}{NC}\sum\nolimits_i \|\sigma(C_i\| \,/\, \|\sigma(D\|$$

$$Dens\_bw(NC) = \frac{1}{NC(NC-1)}\sum\nolimits_i \left[\sum\nolimits_{j,j\neq i} \frac{\sum_{x\in C_i \cup C_j} f(x,u_{ij})}{max\left\{\sum_{x\in C_i} f(x,c_i)\, , \sum_{x\in C_j} f(x,c_j)\right\}}\right]$$

D: data set;
NC: number of clusters;
$\sigma(C_i)$: variance vector of C_i;
$\sigma(D)$: variance vector of D;
C_i: the *i*–th cluster;
c_i: center of C_i;
$f(x, u_{ij})$: density function.

Table 7 presents the internal validation measures notations and their optimal value criteria.

Table 7 – Percentage of users in the same cluster by clustering method

Measure	Notation	Optimal value
Calinski-Harabasz index	*CH*	Max
Davies-Bouldin index	*DB*	Min
Dunn's indices	*D*	Max
S_Dbw validity index	*S_Dbw*	Min

Source: The authors' own elaboration

Table 8 presents the results of the four internal measures evaluated for each clustering algorithm. The K-means clustering outperforms the other two algorithms in three out of the four measures - Calinski-Harabasz, Davies-Bouldin, and S_Dbw. The SOM algorithms perform better for the Dunn index.

Table 8 – Clustering performance for each internal validation index

Measure	K-means	TwoStep	SOM
Calinski-Harabasz	**3307.51**	878.97	3216.89
Davies-Bouldin	**16.40**	21.76	178.00
Dunn index	1.04E-05	7.47E-06	**2.41E-05**
S_Dbw	**15.75**	21.44	17.42

Source: The authors' own elaboration

6 CONCLUSIONS AND FUTURE DIRECTIONS

In this study, we evaluated spatiotemporal patterns from precarious settlement residents by using three different clustering algorithms to the database: K-means, TwoStep, and SOM. The groups are described and evaluated by their similarities and differences, and we also evaluated the feasibility of identifying low-paid employees' travel patterns using the proposed methodology.

We begin reviewing the current literature regarding the uses of smart card data in public transport researches and some travel pattern analysis with smart card data. Afterward, we describe the dataset used and the regions of analysis selected – Paraisópolis, Cantinho do Céu, Parque Taipas and São Francisco Global representing precarious settlements; and Vila Sônia, Parque Interlagos, Jardim São Paulo and Vila Gomes Cardim representing middle-class areas. The idea was to compare areas with residents from different social classes and to evaluate their differences and similarities in travel patterns.

The methodological approach regards a preprocessing step, with the spatial inference for each transaction, the filtering of non-frequent travelers, selecting only adult and student card types, the inference of residence for each cardholder, the inference of activities and transfers, filtering out transfers, and the structuration of the database for the travel pattern clustering. The subsequent step explains each clustering technique chosen to perform, and the calculation of the optimal number of clusters, which resulted in eight.

After the smart card processing, we present the results of the clustering methods. First, results derived from the preprocessing step. Afterward, an evaluation of the eight clusters regarding their features and spatial distribution. Finally, a cluster validation is calculated to evaluate the performance of each clustering algorithm.

Despite an extensive literature regarding the uses of smart card data, not much effort has been invested with the focus on low income or precarious settlements residents, which is a contribution from the present work. It was possible to compare areas with residents from four different precarious settlements in São Paulo with other four from middle-class residents, and we could state from the clustering results that their residents indeed differ when analyzing daily travel patterns. Exploring travel patterns in areas of precarious settlements brought a better understanding of their mobility characteristics.

From the results, apart from clusters with evidence of commuting passengers working in areas of regular labor supply concentration, we could clearly state that at least two clusters had features that suggest an association with low-paid employment. These clusters could be further investigated in future works to help transit authorities to re-evaluate their current services for this population and to provide them with a more suitable and reliable service, or even at a strategic level of transportation planning, with new services for the area according to their destinations. New services could be implemented not necessarily in an axial direction, but towards medium or high-income areas in the surroundings, or even with short distance shuttle services supplying their own region, perhaps with peripheral/circular routes.

The use of three different clustering algorithms for data classification introduced more robustness and reliability to the results, considering that each of the methods has different criteria and assumptions of use, and the results were consistent for the three, with similar clusters formed. These results can imply that they are consistent clusters and not artifacts of the natural sampling variation. Also, the studied areas had relatively similar travel patterns and regularity comparing one to another, and that indicates a regularity in behavior between precarious settlement residents.

The use of DBSCAN to infer residences is also a contribution not previously discussed in the literature. The use of only the main cluster formed from the first validations of each user to calculate the centroid of their residence brings an alternative methodology to this inference, aiming to filtering out eventual outlier validations, and pointing out to an interesting and better result when comparing to the traditional centroid calculation per se.

This work also points out that it is possible to extract semantic meaning from smart card data. From the data processing and clustering results we were able to infer important information about the transit passengers, such as income, land use characteristics, distance, duration and frequency patterns, and the type of work of the user – being low-paid employed or not. These inferences may suggest that smart card data could serve as a complement to conventional travel behavior surveys.

We believe that numerous perspectives arise from this research. First, the time between the user's entry into the bus and the actual crossing of the bus ticket gate is not an easy prediction, and was not considered here for simplification purposes. São Paulo buses have front seats and it can take several minutes for passengers to effectively cross the ticket gate, distorting their travel validation. Also, there was no approach here regarding night shifts of employment. Some premises made during the development of this work, such as using the first boarding of the day for each user to infer their residences, do not consider these night jobs. Including weekends in the analyses to compare the differences in travel patterns with business days would also be a matter to be explored in future works. A deeper evaluation of student card type during the weeks of analysis, especially regarding their evolution of validations during the period of school vacation (in July) could also be investigated.

An additional important issue of future studies regards the inference of land use for each validation, apart from the first of the day. The type of land use can easily vary within a small area, changing the inference made for each validation, and there is a significant variability in the spatial location of validations. An initial idea would be creating a buffer for each validation and assigning the land use with most appearance or with the largest area within the buffer, for example, but much still needs to be explored.

Finally, further works should link the results found here with the performance of the transit network or make service adjustments considering to the spatial analysis here presented, evaluating the public transport accessibility level and detecting if one of the precarious settlement areas are underserved. Based on these conclusions, we hope that transportation policies can also be improved.

7 REFERENCES

AGARD, B.; MORENCY, C.; TRÉPANIER, M. Mining Public Transport User Behaviour From Smart Card Data. **IFAC Proceedings Volumes**, v. 39, n. 3, p. 399–404, 2006.

AGARD, B.; TRÉPANIER, M. Assessing Public Transport Travel Behaviour form Smart Card Data with Advanced Data Mining. **13th WCTR**, 2013.

ALSGER, A.; TAVASSOLI, A.; MESBAH, M.; FERREIRA, L.; HICKMAN, M. 2018. Public transport trip purpose inference using smart card fare data. **Transportation Research Part C: Emerging Technologies,** v. 87, p. 123–137, 2018.

AMAYA, M.; CRUZAT, R.; MUNIZAGA, M. A. Estimating the residence zone of frequent public transport users to make travel pattern and time use analysis. **Journal of Transport Geography**, v. 66, p. 330–339, 2018.

ANDA, C.; FOURIE, P.; ERATH, A. Transport Modelling in the Age of Big Data. **Singapore - ETH Centre: Future Cities Laboratory**, v. Work Report, 2016.

ARBEX, R. O.; ALVES, B. B.; GIANNOTTI, M. A. Comparing Accessibility in Urban Slums Using Smart Card and Bus GPS Data. **Transportation Research Board TRB 95th Annual Meeting**, Washington DC. DVD Compendium, v. 1, 2016.

ARBEX, R. O.; CUNHA, C. B. DA. Estimação da matriz origem-destino e da distribuição espacial da lotação em um sistema de transporte sobre trilhos a partir de dados de bilhetagem eletrônica. **Transportes**, v. 25, n. 5, 2017.

BARRY, J.; NEWHOUSER, R.; RAHBEE, A.; SAYEDA, S. Origin and destination estimation in New York City with automated fare system data. **Transportation Research Record: Journal of the Transportation Research Board**, v. 1817 (-1), p. 183–187, 2002.

BARROSO, L. P.; ARTES, R. Análise Multivariada. **1. ed. Lavras: Região Brasileira da Sociedade Internacional de Biometria**, v. 1, p. 150, 2003.

BACHER, J., WENZIG, K.; VOGLER, M. (2004). SPSS twostep cluster: A first evaluation. (Arbeits- und Diskussionpapiere. 2, 2) Erlange—Nurnberg, University of Friedrich—Alexander, Chair of Sociology. Available from: <https://www.ssoar.info/ssoar/bitstream/handle/document/32715/ssoar-2004-bacher_et_al-SPSS_TwoStep_Cluster_-_a.pdf?sequence=1>. Accessed Apr. 18, 2017. Accessed Jun. 09, 2018.

BACAO, F.; LOBO, V.; PAINHO, M. Self-organizing Maps as Substitutes for K-Means Clustering. **Lecture Notes in Computer Science, SpringerLink**, p. 476–483, 2005.

BAGCHI, M.; WHITE, P. R. The potential of public transport smart card data. **Transport Policy**, v. 12(5), n. 9, p. 464–474, 2005.

BAILEY, K. D. Typologies and Taxonomies: An Introduction to Classification Techniques. **Sage Publications, Thousand Oaks, CA**, 1994.

BLYTHE, P. Improving public transport ticketing through smart cards. **Proceedings of the Institution of Civil Engineers, Municipal Engineer**, v. 157, p. 47–54, 2004.

BOUMAN, P.; VAN DER HURK, E.; KROON, L.; LI, T.; VERVEST, P. Detecting activity patterns from smart card data. **25th Benelux Conference on Artificial Intelligence**, 2013.

BRASIL. Ministério das Cidades. Secretaria Nacional de Habitação. (2010). **Guia para o mapeamento e caracterização de assentamentos precários.** Brasília: Ministério das Cidades.

BRIAND, A. S.; CÔME, E.; TRÉPANIER, M.; OUKHELLOU, L. Analyzing year-to-year changes in public transport passenger behaviour using smart card data. **Transportation Research Part C: Emerging Technologies**, v. 79, p. 274–289, 2017.

CARNEIRO, C. C. Self-Organizing Maps – SOM (Mapas Auto-Organizáveis). **Lecture for the discipline “PTR-5922 – Análise Espacial Aplicada a Transportes” at Polytechnic School at University of São Paulo**. 2015. 63 slides.

CERIN E.; LESLIE, E.; DU TOIT, L.; OWEN, N.; FRANK, L. D. Destinations that matter: associations with walking for transport. **Health Place**, v. 13, n. 3, p. 713–724, 2007.

CHIU, T.; FANG, D.; CHEN, J.; WANG, Y. JERIS, C. A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. **KDD, San Francisco, CA**, 2001.

DELBOSC, A. Why write well? **Transport Reviews**, v. 37, n. 5, p. 545–550, 2017.

DESGRAUPES, B. Clustering Indices. **University Paris Ouest**, **Lab Modal'X**, 2017. Available in: <https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf>. Accessed Aug. 18, 2018.

DEVILLAINE, F.; MUNIZAGA, M.; TRÉPANIER, M. Detection of Activities of Public Transport Users by Analyzing Smart Card Data. **Transportation Research Record: Journal of the Transportation Research Board**, n. 2276, p. 48–55, 2012.

EL MAHRSI, M. K.; CÔME, E.; BARO, J.; OUKHELLOU, L. Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data. **The 3rd International Workshop on Urban Computing (UrbComp 2014)**, 2014.

ESTER, M.; KRIEGEL, H-P.; SANDER, J.; XU, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. **Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining**, p. 226–231, 1996.

FAHAD, A.; ALSHATRI, N.; TARI, Z.; ALAMRI, A.; KHALIL, I.; ZOMAYA, A.; FOUFOU, S.; BOURAS, A. A survey of clustering algorithms for big data: taxonomy and empirical analysis. **IEEE Trans Emerg Topics Comp**. v. 2 (3), p. 267–79, 2014.

FARZIN, J. M. Constructing an Automated Bus Origin-Destination Matrix Using Farecard and Global Positioning System Data in São Paulo, Brazil. **Transportation Research Record: Journal of the Transportation Research Board**, v. 2072 (-1), p. 30–37, 2008.

FONTES, A.; PERO, V.; BERG, A. J. Low-paid employment in Brazil. **International Labour Review**, v. 151, n. 3, p. 193–219, 2012.

FORGY, E. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. **Biometrics**, v. 21, p. 768–769, 1965.

FONZONE, A.; SCHMOCKER, J.-D.; KURAUCHI, F.; HASSAN, S. M. Strategy Choice in Transit Networks. **Journal of the Eastern Asia Society for Transportation Studies**, v. 10, p. 796-815, 2013.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. The elements of statistical learning. **Springer Series in Statistics**, New York, v. 1, 2001.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On Clustering Validation Techniques. **Intelligent Information Systems**, 2001.

HALKIDI, M.; VAZIRGIANNIS, M. Clustering validity assessment: Finding the optimal partitioning of a data set. **IEEE ICDM**, San Jose, CA, p. 187–194, 2001.

HAN, G.; SOHN, K. Clustering the Seoul Metropolitan Area by Travel patterns based on a Deep Belief Network. **3rd MEC International Conference on Big Data and Smart City,** 2016.

HANSON, S.; HUFF, J. Classification issues in the analysis of complex travel behavior. **Transportation**, v. 13, p. 271–293, 1986.

HAH, J.; KAMBER, M. **Data Mining: Concepts and Techniques.** Morgan Kaufmann, 2000.

HATZICHRISTOS, T. Delineation of demographic regions with GIS and computational intelligence. **Environment and Planning B: Planning and Design**, v. 31, p. 39–49, 2004.

HIMANEN V.; JARVI-NYKANEN T.; RAITIO J. Daily travelling viewed by self-organizing maps. **Neural Networks in Transport Applications**, p. 85–110, 1998.

HOOGENDOORN, S. P.; HOOGENDOORN-LANSER, S. Neural networks approach to travel behavior in public transport networks. **TRB Annual Meeting, Washington DC: Transportation Research Board**, 2001.

IBM (2011). IBM SPSS Statistics 20 Algorithms.

IBM, (2016). IBM SPSS Statistics Base 24.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Censo demográfico 2010: Aglomerados subnormais - informações territoriais**. Censo demográfico., Rio de Janeiro, p. 1–251, 2010.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Censo demográfico 2010.** Available in <https://censo2010.ibge.gov.br/>. Accessed Jan. 26, 2018.

JAIN, A.; DUBES, R. **Algorithms for clustering data**. Prentice Hall advanced reference series. Prentice Hall, 1988.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys**, v. 31, n. 3, p. 264–323, 1999.

JAMES, G.; WITTEN. D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning: With Applications in R**. Springer, New York, 2013.

JOSEPH, J.; TORNEY, C.; KINGS, M.; THORNTON, A.; MADDES, J. Applications of machine learning in animal behavior studies. **Animal Behavior**. V. 124, n. December, p. 203–220, 2016.

JUN, C.; DONGYUAN, Y. Estimating smart card commuters origin-destination distribution based on APTS data. **Journal of Transportation Systems Engineering and Information Technology**, v. 13, n. 4, p. 47–53, 2013.

KALAFTIS, M. G.; VLAHOGIANNI, E. I. Statistics versus neural networks in transportation research: differences, similarities and some insights. **Transportation Research Part C Emerging Technologies**, v. 19 (3), p. 387-399, 2011.

KHALILZADDEH, J.; TASCI, D. A. Large sample size, significance level, and the effect size: Solutions to perils of using big data for academic research. **Tourism Management**, v. 62 (5), p. 89–96, 2017.

KIEU, L. M.; BHASKAR, A.; CHUNG, E. A modified Density-Based Scanning Algorithm with Noise for spatial travel pattern analysis from Smart Card AFC data. **Transportation Research Part C: Emerging Technologies**, v. 58, p. 193–207, 2015.

KIEU, L. M.; BHASKAR, A.; CHUNG, E. Mining temporal and spatial travel regularity for transit planning. **Australasian Transport Research Forum 2013 Proceedings**, 2013.

KOHONEN, T. **Self-Organizing Map**. Springer, Berlin, Heidelberg. 1995. (Third, Extended Edition, 2001).

KON, A. "Diversidades nas condições de informalidade do trabalho brasileiro". **Anais do XXXII Encontro Nacional de Economia**, ANPEC, João Pessoa, 2004.

KOUA E. L.; KRAAK M.-J. Alternative visualization of large geospatial datasets. **The Cartographic Journal**, v. 4, p. 217–228, 2004.

KREIN, J. D.; PRONI, Marcelo W. Economia informal: aspectos conceituais e teóricos. Brasília: **OIT- Brasil** (Trabalho decente no Brasil; Documento de trabalho, n. 4), 2010.

LANGLOIS, G.; KOUTSOPOULOS, H. N.; ZHAO, J. Inferring patterns in the multi-week activity sequences of public transport users. **Transportation Research Part C: Emerging Technologies**, v. 64, p. 1–16, 2016.

LATHIA, N.; CAPRA, L. How smart is your smartcard? Measuring travel behaviours, perceptions, and incentives. **Proceedings of the 13th International Conference on Ubiquitous Computing**, UbiComp 2011, p. 291–300, 2011.

LATHIA, N.; QUERCIA, D.; CROWCROFT, J. The hidden image of the city: Sensing community well-being from urban mobility. **Proceedings of the 10th International Conference on Pervasive Computing**, p. 91–98, 2012.

LATHIA, N.; SMITH, C.; FROEHLICH, J.; CAPRA, L. Individuals among commuters: Building personalized transport information services from fare collection systems. **Pervasive and Mobile Computing**, v. 9(5), p. 643–664, 2013.

LEE, S. G.; HICKMAN, M. Trip purpose inference using automated fare collection data. **Public Transport**, v. 6, n. 1, p. 1–20, 2013.

LEGENDRE, P. & LEGENDRE, L. **Numerical Ecology (Developments in Environmental Modelling)**, Elsevier, 1998.

LIN, M.; LUCAS, H. C.; SHMUELI, G. Too big to fail: Large samples and the p-value problem. **Information Systems Research**, v. 24(4), p. 906-917, 2013.

LIU, Y. LI.; Z. XIONG, H.; GAO, X.; WU, J. Understanding of internal clustering validation measures. **IEEE ICDM**, p. 911–916, 2010.

LONG, Y.; THILL, J. C. Combining smart card data and household travel survey to analyze jobs-housing relationships in Beijing. **Computers, Environment and Urban Systems**, v. 53, p. 19–35, 2015.

LYNN, S. Self-organizing maps for customer segmentation. **Talk for Dublin R users group**. Jan. 20, 2014. 49 slides. Available in <https://www.slideshare.net/shanelynn/2014-0117-dublin-r-selforganising-maps-for-customer-segmentation-shane-lynn>. Accessed Apr. 16, 2018.

MA, X.; LIU, C.; WEN, H.; WANG, Y.; WU, Y. J. Understanding commuting patterns using transit smart card data. **Journal of Transport Geography**, v. 58, p. 135–145, 2017.

MA, X.; WU, Y. J.; WANG, Y.; CHEN, F.; LIU, J. Mining smart card data for transit riders' travel patterns. **Transportation Research Part C**, v. 36, p. 1–12, 2013.

MARQUES, E. C. L.; SARAIVA, C. Urban integration or reconfigured inequalities? Analyzing housing precarity in São Paulo, Brazil. **HABITAT INTERNATIONAL**, v. 69, p. 18-26, 2017.

MAULIK, U.; BANDYOPADHYAY, S. Performance evaluation of some clustering algorithms and validity indices. **IEEE PAMI**, v. 24, p. 1650–1654, 2002.

MCNALLY, M. G.; KULKARNI, A. Assessment of influence of land use–Transportation system on travel behavior. **Transportation Research Record**. v. 1607, p. 105–115, 1997.

MORENCY, C.; TREPANIER, M.; AGARD, B. Analysing the Variability of Transit Users Behaviour with Smart Card Data. **2006 IEEE Intelligent Transportation Systems Conference**, p. 44–49, 2006.

MORENCY, C.; TRÉPANIER, M.; AGARD, B. Measuring transit use variability with smart-card data. **Transport Policy**, v. 14, n. 3, p. 193–203, 2007.

MORENO, D.; MARCO, P., & OLMEDA, I. Self-organizing maps could improve the classification of Spanish mutual funds. **European Journal of Operational Research**, v. 147, p. 1039–1054, 2006.

MOSTAFA, M. M. Clustering the ecological footprint of nations using Kohonen's self-organizing maps. **Expert Systems with Applications,** v. 37, p. 2747–2755, 2010.

MOURA, Rodrigo Leandro de; BARBOSA FILHO, Fernando Holanda. Evolução Recente da Informalidade no Brasil: Uma Análise segundo Características da Oferta e Demanda de Trabalho. In: **Encontro Nacional de Economia**, 41, 2012, Rio de Janeiro. Anais. Foz do Iguaçu: ANPEC, 2013.

MUNIZAGA, M. A.; PALMA, C. Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. **Transportation Research Part C: Emerging Technologies**, v. 24, p. 9–18, 2012.

NAMRATHA, M.; PRAJWALA, T. R. A Comprehensive Overview of Clustering Algorithms in Pattern Recognition. **Journal of Computer Engineering,** v. 4, n. 6, p. 23–30, 2012.

OD Pesquisa Origem Destino 2007. (2008). **Síntese das Informações de Pesquisa Domiciliar.** São Paulo: Metrô.

ORTEGA-TONG, M.A., 2013. Classification of London’s Public Transport Users Using Smart Card Data, Thesis MIT. Available in <http://dspace.mit.edu/handle/1721.1/82844>. Accessed Nov. 15, 2017.

ORTÚZAR, J. D.; WILLUMSEN, L. G. **Modelling Transport**. 4. ed. Chichester: John Wiley and Sons, 2011.

PAS. E.I. Weekly travel-activity behavior. **Transportation** v. 15, p. 89–109, 1988.

PELLETIER, M.-P.; TRÉPANIER, M.; MORENCY, C. Smart card data use in public transit: A literature review. **Transportation Research Part C: Emerging Technologies**, v. 19, n. 4, p. 557–568, 2011.

PENDYALA, R.; PARASHAR, A.; MUTHYALAGARI, G. Measuring day-to-day variability in travel characteristics using GPS data. **Proceedings of the 79th Annual Meeting of the Transportation Research Board**, 2000.

PITOMBO, C.S.; KAWAMOTO, E., SOUSA, A.J. An exploratory analysis of relationships between socioeconomic, land use, activity participation variables and travel patterns. **Transport Policy**, v. 18 (2), p. 347–357, 2011.

Quantum GIS Development Team (2016). Quantum GIS Geographic Information System. **Open Source Geospatial Foundation Project**. Available in <http://qgis.osgeo.org>. Accessed Apr. 18, 2017.

QUEIROZ FILHO, A. P. DE. As definições de assentamentos precários e favelas e suas implicações nos dados populacionais: abordagem da análise de conteúdo. **Revista Brasileira de Gestão Urbana**, v. 7, n. 3, p. 340–353, 2015.

R Core Team. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria. URL https://www.R-project.org, 2018.

ROUSSEEUW, P. J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. **Computational and Applied Mathematics**, v. 20, p. 53–65, 1987.

ROUSSINOV, D. CHEN, H. A scalable self-organizing map algorithm for textual classification: A neural network approach to thesaurus generation. **CC-AI—Commun., Cogn. Artif. Intell**., v. 15, p. 81–111, 1998.

SARASWATHI, S.; SHEELA, M. I. A Comparative Study of Various Clustering Algorithms in Data Mining. **International Journal of Computer Science and Mobile Computing**, v.3, n. 11, p. 422–428, 2014.

SHMUELI, D.; SALOMON, I.; SHEFER, D. Neural network analysis of travel behavior: evaluating tools for prediction. **Transportation Research Part C: Emerging Technologies**, v. 4 (3), p. 151-166, 1996.

SILVEN, O.; NISKANEN, M.; KAUPPINEN, H. Wood inspection with nonsupervised clustering. **Machine Vision and Applications**, v. 13, p. 275–285, 2003.

SISODIA, D.; SINGH, L.; SISODIA, S.; SAXENA, K. Clustering Techniques: A Brief Survey of Different Clustering Algorithms. **International Journal of Latest Trends in Engineering and Technology (IJLTET),** v. 1, n. 3, 2012.

SKUPIN A.; HAGELMAN, R. Visualizing demographic trajectories with self-organizing maps. **GeoInformatica**, v. 9, p. 159–179, 2005.

SPTRANS. **São Paulo Transporte (SPTrans)**. Available in <http://www.sptrans.com.br/a_sptrans/>. Accessed Nov. 27, 2017.

TIAN, J.; AZARIAN, M.; PECHT, M. Anomaly Detection Using Self-Organizing Maps-Based K-Nearest Neighbor Algorithm. **European Conference of the Prognostics and Health Management Society 2014**. Nantes, France, 2014.

TRÉPANIER, M.; TRANCHANT, N.; CHAPLEAU, R. Individual trip destination estimation in a transit smart card automated fare collection system. **Journal of Intelligent Transportation Systems**, v. 11(1), p. 1–14, 2007.

UTSUNOMIYA, M.; ATTANUCCI, J.; WILSON, N. H. M. Potential uses of transit smart card registration and transaction data to improve transit planning, **Transportation Research Record**, v. 1971, p. 119–126, 2006.

VAN WEE, B.; BANISTER, D. How to Write a Literature Review Paper? **Transport Reviews**, v. 36, n. 2, p. 278–288, 2016.

VELICKOV, S.; SOLOMATINE, D. Predictive data mining: practical examples. **2nd Joint Workshop on Applied AI in Civil Engineering, Cottbus, Germany**, 2000.

VESANTO, J.; ALHONIEMI, E. Clustering of the Self-Organizing Map. **IEEE Transactions on Neural Networks**, v. 11, n. 3, p. 586–600, 2000.

WILSON, N. H. M.; ZHAO, J.; RAHBEE, A. The potential impact of automated data collection systems on urban public transport planning. **Schedule-Based Modeling of Transportation Networks, of Operations Research/Computer Science Interfaces Series**, v. 46, p. 1–25, 2009.

WITTEN I. E.; FRANK. E. **Data Mining: Practical Machine Learning Tools and Techniques**. 2ed. Morgan Kaufmann, 2005.

XU, R.; WUNSCH. D. Survey of clustering algorithms. **IEEE Transactions on Neural Nets**, v. 16, n. 3, p. 645–678, 2005.

YAN, J., THILL, J.-C. Visual data mining in spatial interaction analysis with self-organizing maps. **Environment and Planning B**, v. 36, p. 466–486, 2009.

YU, C.; HE, Z. Travel Pattern Recognition using Smart Card Data in Public Transit. **International Journal of Emerging Engineering Research and Technology**, v. 4, n. 7, p. 6–13, 2016.

ZHAO, J.; QU, Q.; ZHANG, F.; XY, C.; LIU, S. Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data. **IEEE Transactions on Intelligent Transportation Systems**, p. 1–12, 2017.

ZHAO, J.; RAHBEE, A.; WILSON, N. H. M. Estimating a rail passenger trip Origin-Destination matrix using Automatic Data Collection Systems. **Computer-Aided Civil and Infrastructure Engineering**, v. 22(5), p. 376–387, 2007.

ZHONG, C.; MANLEY, E.; ARISONA, S. M.; BATTY, M.; SCHMITT, G. Measuring variability of mobility patterns from multiday smart-card data. **Journal of Computational Science**, v. 9, p. 125–130, 2015.

ZHOU, Y.; FANG, Z.; ZHAN, Q.; HUANG, Y.; FU, X. Inferring Social Functions Available in the Metro Station Area from Passengers' Staying Activities in Smart Card Data. **ISPRS International Journal of Geo-Information**, v. 6, n. 12, p. 394, 2017.

ZHOU, J.; MURPHY, E.; LONG, Y. Commuting efficiency in the Beijing metropolitan area: An exploration combining smartcard and travel survey data. **Journal of Transport Geography**, v. 41, p. 175–183, 2014.

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

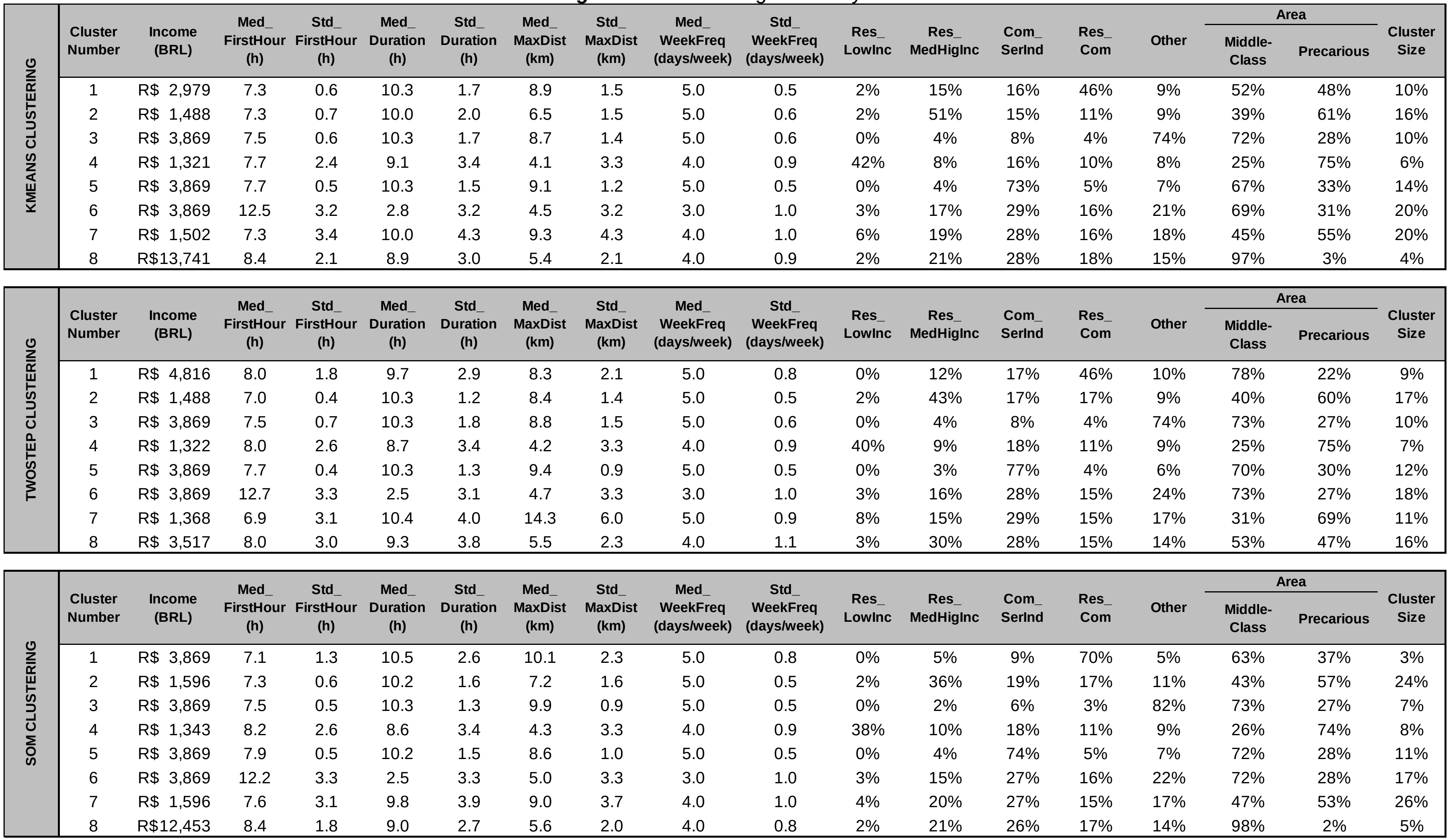
ZHENG, Y.; CAPRA, L.; WOLFSON, O.; YANG, H. Urban computing: Concepts, methodologies, and applications. **Transaction on Intelligent Systems and Technology**, v. 5, n. 3, article 38, 55 pages, 2014.

ZOLTAN, J.; MCKERCHER, B. Analysing intra-destination movements and activity participation of tourists through destination card consumption. **Tourism Geographies**, v. 17 (1), p. 19–35, 2015.

APPENDIX

The clustering results by method (Figure 42), groups formed from TwoStep clustering algorithm (from Figure 43 to Figure 50), groups formed from SOM clustering algorithm (from Figure 51 to Figure 58), activity distribution by cluster from TwoStep algorithm (from Figure 59 to Figure 66) and activity distribution by cluster from SOM algorithm (from Figure 67 to Figure 74).

Figure 42 – Clustering results by method



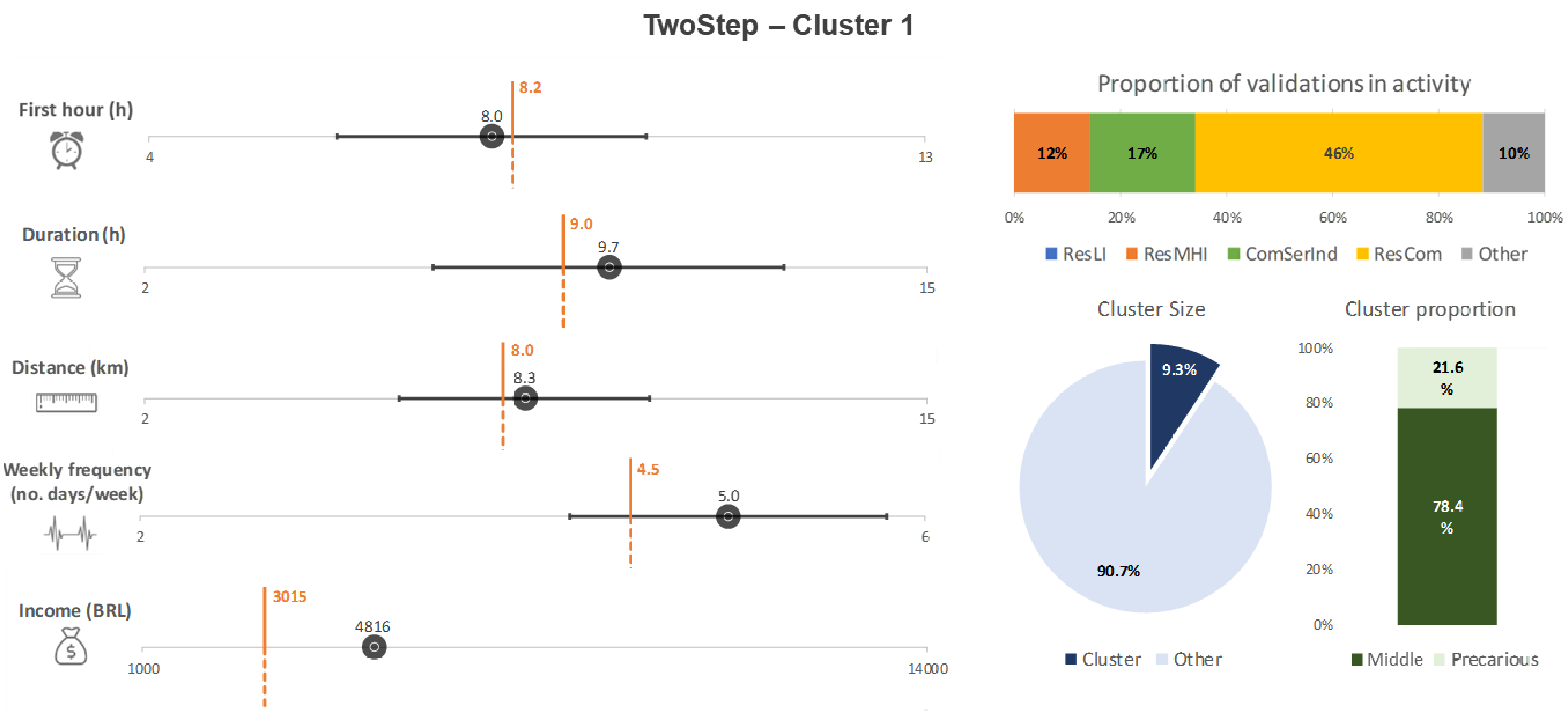
	Cluster Number	Income (BRL)	Med_ FirstHour (h)	Std_ FirstHour (h)	Med_ Duration (h)	Std_ Duration (h)	Med_ MaxDist (km)	Std_ MaxDist (km)	Med_ WeekFreq (days/week)	Std_ WeekFreq (days/week)	Res_ LowInc	Res_ MedHigInc	Com_ SerInd	Res_ Com	Other	Area		Cluster Size
																Middle-Class	Precarious	
KMEANS CLUSTERING	1	R$ 2,979	7.3	0.6	10.3	1.7	8.9	1.5	5.0	0.5	2%	15%	16%	46%	9%	52%	48%	10%
	2	R$ 1,488	7.3	0.7	10.0	2.0	6.5	1.5	5.0	0.6	2%	51%	15%	11%	9%	39%	61%	16%
	3	R$ 3,869	7.5	0.6	10.3	1.7	8.7	1.4	5.0	0.6	0%	4%	8%	4%	74%	72%	28%	10%
	4	R$ 1,321	7.7	2.4	9.1	3.4	4.1	3.3	4.0	0.9	42%	8%	16%	10%	8%	25%	75%	6%
	5	R$ 3,869	7.7	0.5	10.3	1.5	9.1	1.2	5.0	0.5	0%	4%	73%	5%	7%	67%	33%	14%
	6	R$ 3,869	12.5	3.2	2.8	3.2	4.5	3.2	3.0	1.0	3%	17%	29%	16%	21%	69%	31%	20%
	7	R$ 1,502	7.3	3.4	10.0	4.3	9.3	4.3	4.0	1.0	6%	19%	28%	16%	18%	45%	55%	20%
	8	R$13,741	8.4	2.1	8.9	3.0	5.4	2.1	4.0	0.9	2%	21%	28%	18%	15%	97%	3%	4%

	Cluster Number	Income (BRL)	Med_ FirstHour (h)	Std_ FirstHour (h)	Med_ Duration (h)	Std_ Duration (h)	Med_ MaxDist (km)	Std_ MaxDist (km)	Med_ WeekFreq (days/week)	Std_ WeekFreq (days/week)	Res_ LowInc	Res_ MedHigInc	Com_ SerInd	Res_ Com	Other	Area		Cluster Size
																Middle-Class	Precarious	
TWOSTEP CLUSTERING	1	R$ 4,816	8.0	1.8	9.7	2.9	8.3	2.1	5.0	0.8	0%	12%	17%	46%	10%	78%	22%	9%
	2	R$ 1,488	7.0	0.4	10.3	1.2	8.4	1.4	5.0	0.5	2%	43%	17%	17%	9%	40%	60%	17%
	3	R$ 3,869	7.5	0.7	10.3	1.8	8.8	1.5	5.0	0.6	0%	4%	8%	4%	74%	73%	27%	10%
	4	R$ 1,322	8.0	2.6	8.7	3.4	4.2	3.3	4.0	0.9	40%	9%	18%	11%	9%	25%	75%	7%
	5	R$ 3,869	7.7	0.4	10.3	1.3	9.4	0.9	5.0	0.5	0%	3%	77%	4%	6%	70%	30%	12%
	6	R$ 3,869	12.7	3.3	2.5	3.1	4.7	3.3	3.0	1.0	3%	16%	28%	15%	24%	73%	27%	18%
	7	R$ 1,368	6.9	3.1	10.4	4.0	14.3	6.0	5.0	0.9	8%	15%	29%	15%	17%	31%	69%	11%
	8	R$ 3,517	8.0	3.0	9.3	3.8	5.5	2.3	4.0	1.1	3%	30%	28%	15%	14%	53%	47%	16%

	Cluster Number	Income (BRL)	Med_ FirstHour (h)	Std_ FirstHour (h)	Med_ Duration (h)	Std_ Duration (h)	Med_ MaxDist (km)	Std_ MaxDist (km)	Med_ WeekFreq (days/week)	Std_ WeekFreq (days/week)	Res_ LowInc	Res_ MedHigInc	Com_ SerInd	Res_ Com	Other	Area		Cluster Size
																Middle-Class	Precarious	
SOM CLUSTERING	1	R$ 3,869	7.1	1.3	10.5	2.6	10.1	2.3	5.0	0.8	0%	5%	9%	70%	5%	63%	37%	3%
	2	R$ 1,596	7.3	0.6	10.2	1.6	7.2	1.6	5.0	0.5	2%	36%	19%	17%	11%	43%	57%	24%
	3	R$ 3,869	7.5	0.5	10.3	1.3	9.9	0.9	5.0	0.5	0%	2%	6%	3%	82%	73%	27%	7%
	4	R$ 1,343	8.2	2.6	8.6	3.4	4.3	3.3	4.0	0.9	38%	10%	18%	11%	9%	26%	74%	8%
	5	R$ 3,869	7.9	0.5	10.2	1.5	8.6	1.0	5.0	0.5	0%	4%	74%	5%	7%	72%	28%	11%
	6	R$ 3,869	12.2	3.3	2.5	3.3	5.0	3.3	3.0	1.0	3%	15%	27%	16%	22%	72%	28%	17%
	7	R$ 1,596	7.6	3.1	9.8	3.9	9.0	3.7	4.0	1.0	4%	20%	27%	15%	17%	47%	53%	26%
	8	R$12,453	8.4	1.8	9.0	2.7	5.6	2.0	4.0	0.8	2%	21%	26%	17%	14%	98%	2%	5%

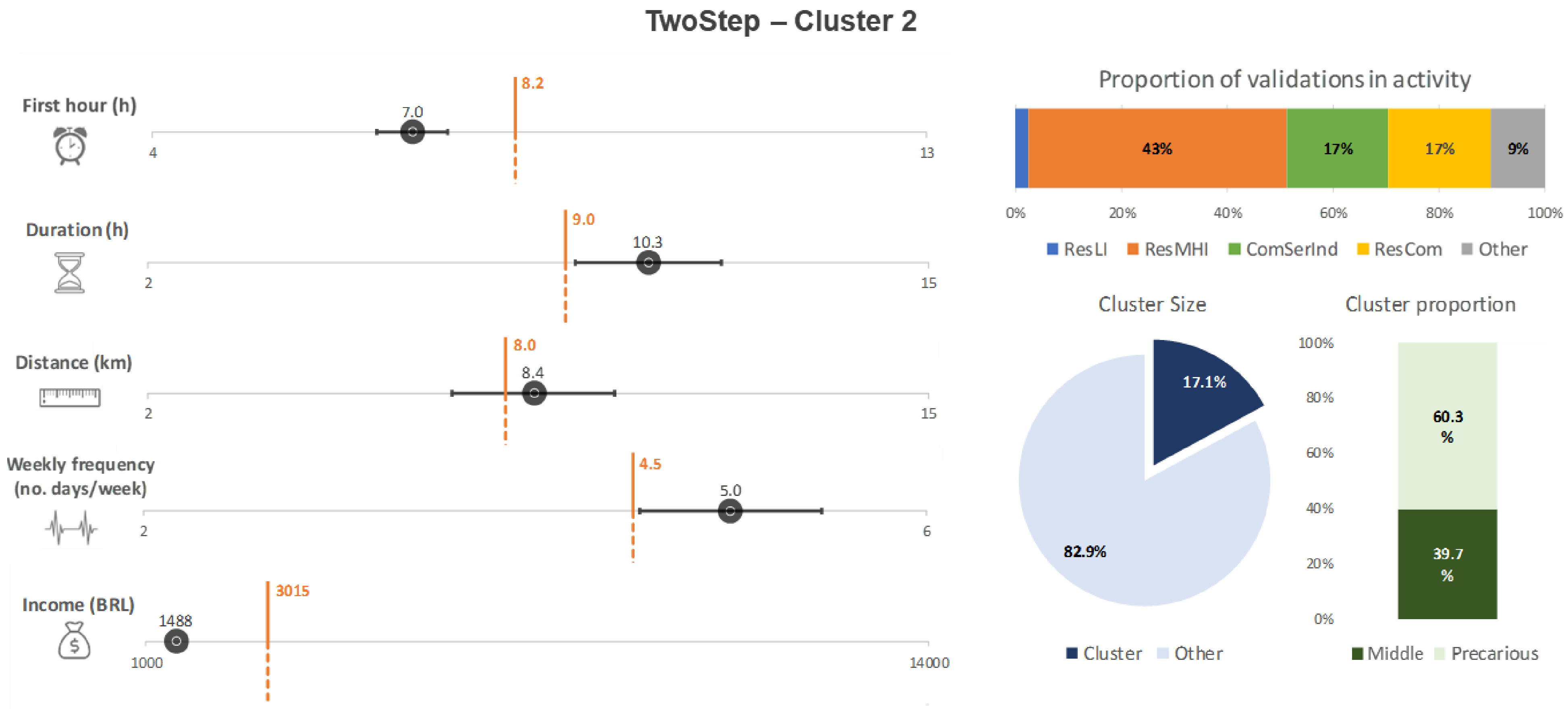
Source: The authors' own elaboration

Figure 43 – Groups formed from the TwoStep clustering algorithm – Cluster 1



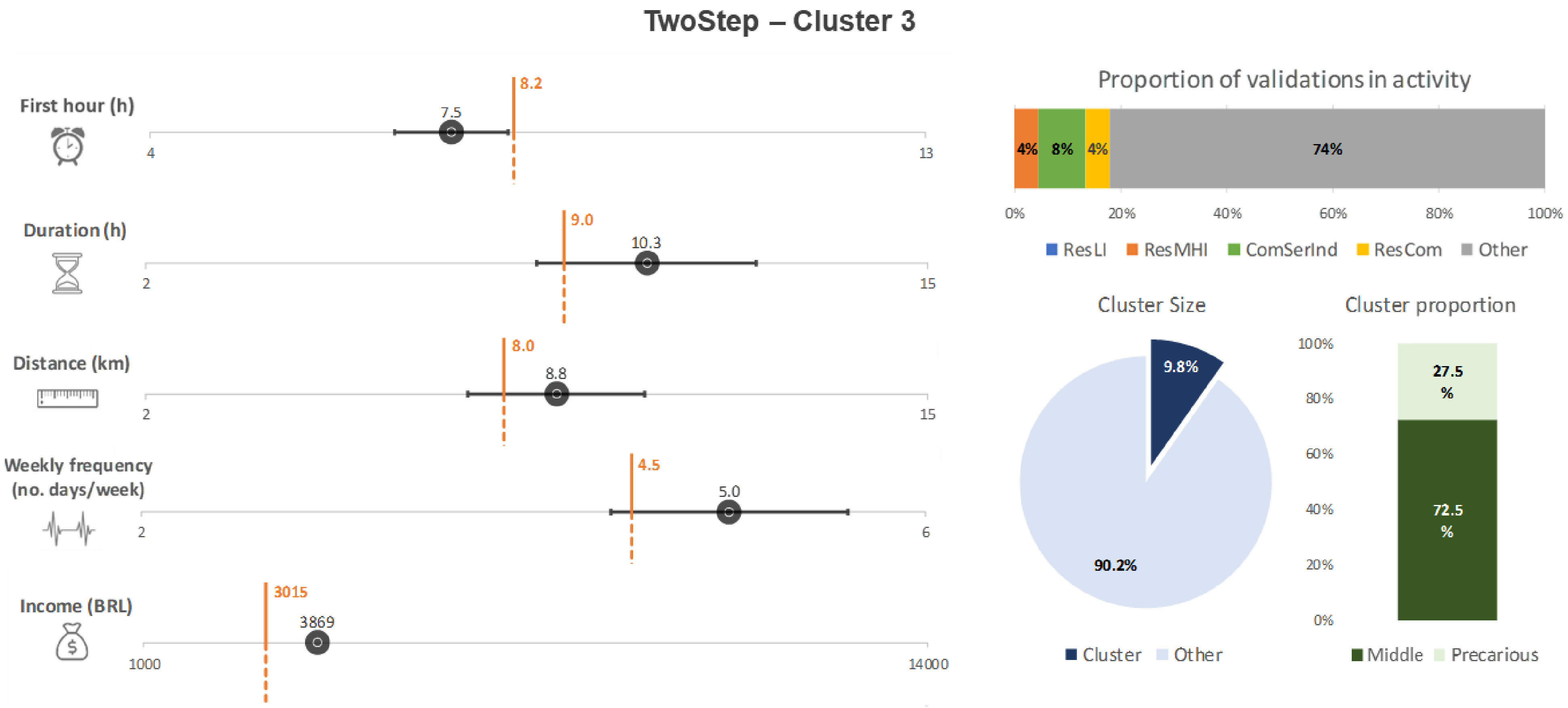
Source: The authors' own elaboration

Figure 44 – Groups formed from the TwoStep clustering algorithm – Cluster 2



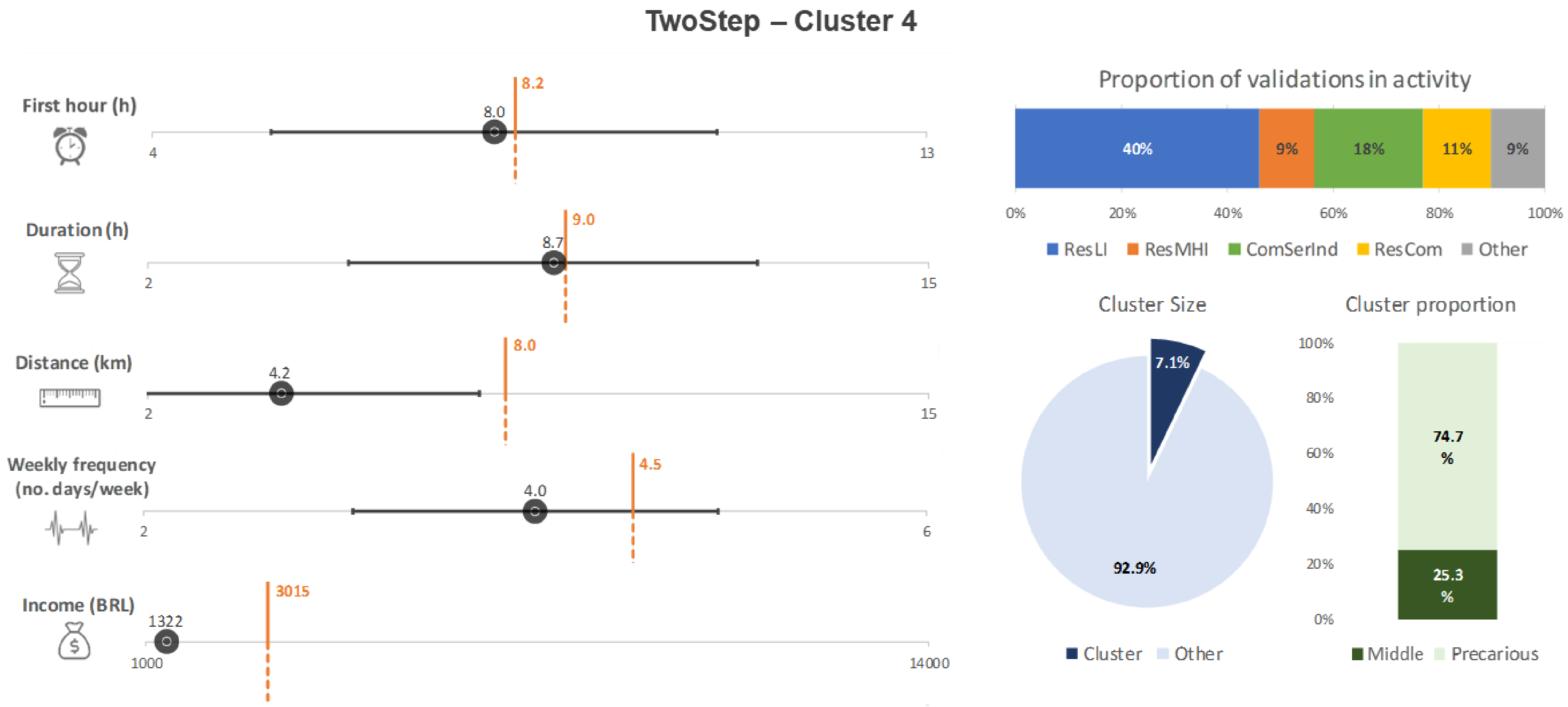
Source: The authors' own elaboration

Figure 45 – Groups formed from the TwoStep clustering algorithm – Cluster 3



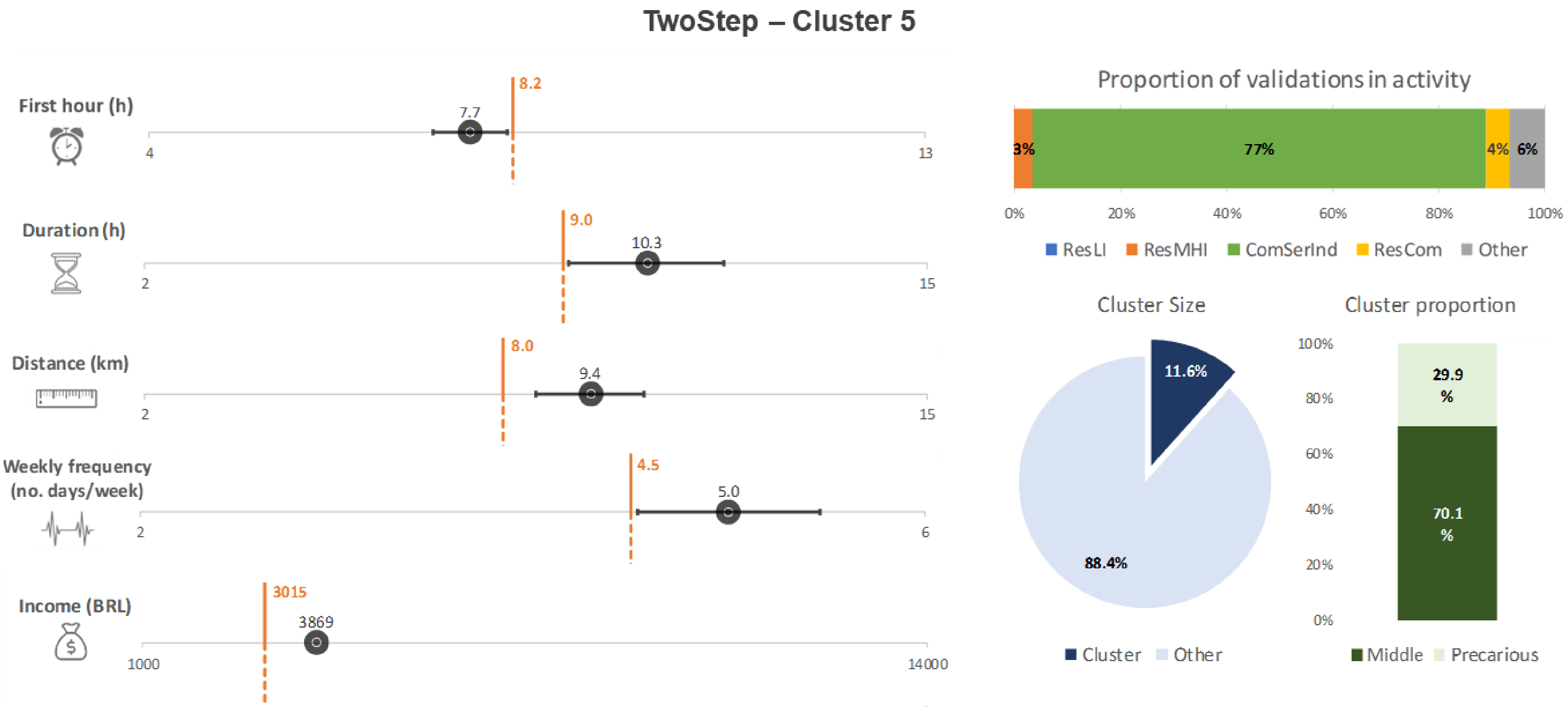
Source: The authors' own elaboration

Figure 46 – Groups formed from the TwoStep clustering algorithm – Cluster 4



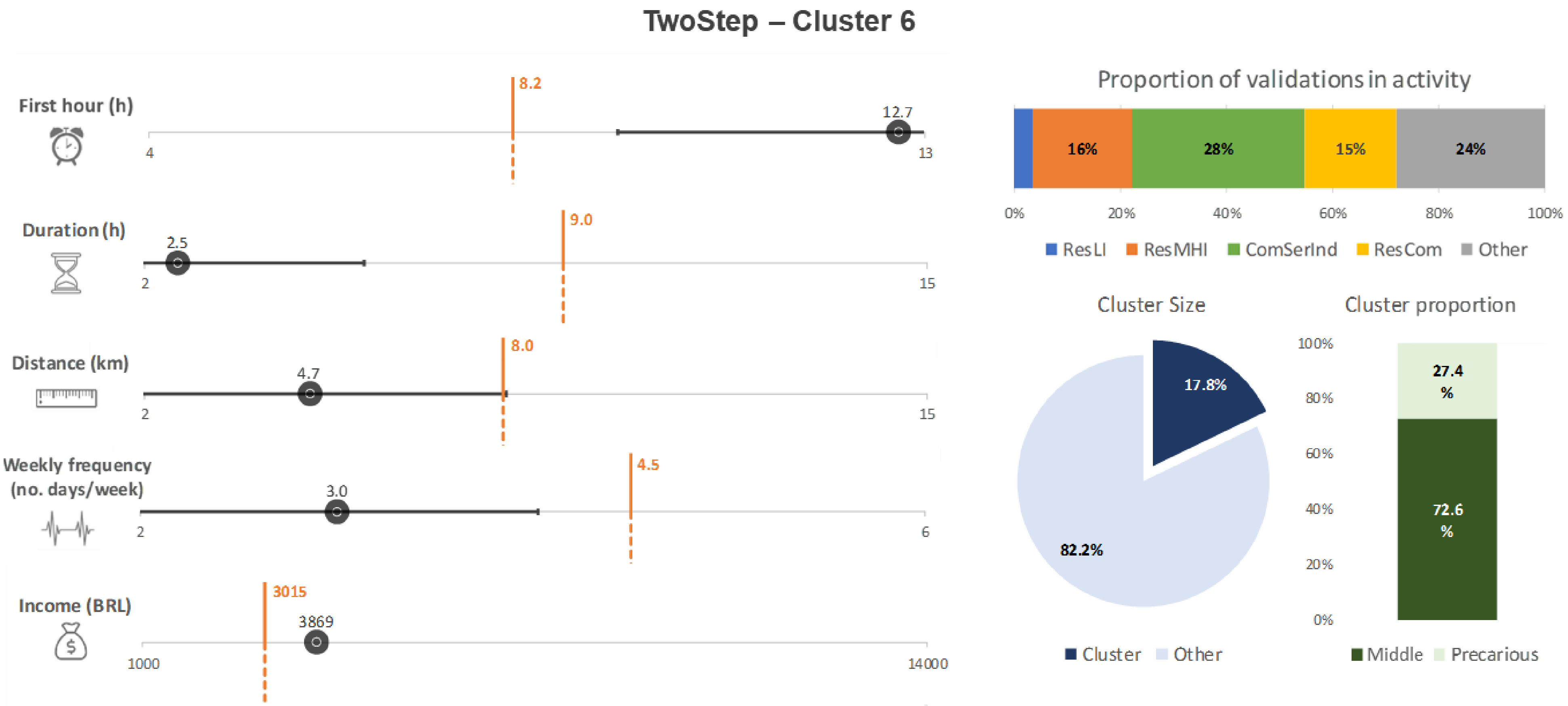
Source: The authors' own elaboration

Figure 47 – Groups formed from the TwoStep clustering algorithm – Cluster 5



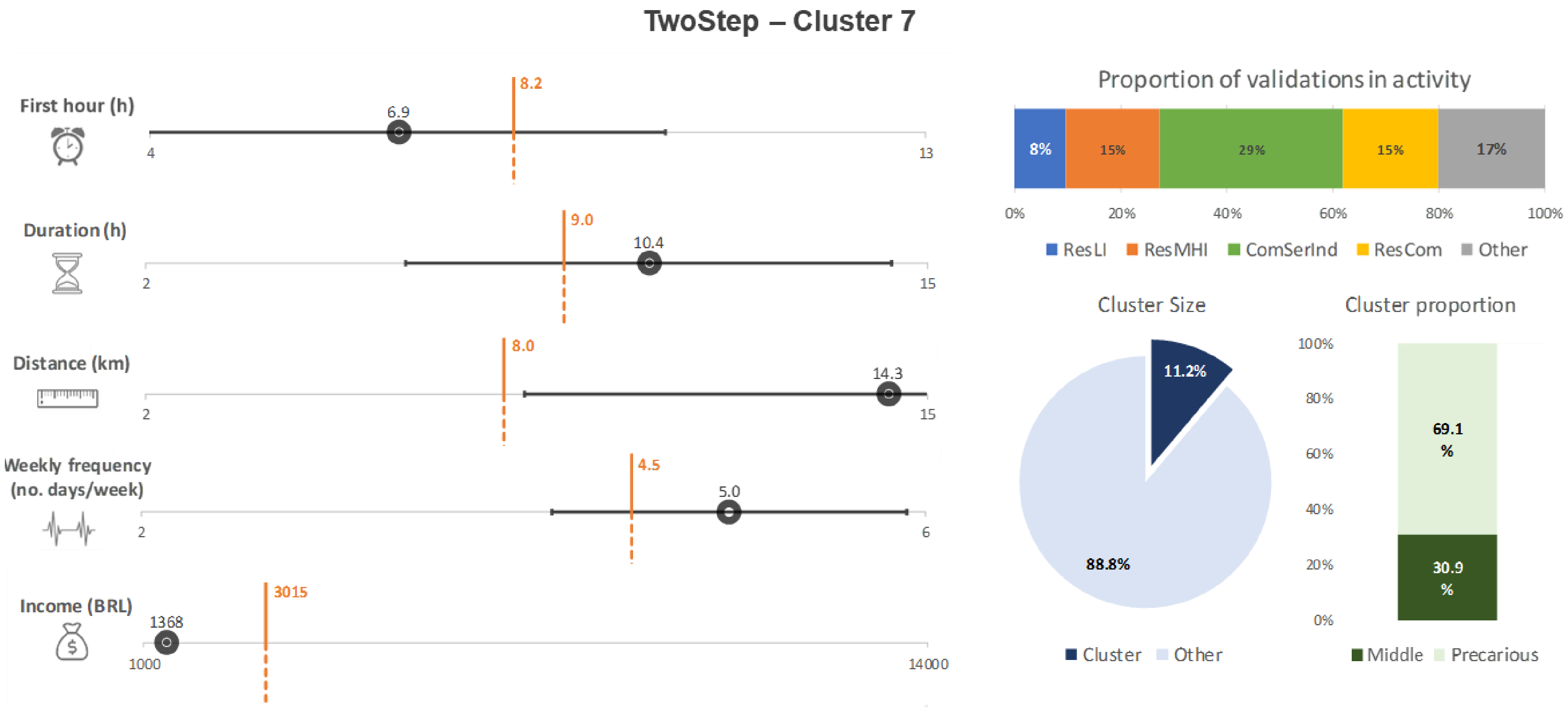
Source: The authors' own elaboration

Figure 48 – Groups formed from the TwoStep clustering algorithm – Cluster 6



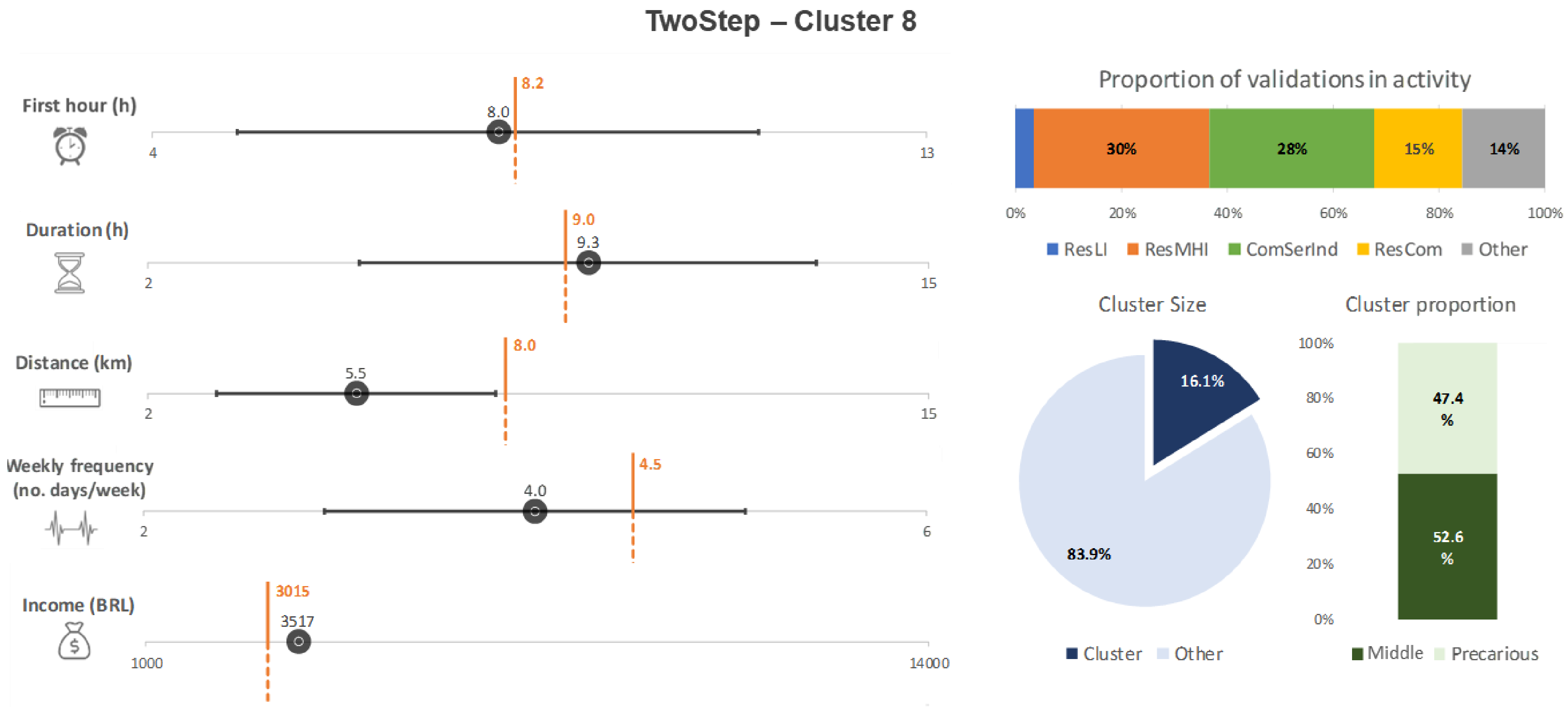
Source: The authors' own elaboration

Figure 49 – Groups formed from the TwoStep clustering algorithm – Cluster 7



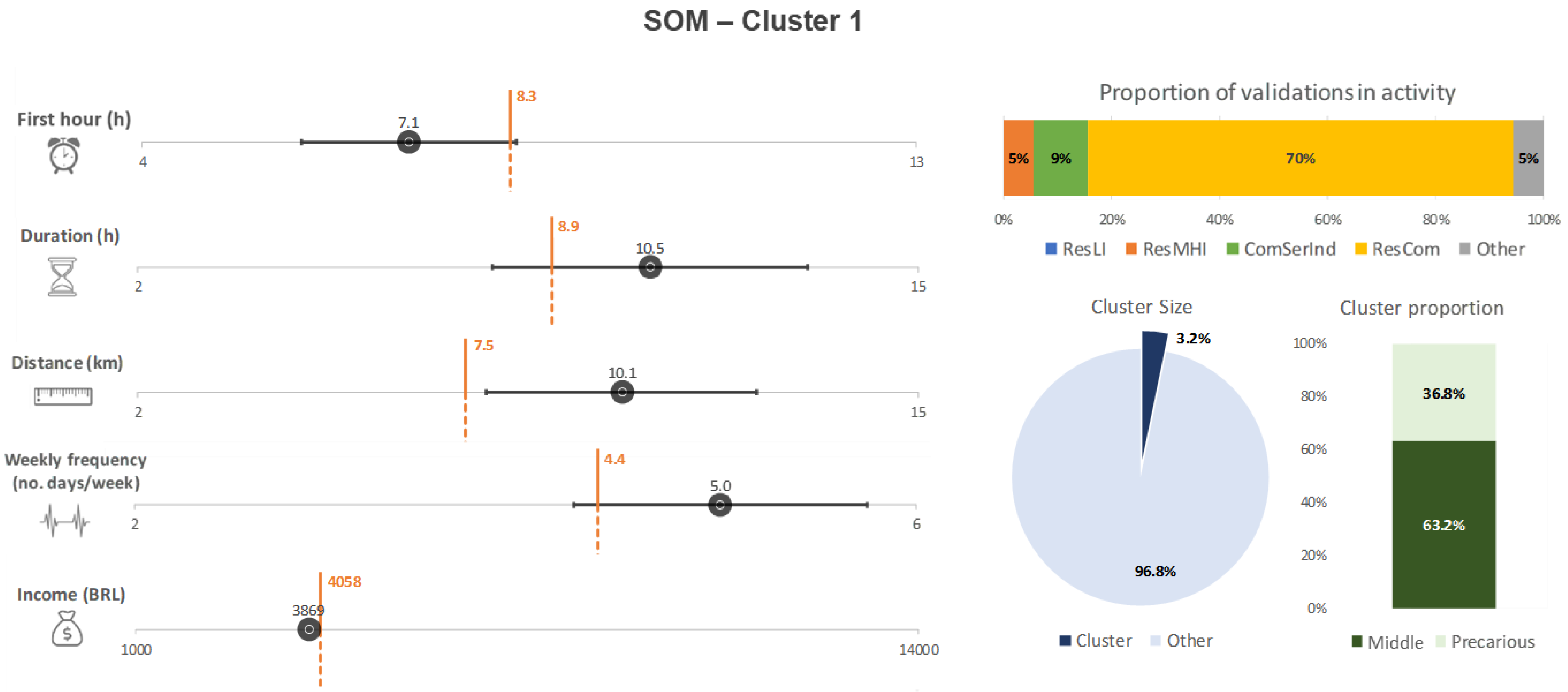
Source: The authors’ own elaboration

Figure 50 – Groups formed from the TwoStep clustering algorithm – Cluster 8



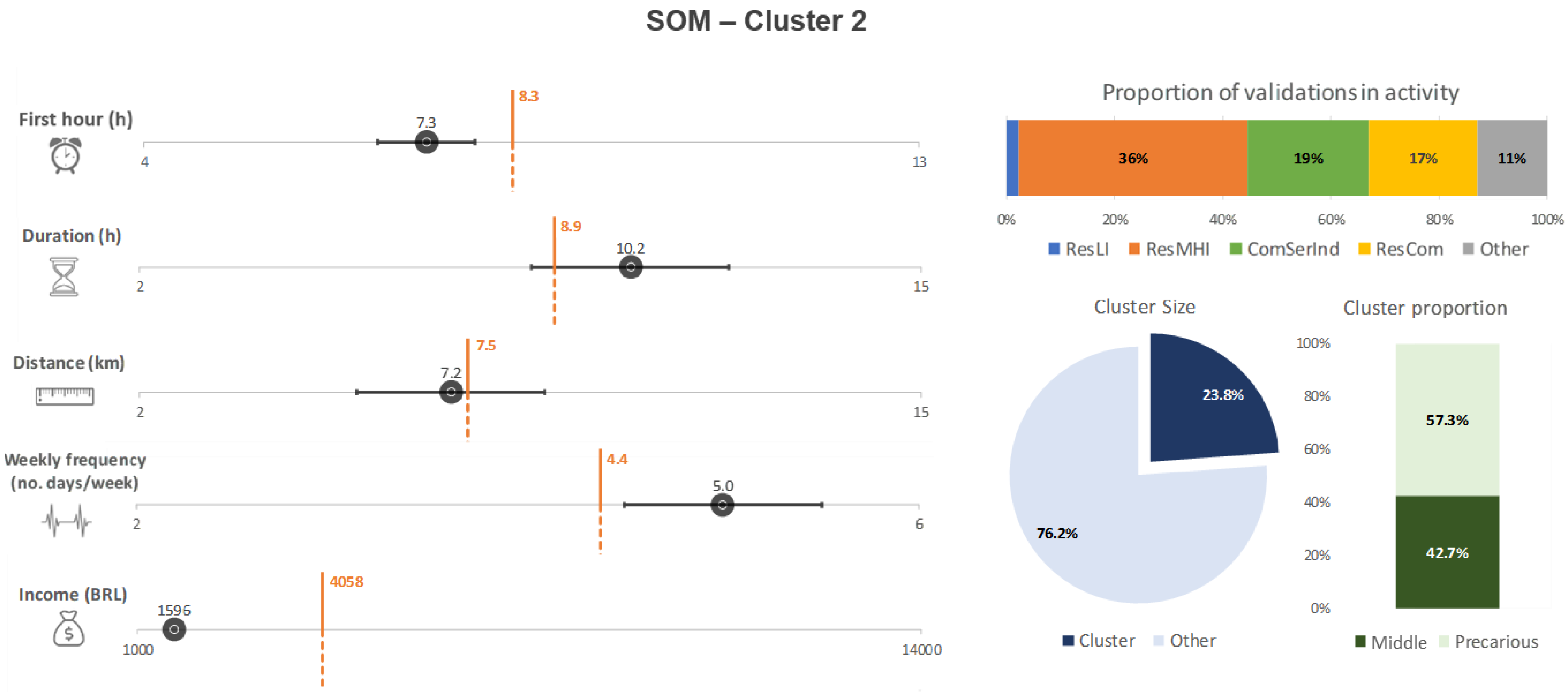
Source: The authors' own elaboration

Figure 51 – Groups formed from the SOM clustering algorithm – Cluster 1



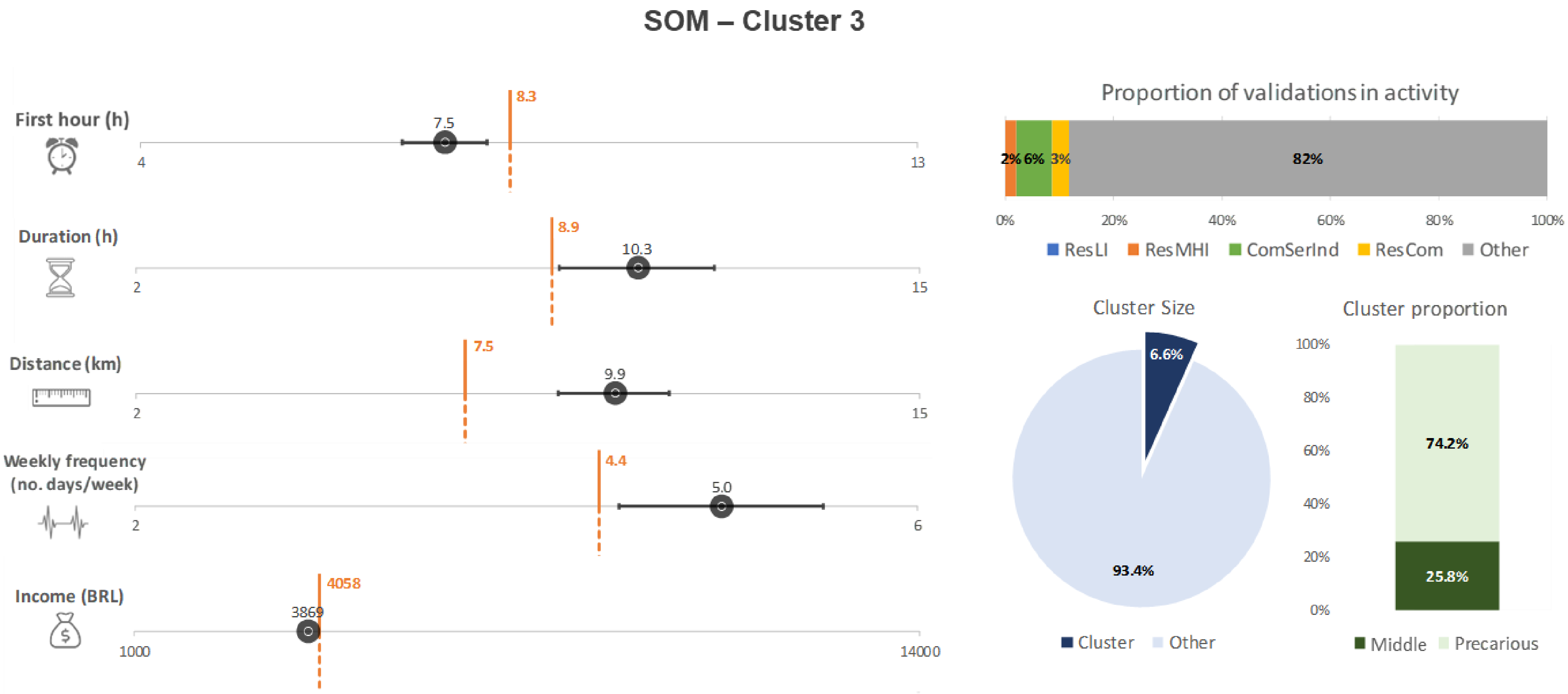
Source: The authors' own elaboration

Figure 52 – Groups formed from the SOM clustering algorithm – Cluster 2



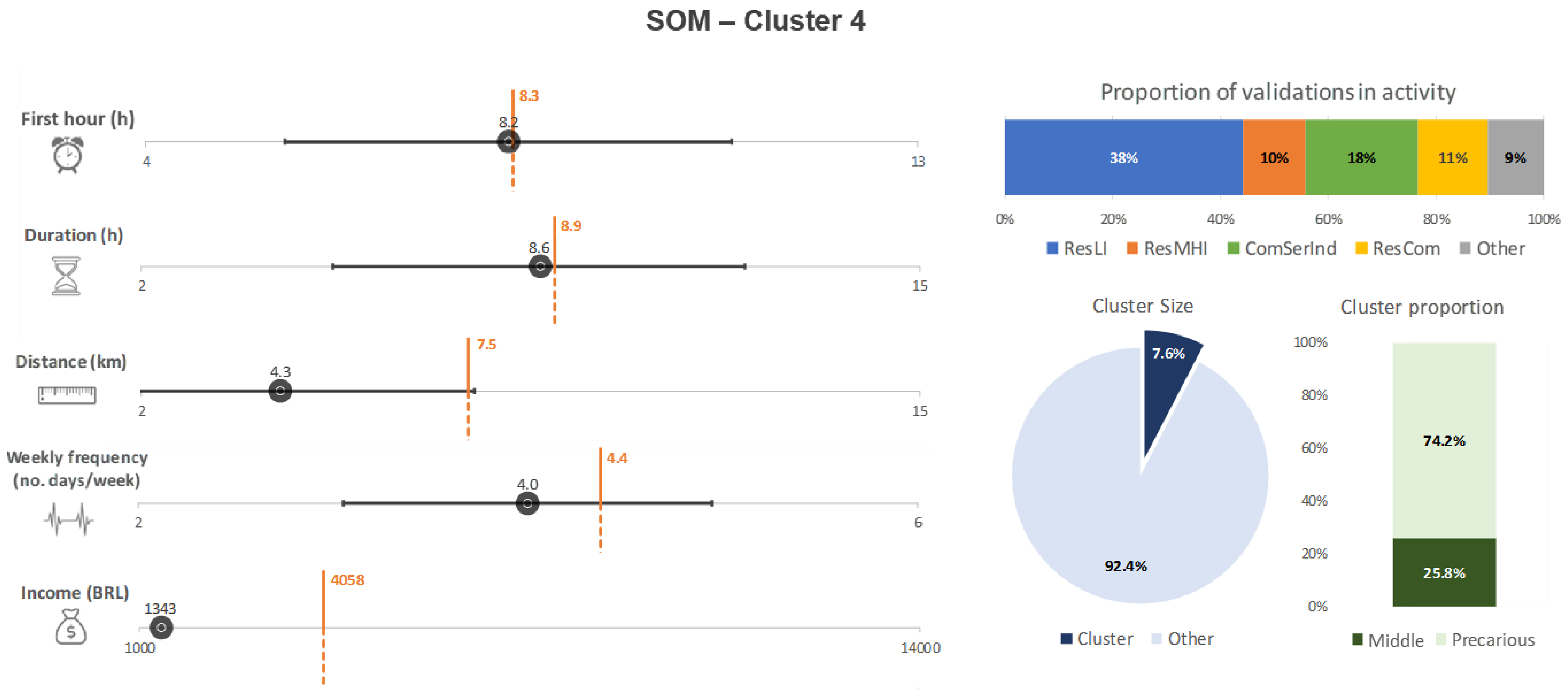
Source: The authors’ own elaboration

Figure 53 – Groups formed from the SOM clustering algorithm – Cluster 3



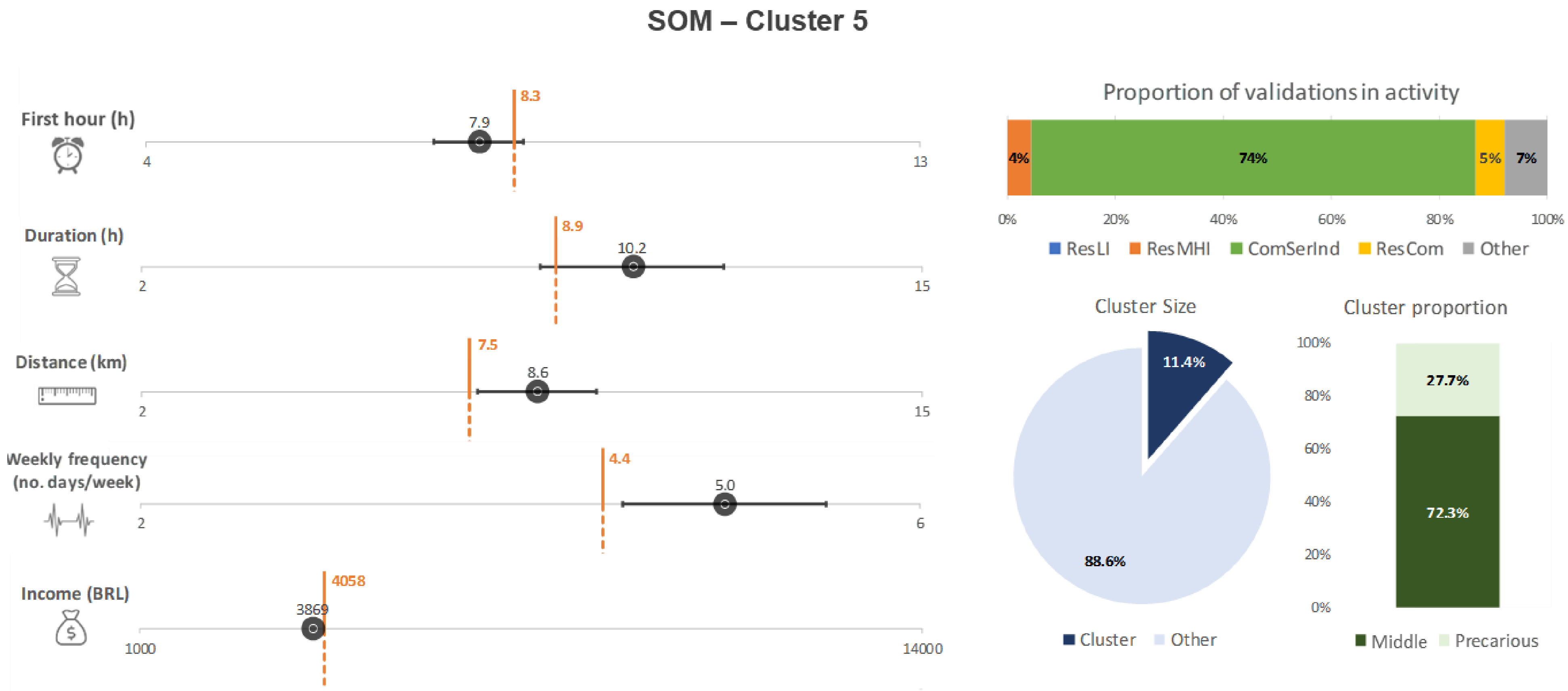
Source: The authors’ own elaboration

Figure 54 – Groups formed from the SOM clustering algorithm – Cluster 4



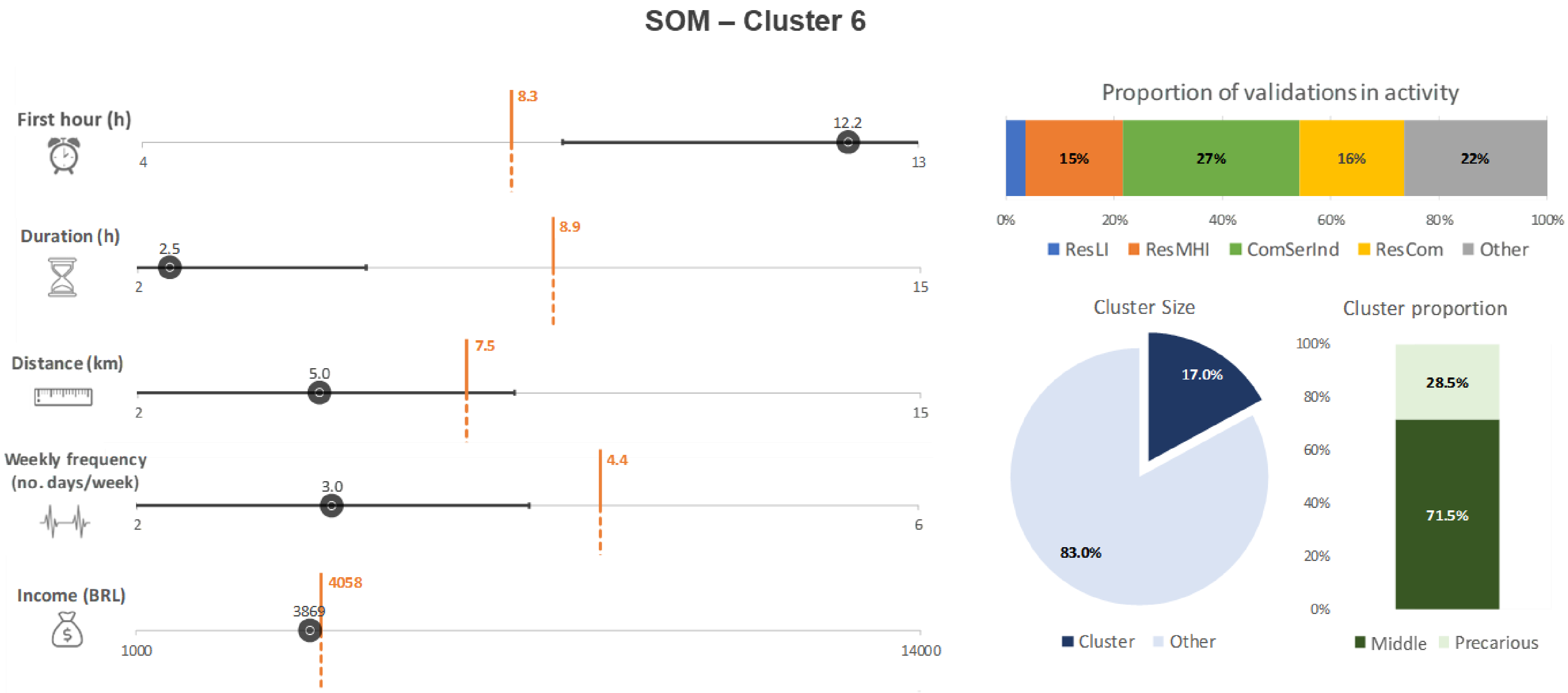
Source: The authors' own elaboration

Figure 55 – Groups formed from the SOM clustering algorithm – Cluster 5



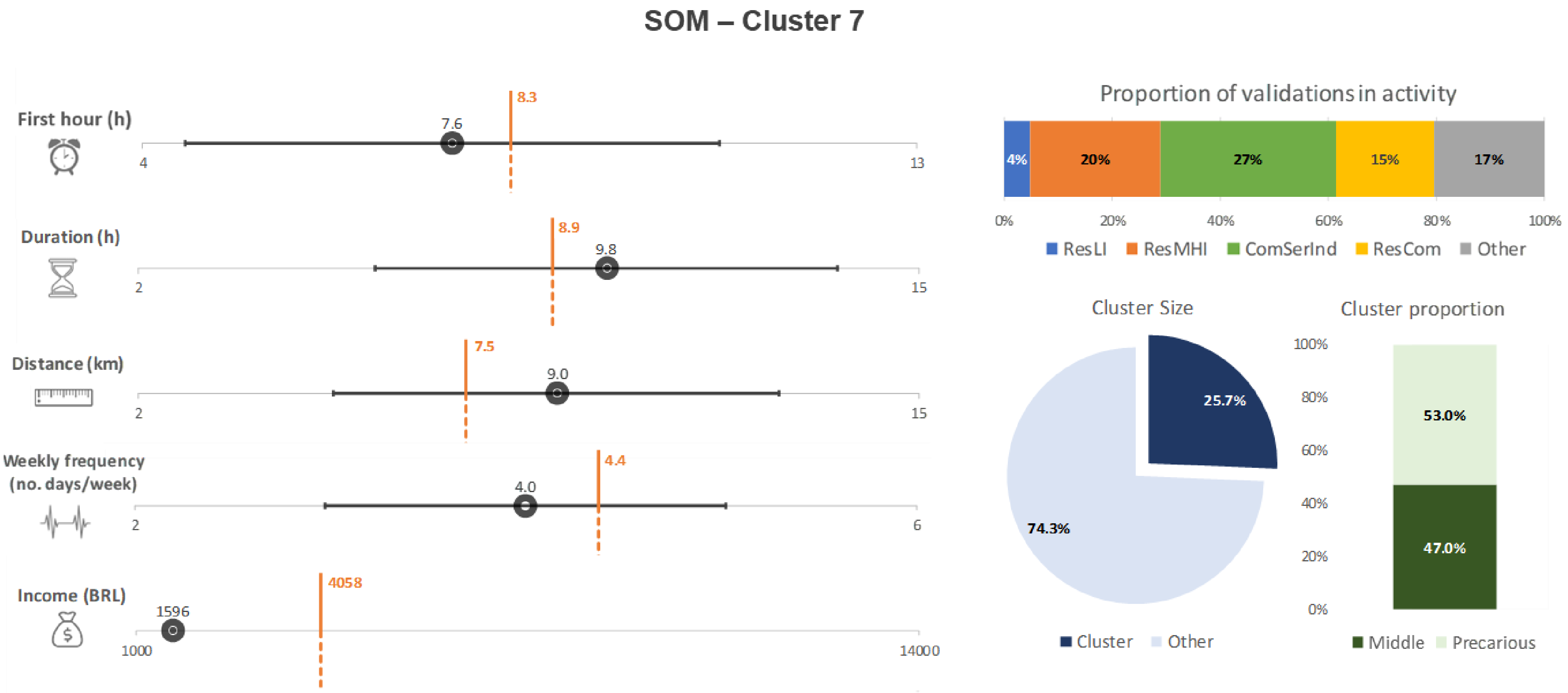
Source: The authors’ own elaboration

Figure 56 – Groups formed from the SOM clustering algorithm – Cluster 6



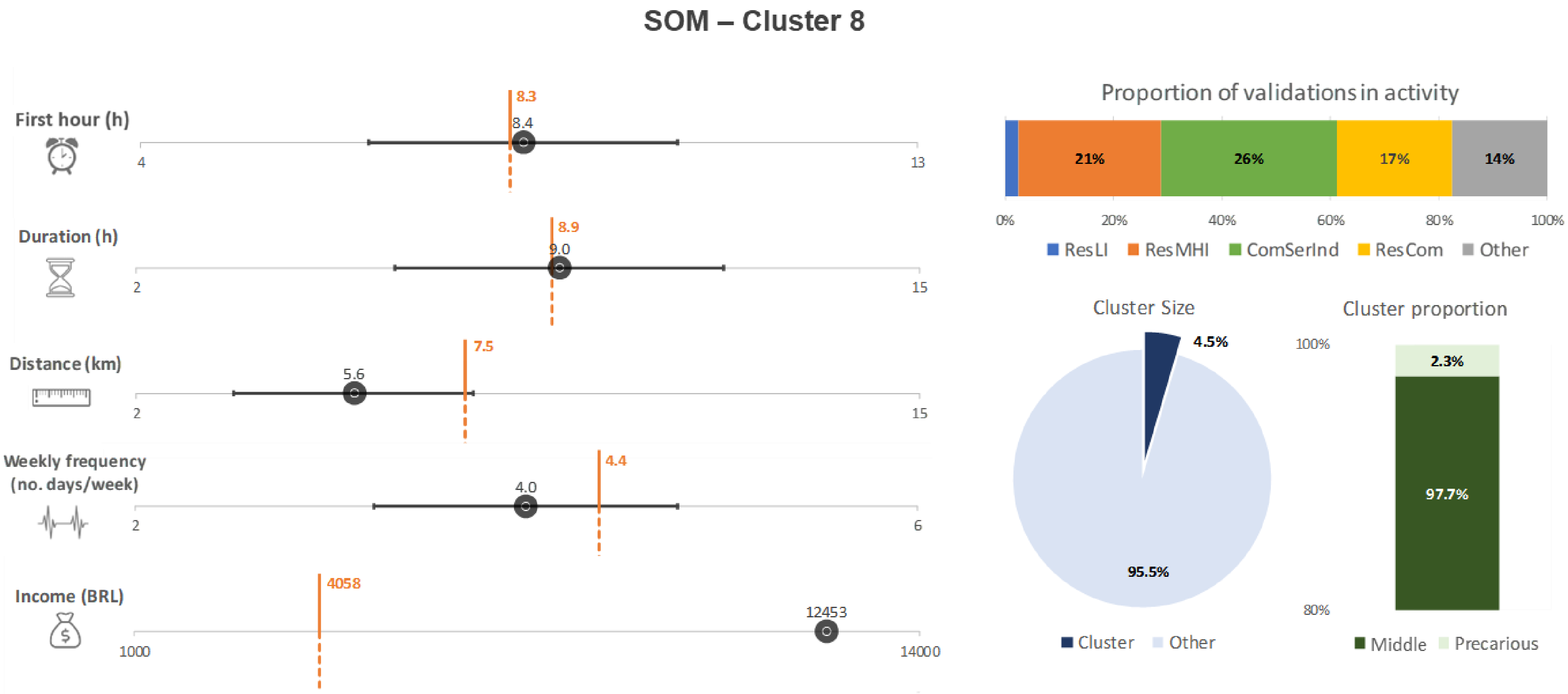
Source: The authors' own elaboration

Figure 57 – Groups formed from the SOM clustering algorithm – Cluster 7



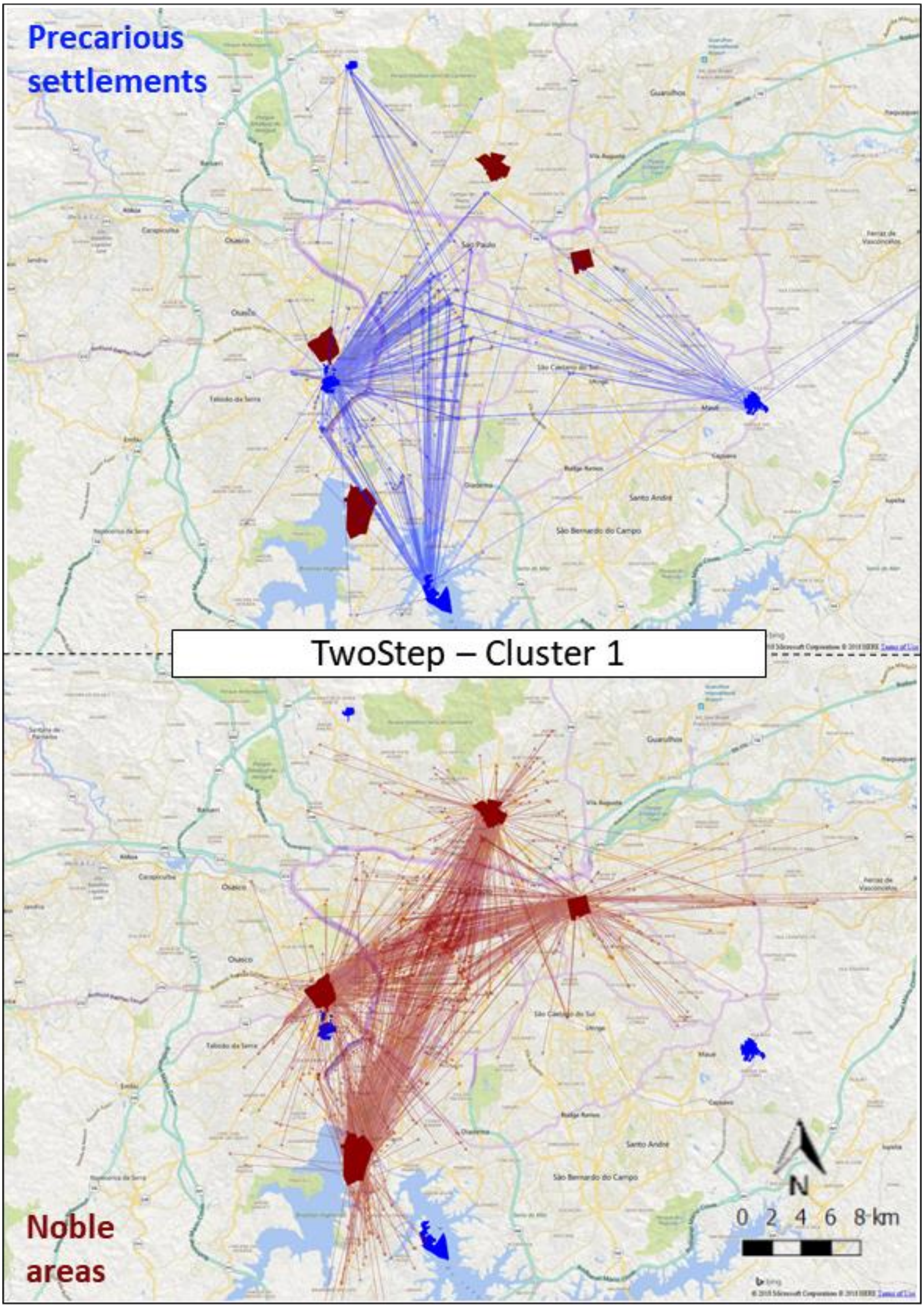
Source: The authors' own elaboration

Figure 58 – Groups formed from the SOM clustering algorithm – Cluster 8



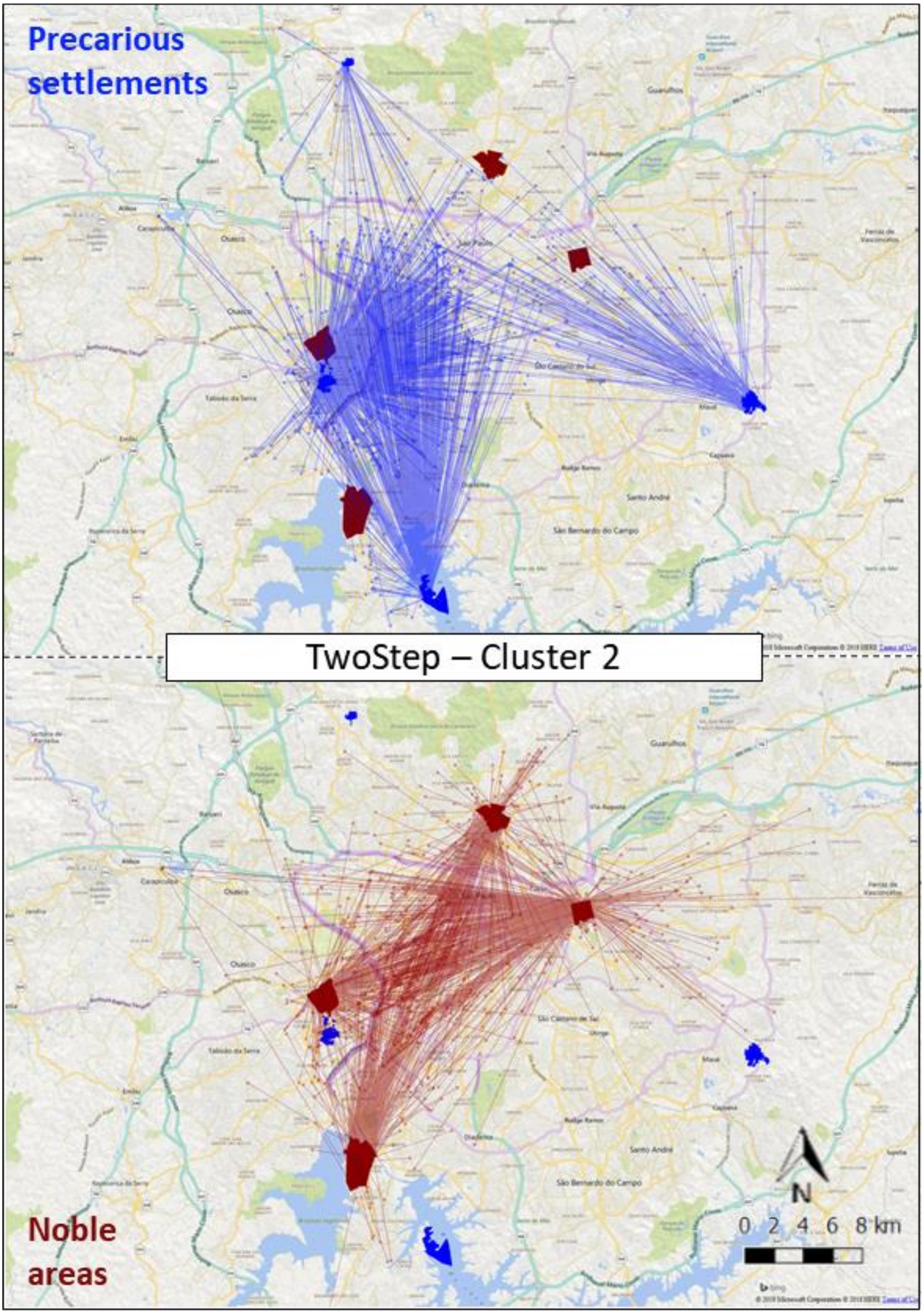
Source: The authors’ own elaboration

Figure 59 – Activity distribution by cluster – TwoStep / Cluster 1



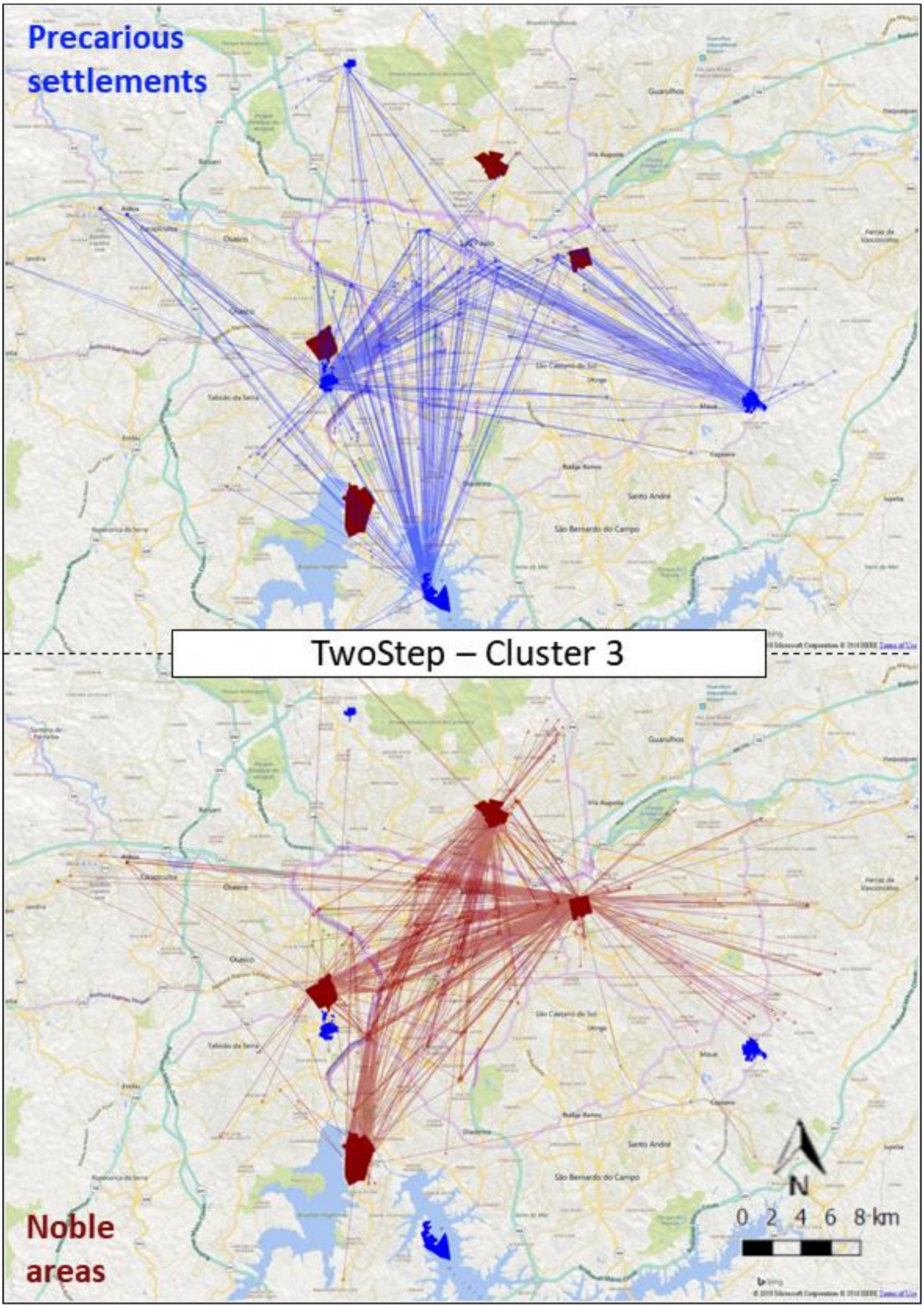
Source: The authors' own elaboration

Figure 60 – Activity distribution by cluster – TwoStep / Cluster 2



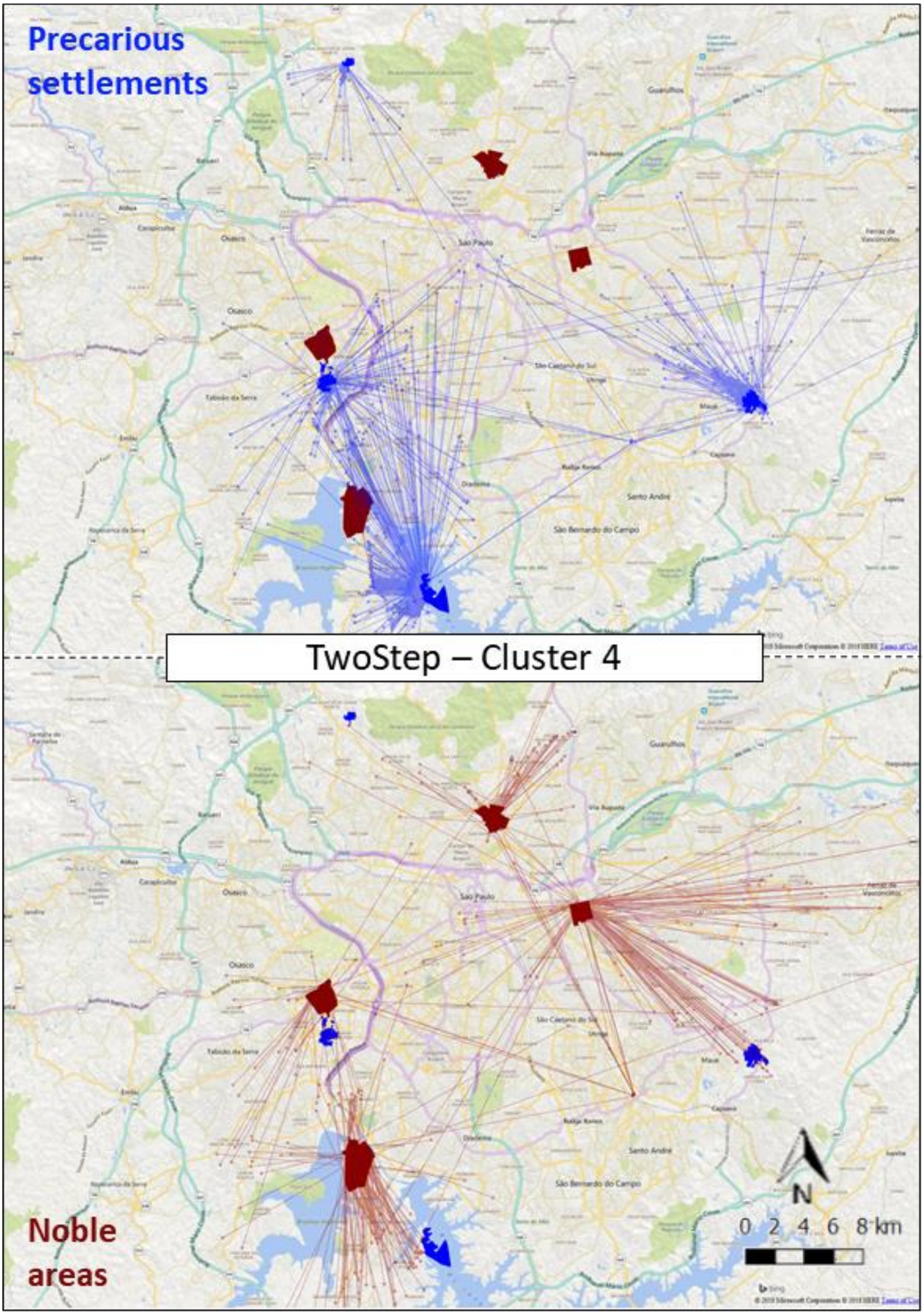
Source: The authors' own elaboration

Figure 61 – Activity distribution by cluster – TwoStep / Cluster 3



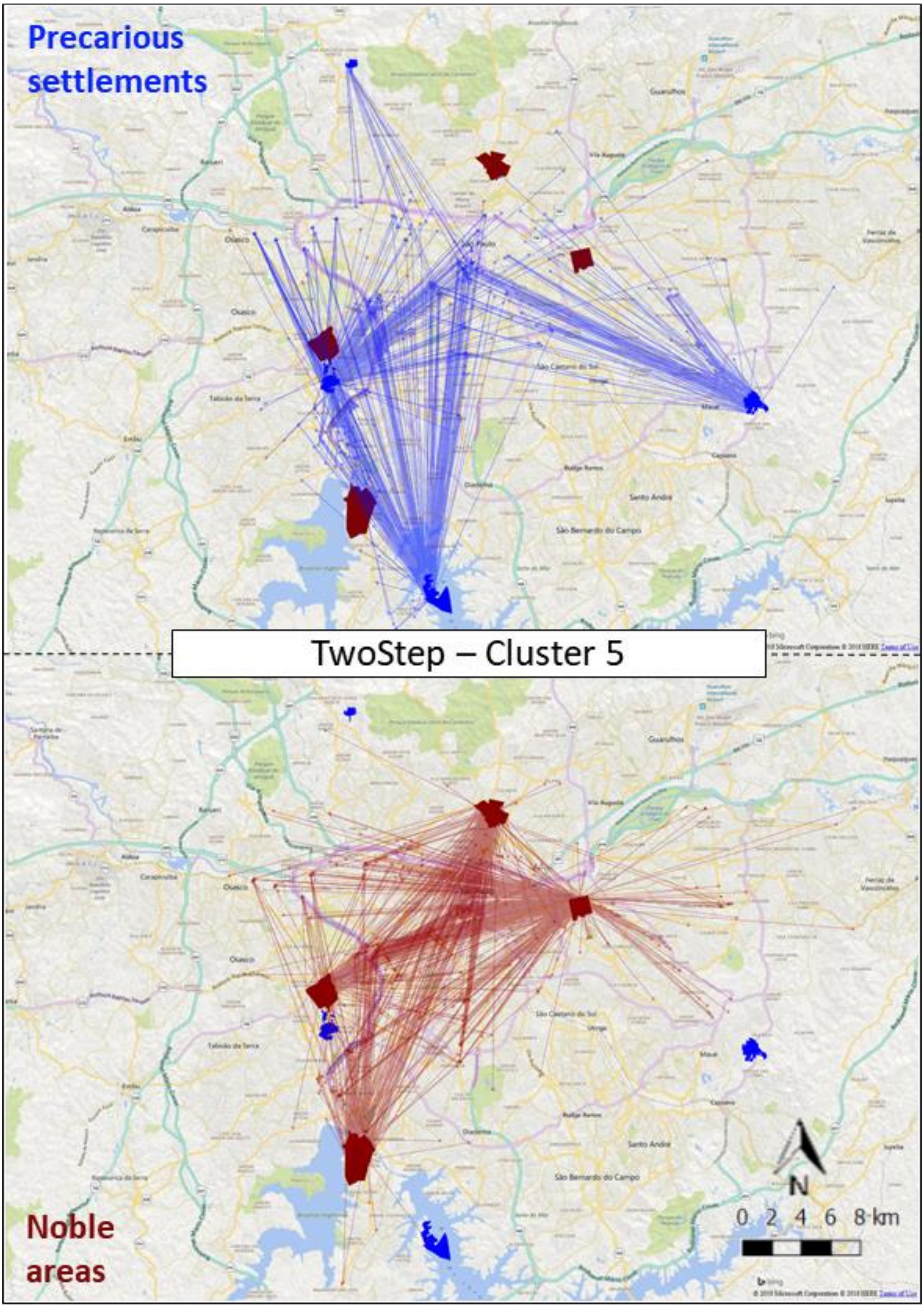
Source: The authors' own elaboration

Figure 62 – Activity distribution by cluster – TwoStep / Cluster 4



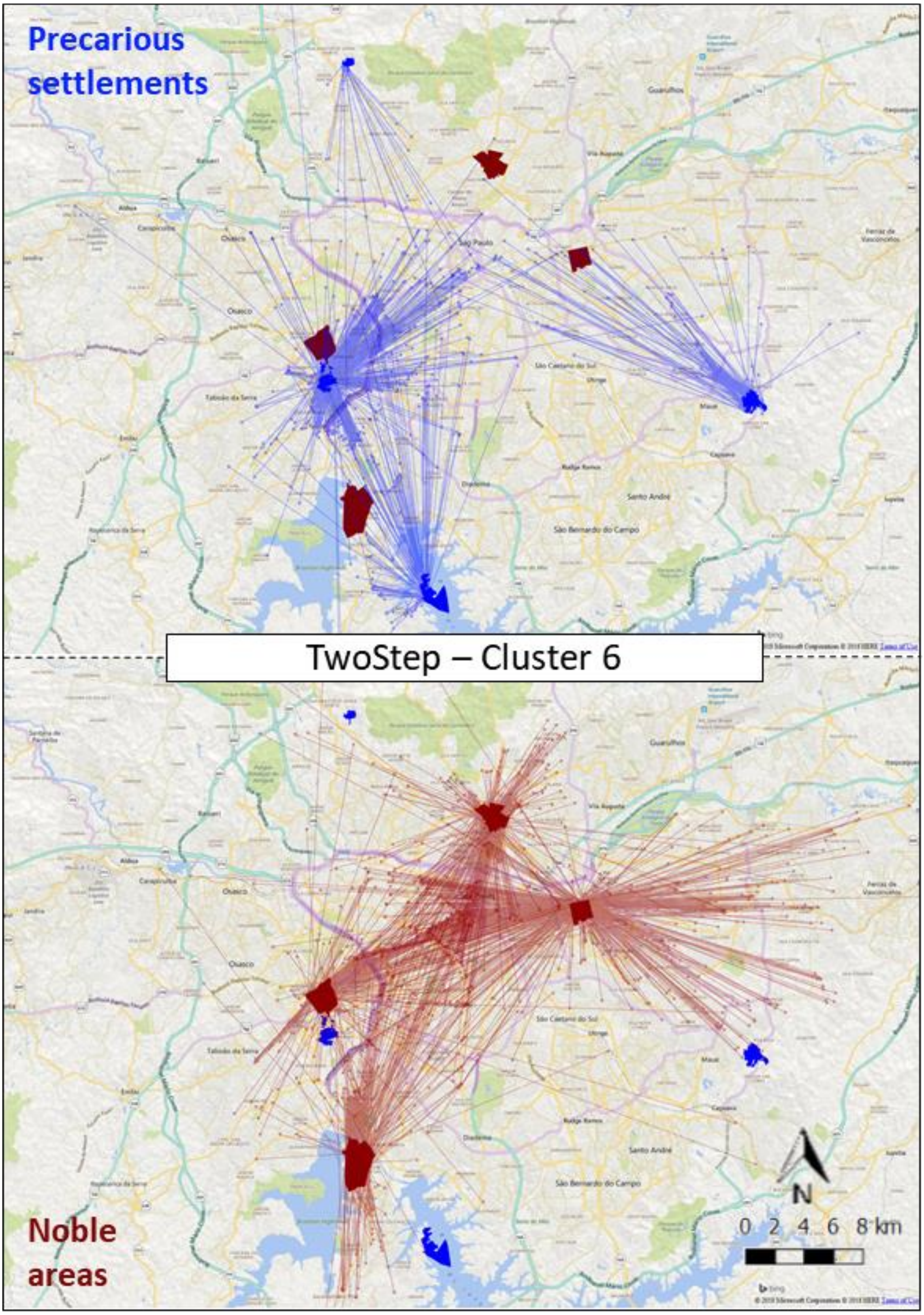
Source: The authors' own elaboration

Figure 63 – Activity distribution by cluster – TwoStep / Cluster 5



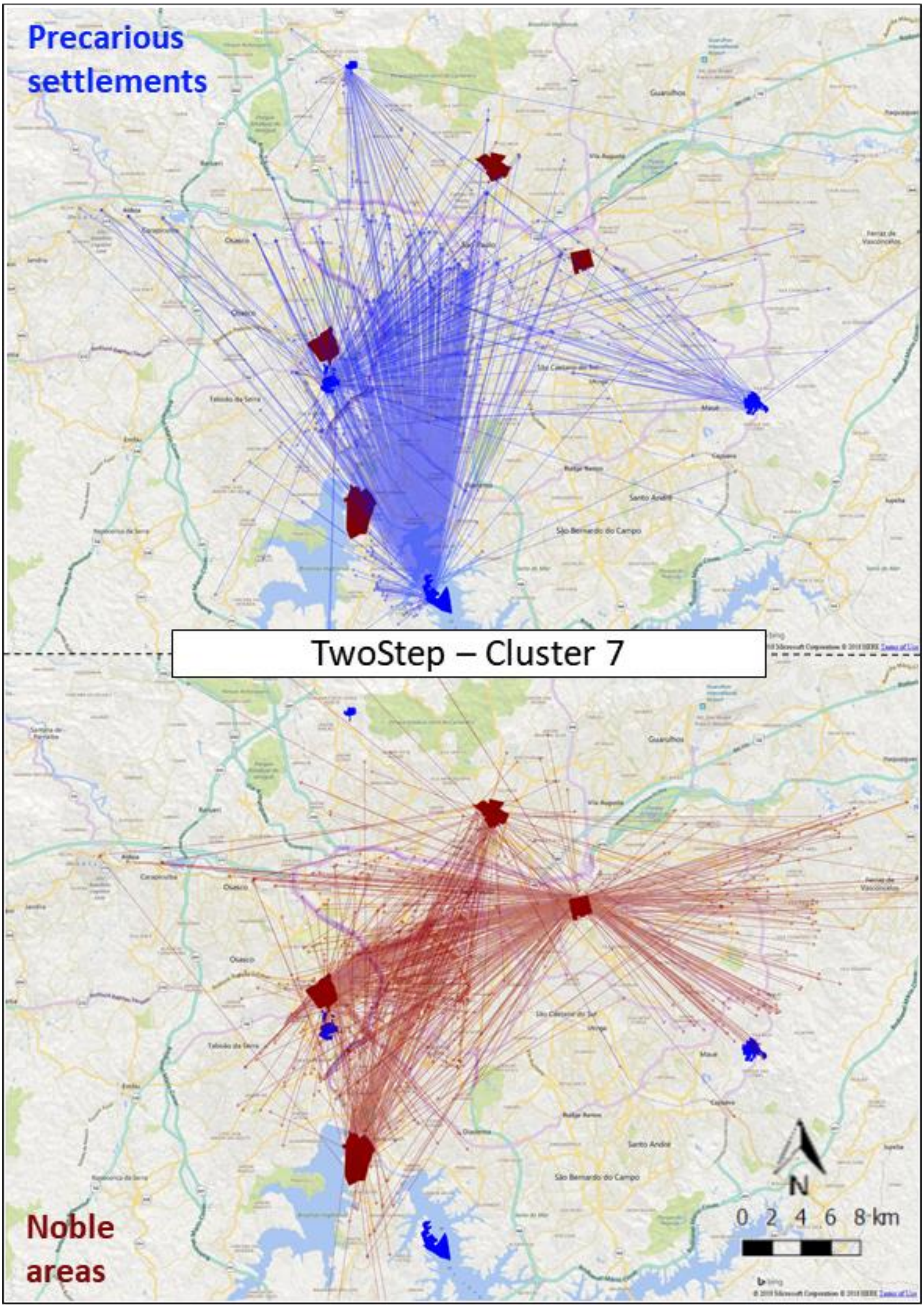
Source: The authors’ own elaboration

Figure 64 – Activity distribution by cluster – TwoStep / Cluster 6



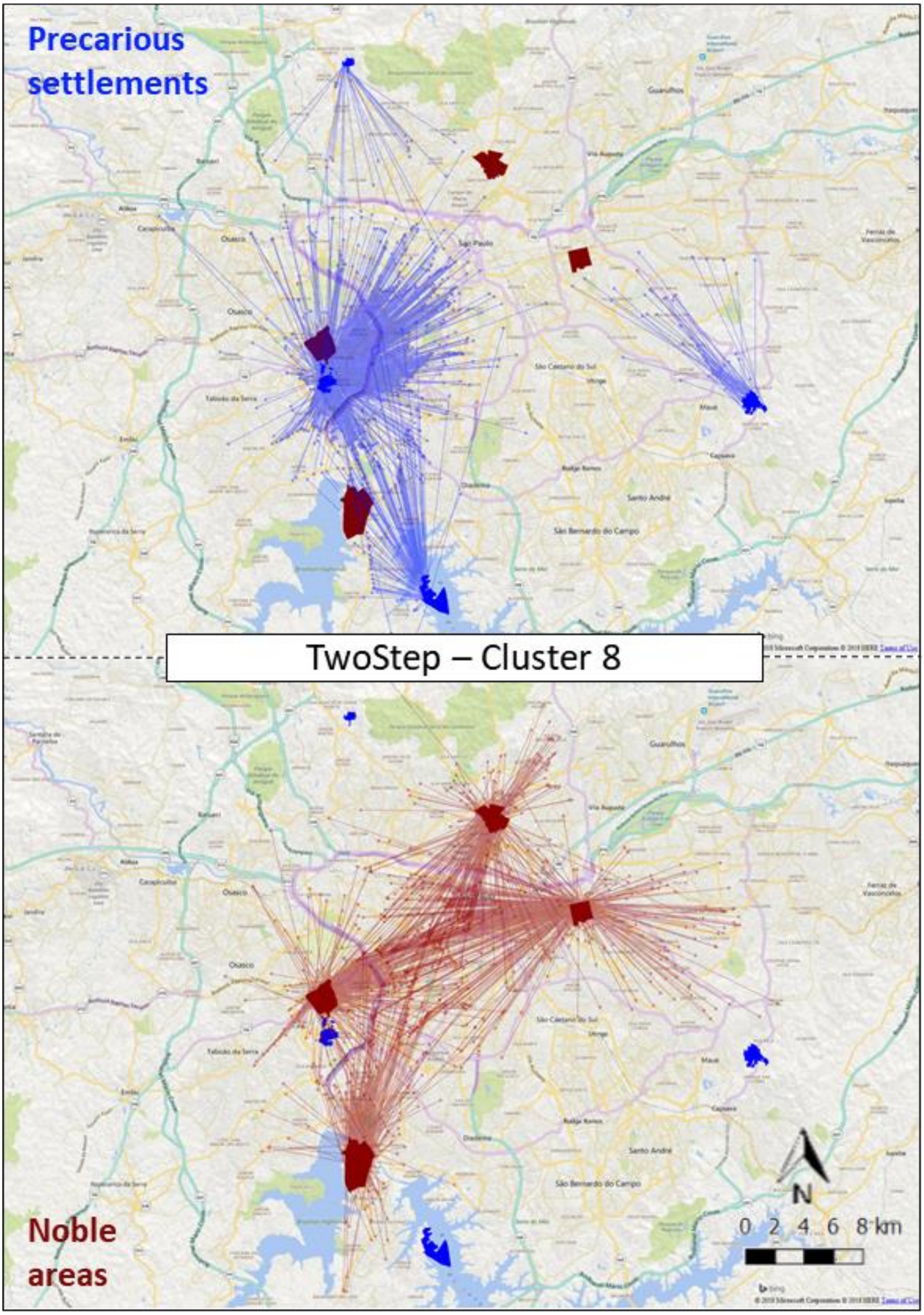
Source: The authors' own elaboration

Figure 65 – Activity distribution by cluster – TwoStep / Cluster 7



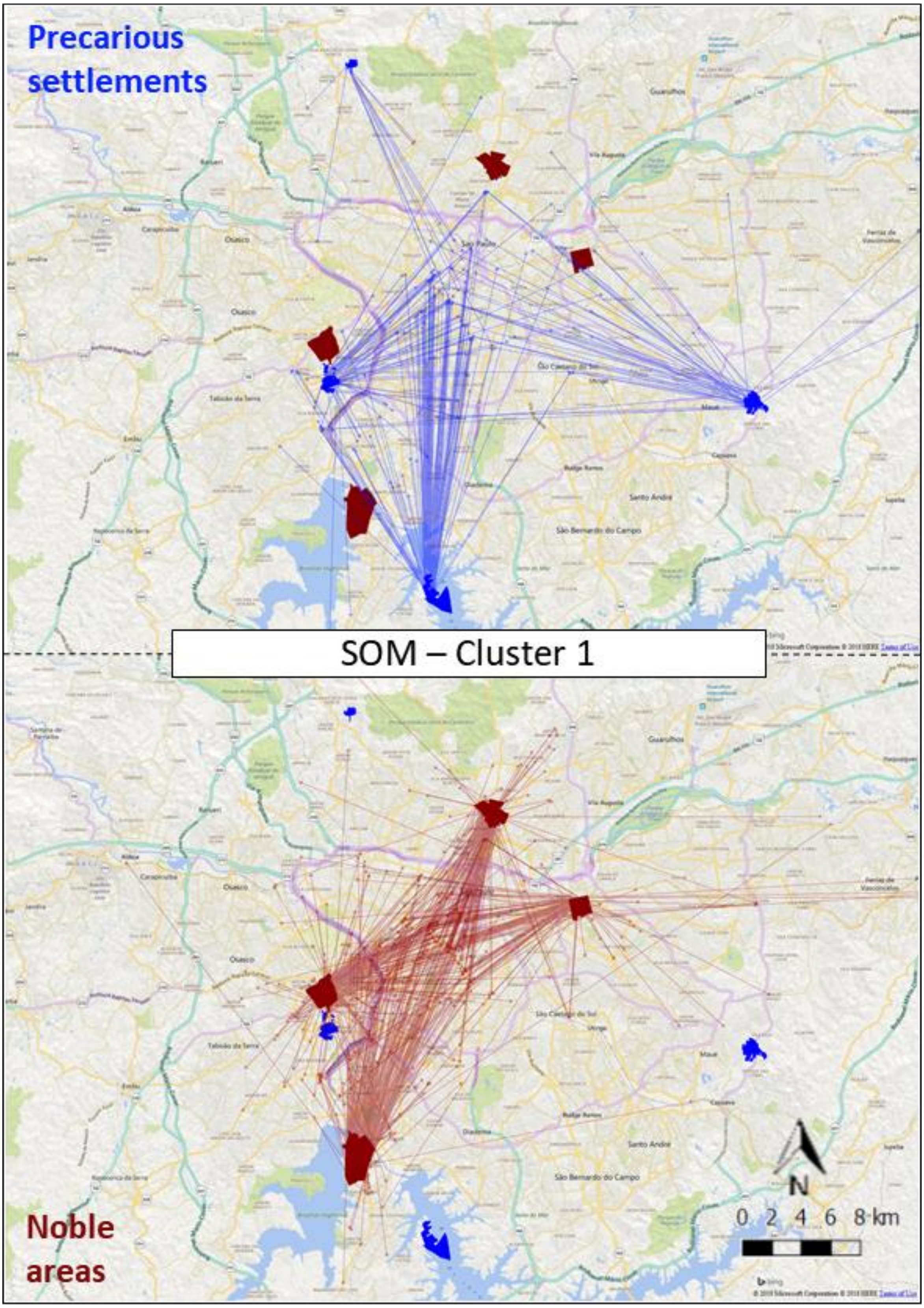
Source: The authors’ own elaboration

Figure 66 – Activity distribution by cluster – TwoStep / Cluster 8



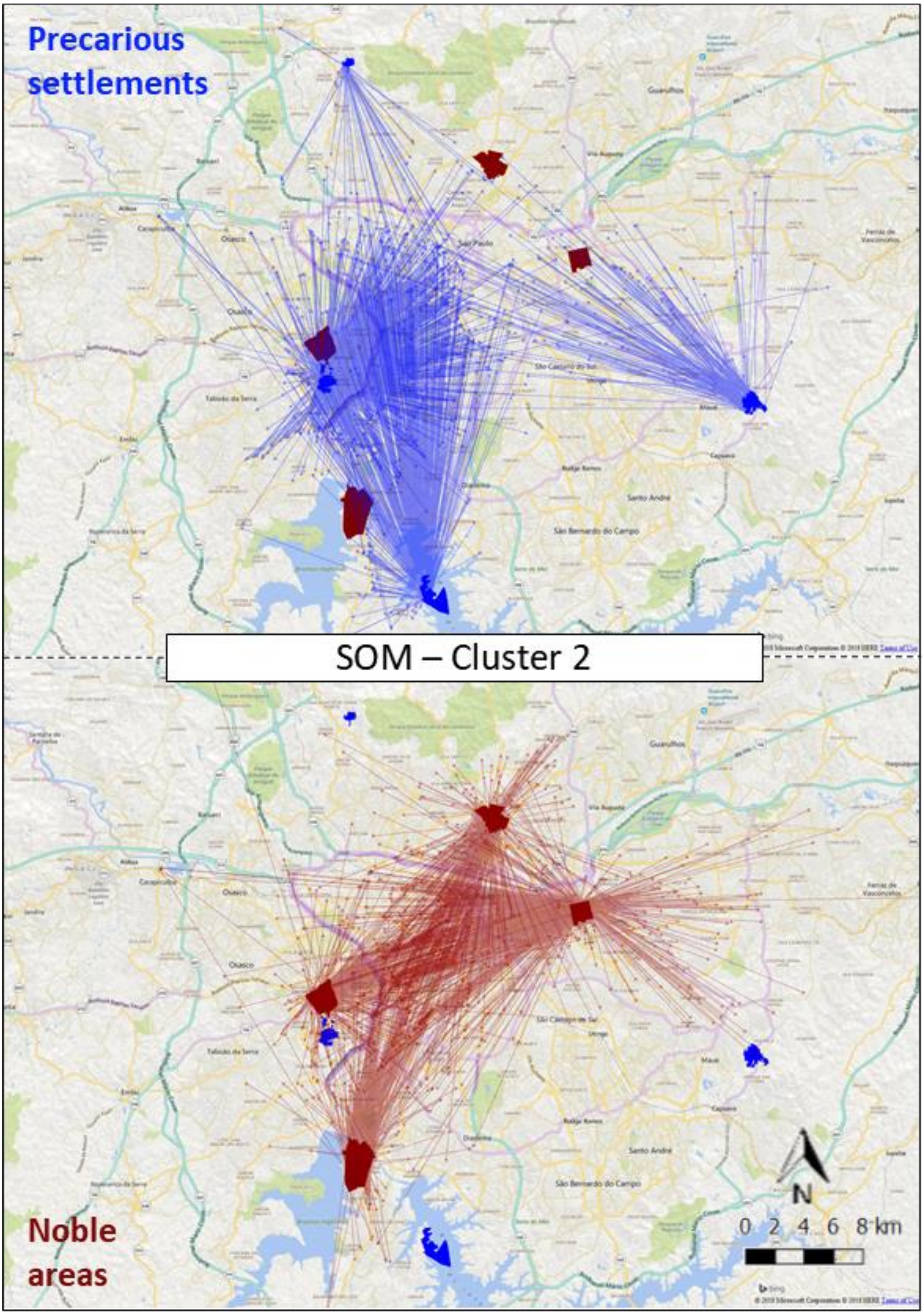
Source: The authors’ own elaboration

Figure 67 – Activity distribution by cluster – SOM / Cluster 1



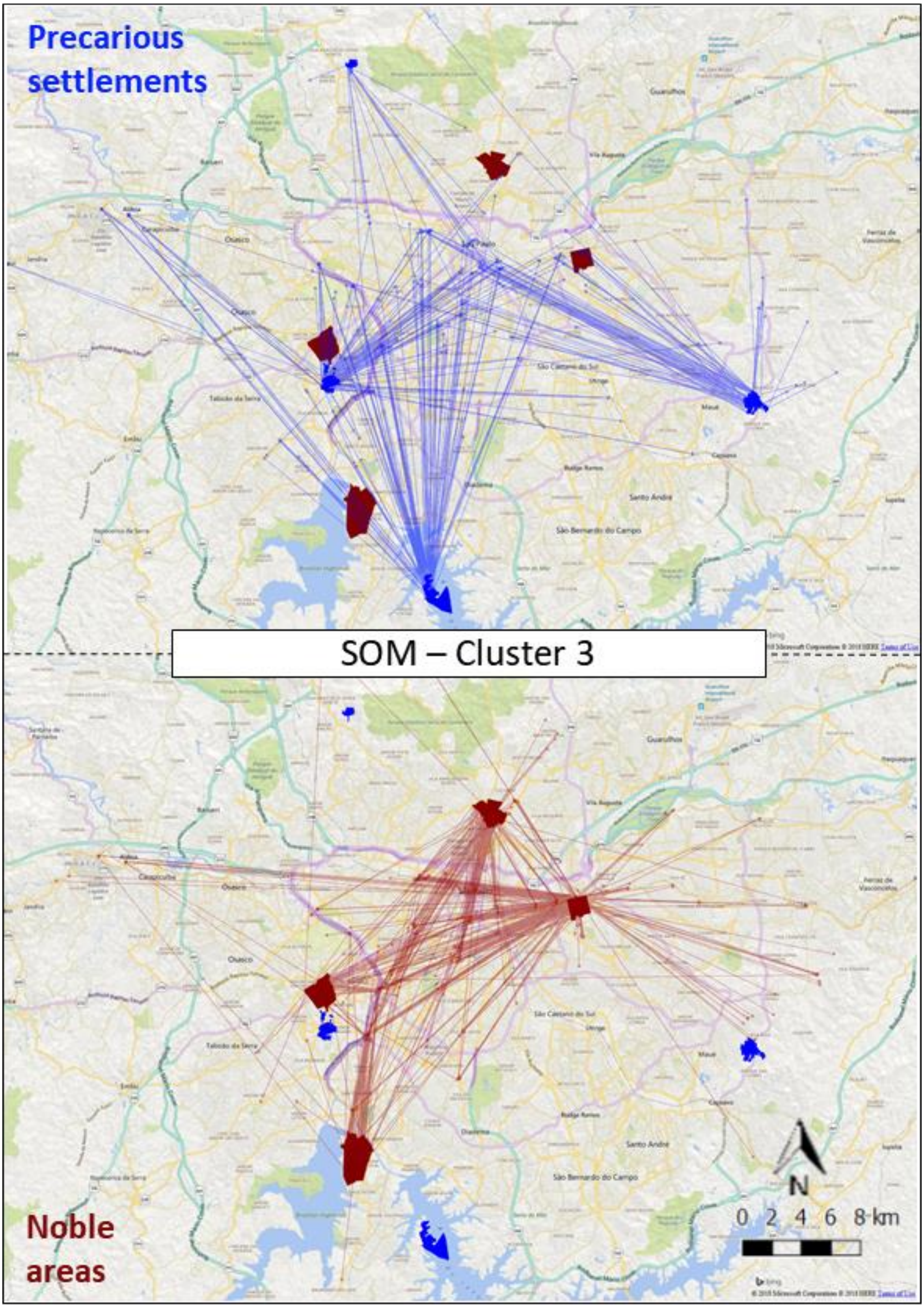
Source: The authors' own elaboration

Figure 68 – Activity distribution by cluster – SOM / Cluster 2



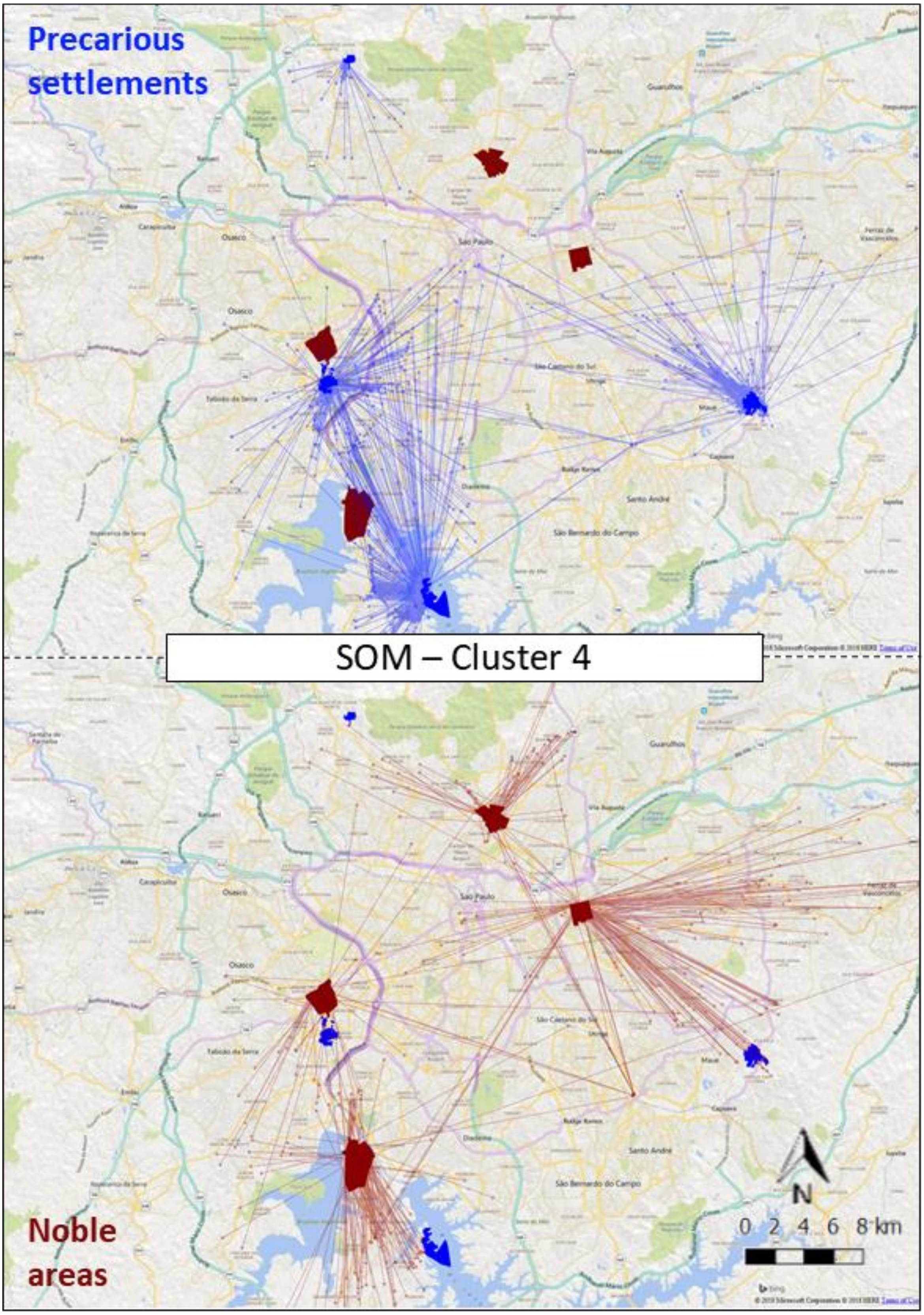
Source: The authors' own elaboration

Figure 69 – Activity distribution by cluster – SOM / Cluster 3



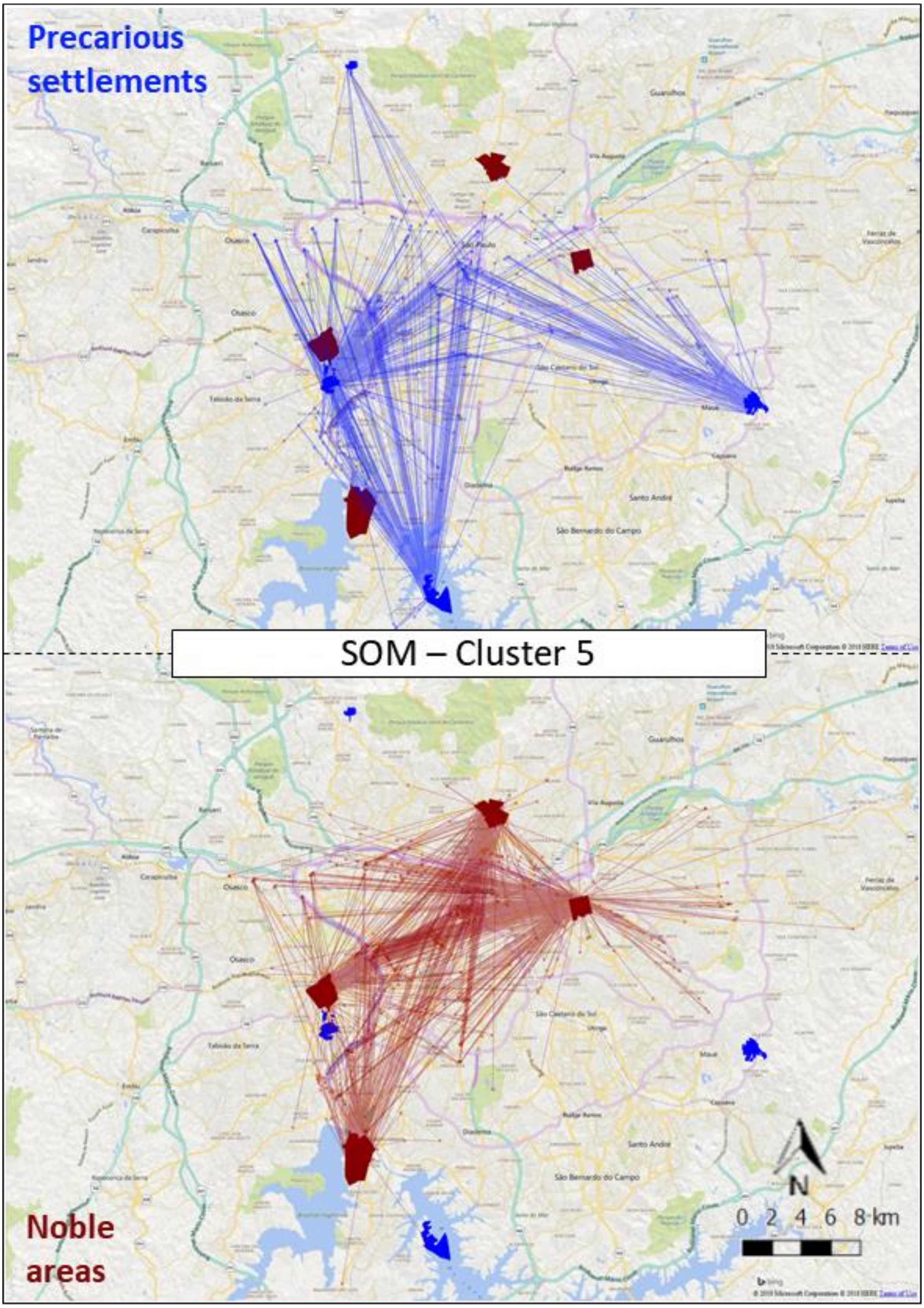
Source: The authors' own elaboration

Figure 70 – Activity distribution by cluster – SOM / Cluster 4



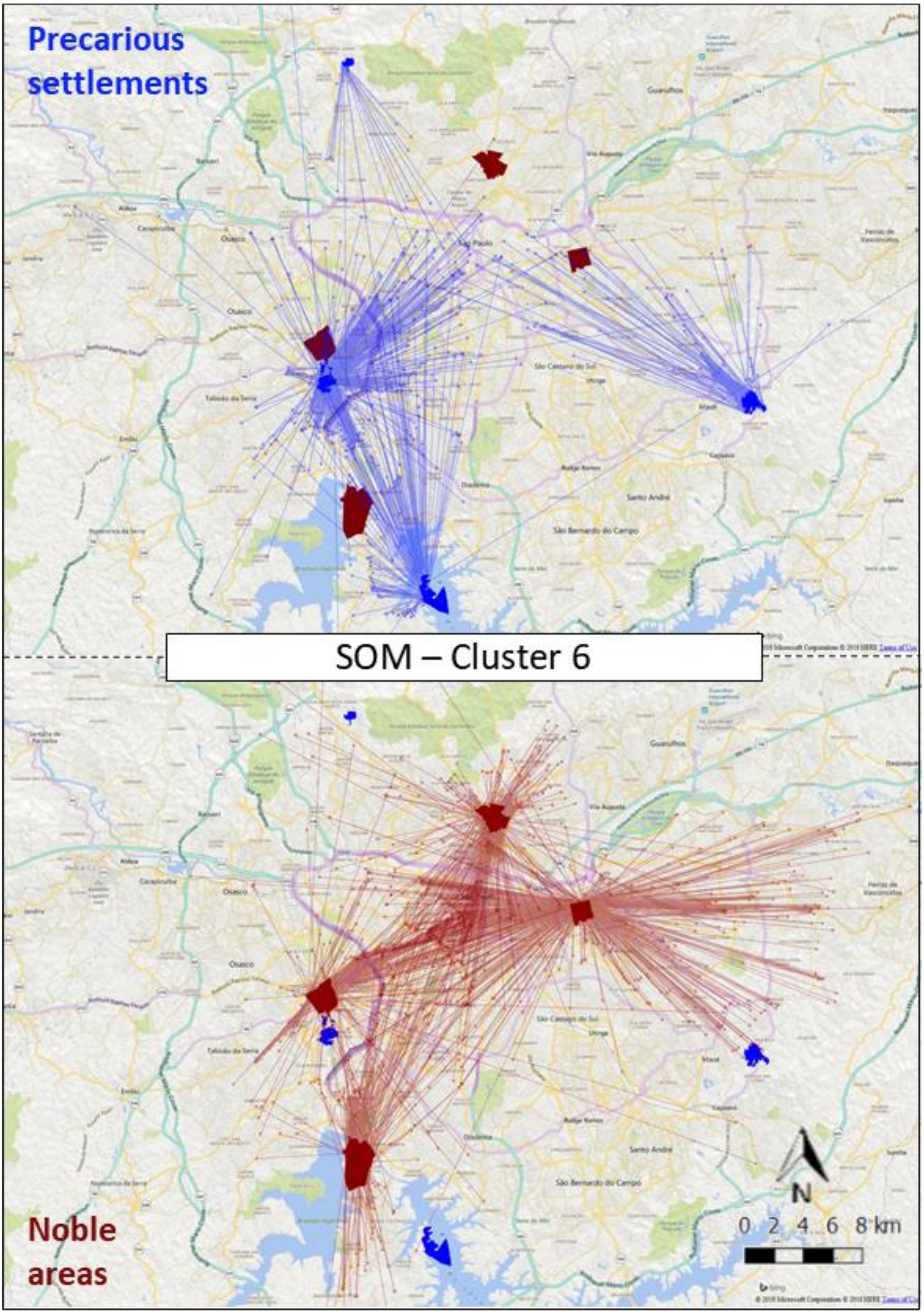
Source: The authors' own elaboration

Figure 71 – Activity distribution by cluster – SOM / Cluster 5



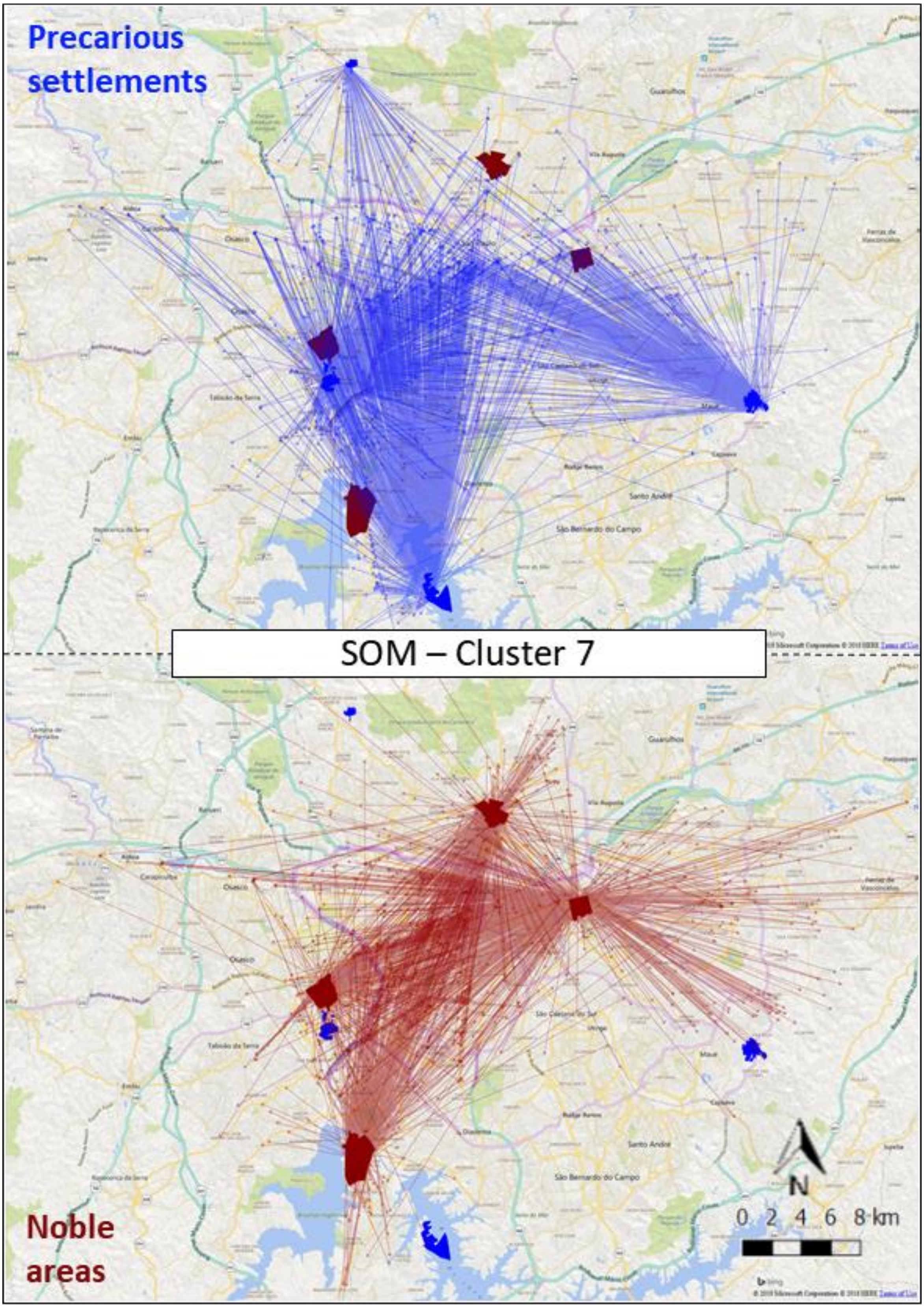
Source: The authors' own elaboration

Figure 72 – Activity distribution by cluster – SOM / Cluster 6



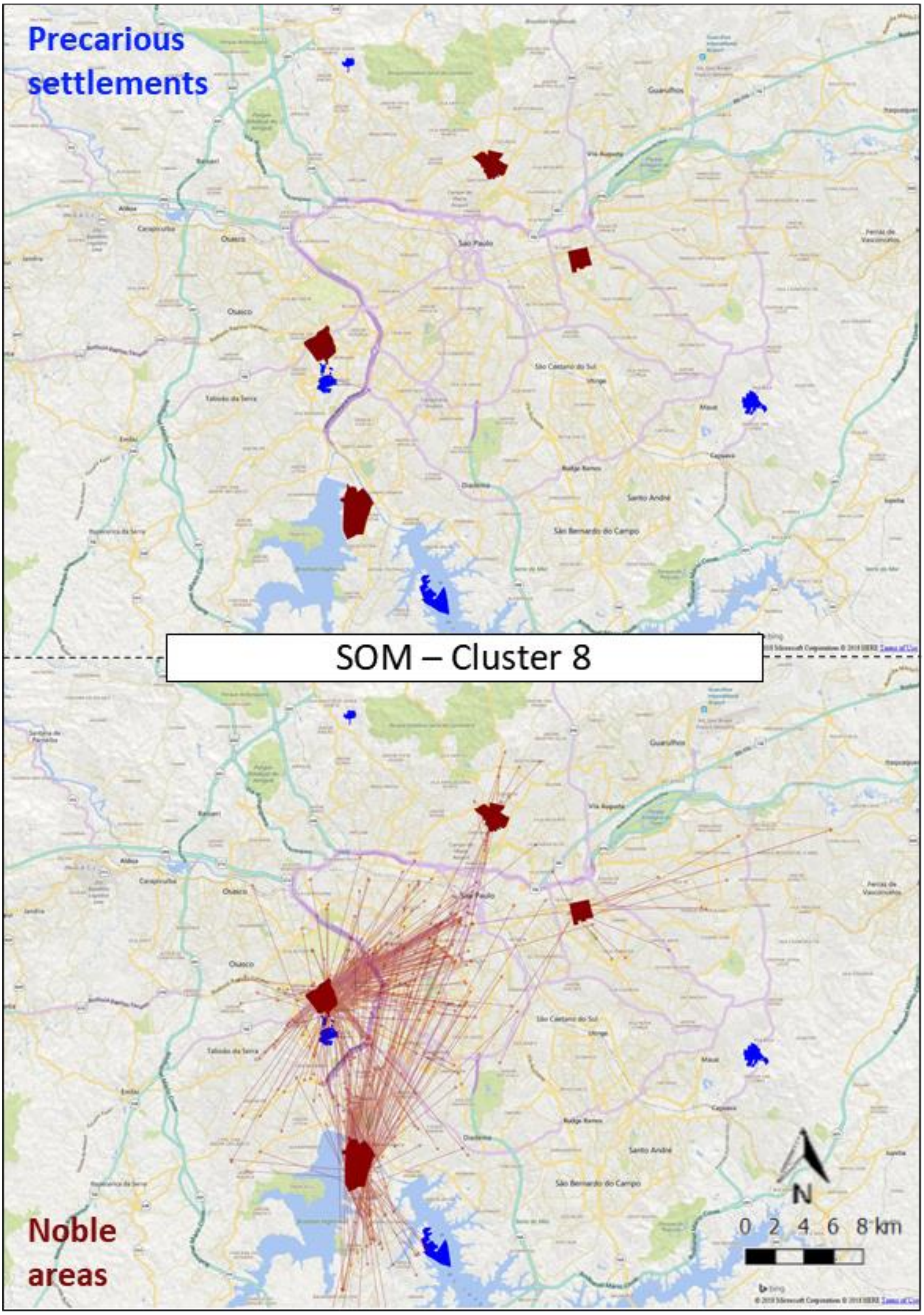
Source: The authors' own elaboration

Figure 73 – Activity distribution by cluster – SOM / Cluster 7



Source: The authors' own elaboration

Figure 74 – Activity distribution by cluster – SOM / Cluster 8



Source: The authors' own elaboration