Spatial self-organization in Santiago: Methods and Applications

Raimundo Sánchez Undurraga

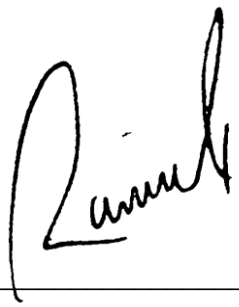
Faculty of Engineering and Science
Centre for Territorial Intelligence
University Adolfo Ibáñez

This dissertation is submitted for the degree of Doctor of Philosophy
In the subject of
Complex Systems Engineering
December 2015

Dedicado a la memoria de Ismael, Clemente, Rosita, Enrique, Nonato y Lemmy

Declaration

This dissertation is the result of original work, done in collaboration, except where specifically indicated in the text. It has not been previously submitted, in part or whole, to any university of institution for any degree, diploma, or other qualification.

Signed: 

Date: December 31th

Raimundo Sánchez Undurraga

Adolfo Ibáñez University

Abstract

Assembling spatial units into meaningful clusters is a challenging task, as it must cope with a consequential computational complexity while controlling for the modifiable areal unit problem (MAUP), spatial autocorrelation and attribute multicollinearity. Nevertheless, we sustain that these effects can reveal significant interactions among diverse spatial phenomena, such as segregation and economic specialization, but most methods treat this apparent disorder as noise.

In order to address this issue, we have developed a hierarchical regionalization algorithm that is sensitive to scalar variations of multivariate spatial correlations, recalculating PCA scores at all aggregation steps in order to account for differences in the span of autocorrelation effects for diverse variables. In such a way, we intend to provide a method that minimizes the information loss associated with both MAUP zoning and scale effects, while providing results that allow studying the self-organization of spatial patterns avoiding arbitrary zoning decisions. This algorithm produces a hierarchical cartography, which has multiple applications, where two particular cases were studied in Santiago de Chile.

With these settings, the scalar evolution of several social distress measures is compared between empirical and 120 random datasets. Remarkably, adjusting several indicators with real and simulated data allows for a clear definition of a stopping rule for spatial hierarchical clustering. Indeed, increasing correlations with scale in random datasets are spurious MAUP effects, so they can be discounted from real data results in order to identify an optimal clustering level, as defined by the maximum of authentic spatial self-organization. This allows to single out the most socially distressed areas in Greater Santiago, thus providing relevant socio-spatial insights from their cartographic and statistical analysis, which agrees to independent diagnostics

On the other hand, despite the abundance of works in hedonic mass appraisal, the potential of implementing hierarchical structures to market segmentation has not been fully explored. The purpose of this research is to fill this gap in the literature by studying the impact of incorporating complex architectures to predictive models, such as: econometrics models, artificial neural networks and hybrid models of combined forecasts. Our results confirm that all models exceed their predictive capability when applied in a hierarchical framework

In sum, a useful methodology is developed to systematically explore the black box of spatial interdependence and multiscalar self-organizing phenomena, while linking these questions to relevant real world issues.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Dr. Marcelo Villena for the continuous support of my Ph.D. study and related research, for his friendship, motivation, and immense knowledge. His guidance helped me in all the time of research and development of this thesis.

Besides my advisor, I would like to thank Dr. Matias Garretón, for the stimulating discussions, for those endless work sessions in several parts of this research, and for all the fun we have had working together in the last years.

My sincere thanks also goes to Dr. Luis Valenzuela, and all the center for Territorial Intelligence, who provided me an opportunity to join their team, and who gave access to their facilities and information. Without they precious support it would not be possible to conduct this research.

Thanks to Dr. Alejandro Jadresic and the Engineering and Sciences Faculty for hosting me for all these years and for giving me so many opportunities.

Also I thank my friends in the PhD program, both professors and classmates. In particular, I am grateful to Dr. Martin Hilbert for enlightening me the first glance of research and helping me with doctoral fellowship.

Last but not the least, I would like to thank my family for supporting me throughout this process and in my life in general.

Contents

1. Introduction

In the past 30 years the advance of geographic information systems (GIS) has enabled the capture, storage and processing of large volumes of spatial data, as never seen before in human history and science. These advances have allowed us to develop models to better understand the dynamics that exist in our geographic space and we could not see in detail because of technological and methodological limitations. The development of spatial analysis methods in recent years has facilitated the creation of several specialized magazines and research centers, graduate programs and even funding sources. The study of the properties of geographical space is a fertile line of research that directly impacts on urban planning and productive development.

1.1. Spatial self-organization

A natural laboratory for the study of these complex phenomena is the city, where several processes show emergence behavior with no apparent central planning. Segregation processes offer a good example of these issues, evolving with self-sustaining dynamics that involve correlated attributes which are locally reinforced (Schelling 1969, Massey & Denton, 1988; Fujita & Thisse, 2013). The interaction between these processes and the geographic space produces dancing landscapes (Miller, 1991), which makes it more difficult to predict. Therefore, there are great opportunities in the development of data based predictive tools for understanding self-organized processes within the city.

The case of Greater Santiago (GS) provides a conspicuous illustration of the historical production of cumulative socio-spatial inequalities at a metropolitan scale (De Mattos, 2002; Hidalgo, 2007). It is clear that income inequality can be traced to unequal access to resources, especially land and education, and also shows an unequal provision of infrastructure and facilities, such as health care and welfare state arrangements. When inequality is analyzed by administrative division, some weak patterns emerge, as can be seen in Figure 1. There are districts around the center of Santiago that have high levels of inequality like the municipalities of *Santiago*, *Providencia* and *Ñuñoa*. In contrast, the northeast sector of the capital has the lowest levels of inequality in general. Segregation measures are strongly affected by the scale of data aggregation, potentially leading to severe biases when comparing regions of different sizes (Krupka, 2007). Aggregating data in administrative units hides relevant information regarding high/low inequality clusters, and other phenomena that occurs at different scales. The appropriate definition of spatial boundaries is a major challenge in geographic analysis (Gehlke & Biehl, 1939; Openshaw & Taylor, 1979; Guo, 2008; Duque et al, 2012), so complex systems like this should be analyzed from its most basic unit.

Besides its computational complexity, this task must consider a combination of three interdependent spatial effects that are unaccounted for in standard statistical methods. These are the 'Modifiable Areal Unit Problem' (MAUP), spatial autocorrelation and local

coproduction of different attributes, which leads to multicollinearity (Lefebvre, 1974; Openshaw & Taylor, 1979, Anselin, 1995). Rather than considering these topological effects as error sources, we sustain that they provide relevant information about how spatial patterns self-organize from apparently disorganized social phenomena.

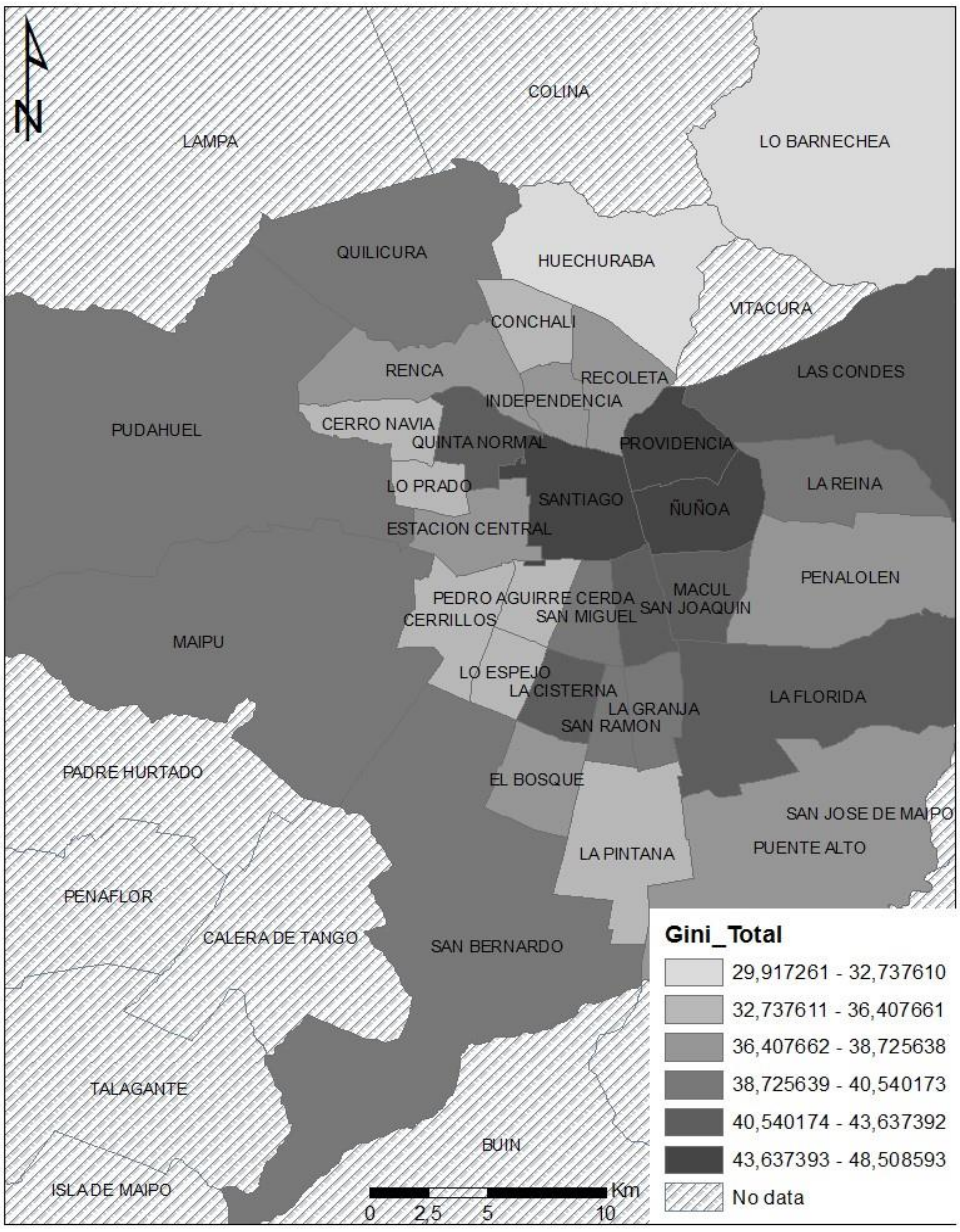


Figure 1: Spatial distribution for Gini index by districts (Source CASEN 2011)

1.2. Objectives

Regionalization, understood as a method for partitioning space in homogeneous and geographically continuous zones, is a convenient strategy to address the aforementioned issues. Remarkably, just before providing a rigorous analysis of MAUP (Openshaw & Taylor, 1979), Openshaw (1977) developed a spatially constrained hierarchical algorithm, explicitly stating the relationship between aggregation biases and optimal-zone design. As can be seen in Table 1, most of prior and subsequent research on spatial clustering has been focused on the development and improvement of a wide variety of algorithms, including contiguous regions restrictions and multiple attributes, but none offers a proper dealing of the MAUP scale effect (Berry, 1961; Lankford, 1969; Monmonier, 1973; Perruchet, 1983; Openshaw & Rao, 1995; Duque et al, 2007; Guo, 2008; Mu & Wang, 2008). Works like Clark et al. (2015) propose strategies to deal with grouping biases in conventional clustering, so we aim to extend that logic to regionalization algorithms.

Table 1: Summary Table

Authors	Spatial information	Conjoint regions	Multiple Attributes	Scale adjustment
Openshaw, 1977	Yes	No	No	No
Perruchet, 1983	Yes	No	No	No
Openshaw & Rao, 1995	Yes	No	Yes	No
Webster & Burrough, 1972	Yes	No	Yes	No
Mu & Wang, 2008	Yes	Yes	Yes	No
Clark et al, 2015	Yes	No	Yes	Yes
Duque et al, 2007	Yes	Yes	Yes	No
Guo, 2008	Yes	Yes	Yes	No
Spielman & Folch, 2015	Yes	Yes	Yes	No

Source: Author

In order to address these questions, we have developed a hierarchical regionalization algorithm designed for parallel bottom-up clustering from local minima, in iterative steps that construct successive scale levels. In such a way, we intend to provide a method able to minimize the information loss associated with both MAUP zoning and scale effects, while providing results that allow studying the self-organization of spatial patterns avoiding arbitrary zoning decisions. Our aim is to capture autocorrelation effects inside of self-produced units, which would minimize the internal variance of multiple attributes while simultaneously maximizing the variance among units. In this way, the information of spatial patterns would be preserved, instead of treating this apparent disorder like noise.

Therefore, this work has several specific objectives. First, we seek to overcome the aforementioned spatial effects by a regionalization method, second and third, to apply this algorithm to analyzing 2 different case-studies in Greater Santiago, hedonic prices

estimation and social distress diagnosis. Both applications are sustained on the hypothesis that the selected variables are spatially coproduced, yet with varying strengths and combinations that are scale-dependent. Therefore, a fourth specific objective is to highlight the relevance of MAUP and spatial self-organization for a better understanding of regionalization methods.

This thesis is organized as follows: a review on regionalization methods; case study: identifying the best level of analysis for social distress; case study: hierarchical systems for hedonic appraisal; and a discussion on the main findings and research perspectives.

2. Regionalization methods

Regionalization is as a process of space partitioning in homogeneous and geographically continuous zones, through the optimization of an objective function under constraints, while guaranteeing that each elementary entity is unambiguously assigned to one zone (Openshaw & Rao, 1995; Guo & Wang, 2011). Besides being appropriate to address MAUP, these methods are useful for optimal zonal design, improving spatial data aggregation for anonymity, for the statistical significance of the collected information, for spatial data mining or for an adequate cartographical representation (Openshaw, 1977; Pilevar & Sukumar, 2005; Duque et al, 2007).

2.1. The Modifiable Area Unit Problem

Geographic space is a dynamic matrix which can reinforce natural or social phenomena which take place in it and their interactions (Lefebvre, 1974). Thus, general assumptions of statistical independence do not hold in geographic analysis, mainly due to spatial autocorrelation and local multicollinearity. Auto-correlated variables can be self-organized into systematic patterns, as local attributes influence the reproduction of the same phenomenon in neighboring areas (Goodchild, 1986; Getis & Ord, 1992; Anselin, 1995).

For example, the arrival of high income residents usually contributes to an escalation of real estate prices in a neighborhood, increasing the odds for low income residents to leave (Smith, 2002). Local multicollinearity arises when different attributes are coproduced or are mutually interdependent. For example, unemployment tends to reduce income and can be related to higher crime rates, which may stigmatize neighborhoods, restricting job access and thus generating a vicious circle (Galster, 2012). In sum, spatial attributes can be influenced by themselves and by correlated variables, biasing statistical analysis and generating spurious regression coefficients (Lauridsen & Mur, 2006; Mur et al., 2010, Openshaw & Taylor 1979).

Geographers have been aware of these issues at least since Gehlke & Biehl (1939) observed that grouping had major effects over correlation coefficients using US census information, which were consistent with those observed in random data. These kind of effects were systematically analyzed by Openshaw & Taylor (1979), who coined the term 'Modifiable Areal Unit Problem' (MAUP). In fact, "when data are gathered according to different boundary definitions, different data sets are generated. Analyzing these data sets will likely provide inconsistent results" (Wong, 2004:571). This problem arises either if different entities are modified while maintaining a similar size - the zoning effect - or if smaller units are aggregated into larger units - the scale effect -. Both aspects of MAUP are intertwined with spatial autocorrelation and local multicollinearity. Indeed, an auto-correlated variable may present high average values in a small unit that contains a local concentration, while being diluted in a larger area, leading to a scale effect. Besides, two overlapping units of the same scale, one fully encompassing a local concentration and the other containing just a portion of it, would have different densities of the same variable, a zoning effect. Both observations also hold for a set of correlated variables, thus producing multivariate MAUP

effects through local multicollinearity. In sum, a theoretical connection exists between spatial interactions and the statistical inconsistencies produced by MAUP.

Firstly, spatial phenomena usually show continuous geographic variations, but the information used to measure them is often gathered or aggregated in arbitrary boundaries. On the one hand, important biases can be thus introduced in spatial analysis. On the other hand, spatial data aggregation can be an unavoidable procedure in order to ensure anonymity or statistical significance of collected data. In any case, determining the shape and scale of a geographic entity is a difficult problem, for which there might be different optimal solutions, depending on the variables which are considered.

Secondly, the spatial distribution of a variable can be self-organized in systematic patterns, produced by the effect of local attributes over the same variables in neighboring areas. For example, the arrival of high income residents usually contributes to an increase in real estate prices in a neighborhood, increasing the odds for low income residents to leave and to be replaced by higher income households (Smith, 2002). This well-known effect of spatial autocorrelation is observed in many geographic phenomena and violates the assumption of independence among values of different observations, which is necessary for standard statistical analysis (Getis & Ord, 1992; Anselin, 1995; Amaral & Anselin, 2013).

Spatial autocorrelation, either positive or negative, is intertwined with the MAUP. As the influence of an attribute over its own spatial distribution varies with distance, the effect of local concentration could increase the mean value of this variable in a small unit, while being diluted in a larger area, leading to a MAUP scale effect. Besides, two overlapping units of the same scale, one fully encompassing a local concentration and the other containing just a half of it, would have different distributions of the same variable, a MAUP zoning effect. Thus, a direct theoretical connection can be established between autocorrelation effects and the statistical inconsistencies produced by the MAUP.

Thirdly, attributes in space can be mutually interdependent or coproduced. For example, unemployment tends to reduce income and can be related with higher crime rates, which may lead to stigmatization of certain neighborhoods, preventing its residents to get new jobs (Galster, 2012). Thus, spatial attributes can be influenced by their own distribution and by those of correlated variables. Owing to these biases, regression coefficients would be spurious results generated by spatial multicollinearity (Lauridsen & Mur, 2006; Mur et al., 2010). Hence, MAUP biases may affect the measurement of a whole set of attributes.

This brief account of three major issues in spatial statistics highlights the relevance of developing adequate methods for zonal systems design, in order to reliably determine homogeneous zones at multiple scales (Duque et al, 2007; Guo & Wang, 2011, Mu & Wang, 2008). Particularly, the measurement of segregation and related urban phenomena is very sensitive to the spatial definition of statistical aggregates, as socially uniform areas may be well represented by entities such as census tracts in some cases while being inadequately mingled in others (Krupka, 2007). Thus, the definition of homogeneous areas can be useful to produce more accurate estimates of diverse spatial indicators (Spielman & Folch, 2015), while revealing patterns of spatial autocorrelation and local multicollinearity.

Reciprocally, the analysis of self-organizing spatial phenomena is fundamental to understand the behavior of regionalization algorithms, which are conceptually related to the well-known Ward's (1963) hierarchical clustering method, but adding the additional restriction of producing geographically continuous clusters at all scales.
In the next section, we will discuss several lines of research that have considerably advanced in this topic.

2.2. Literature review

Regionalization is a particular case of spatial clustering, which stems from general data clustering methods. The partitioning logic is essentially the lenses with which we perceive reality, which varies among different complex systems. This starts with philosophy over Plato's cave and Kant, and goes from probability theory (Kolmogorov's generating partition), or number theory (combinatorics), to dynamical systems, psychology and neuroscience, biology, social networks, artificial intelligence and many others.

Several statistical approaches have been adapted to spatial clustering, without satisfying regionalization constraints. Two-step procedures generate homogeneous groups through statistical clustering and then assemble the contiguous units from the same types, usually producing fragmented aggregates (Fischer, 1980; Openshaw, 1973). Standard clustering algorithms have been applied to spatial entities, combining their geographic coordinates with other attributes, thus increasing the heterogeneity of the clusters or tending to produce circular regions (Murray & Shyhy, 2000; Webster & Burrough, 1972). Henriques et al (2012) propose an interesting variation of these approaches using Kohoonen neural maps, and subsequent treatment of their output space can improve the results (Feng et al, 2014). Density-based and grid-based algorithms aggregate points or areas which are contained under a suitable density threshold (Hartigan, 1975; Pilevar & Sukumar, 2005; Sander et al, 1998). These methods are able to detect arbitrarily shaped clusters, but they are very sensitive to the selected threshold (Kriegel et al, 2011) and a proportion of the observations may be classified as outliers.

Recent works have developed an interesting approach to spatial clustering, considering multiscalar context measures around singular locations. Spielman & Logan (2013) use individual data of a nineteenth century census to elaborate profiles describing ethnical and socioeconomic variations with distance, around each person. Then, each location is assigned a probability of belonging to six classes through a model-based clustering procedure, allowing defining neighborhoods' cores and edges. Clark et al (2015) provide a detailed description of Los Angeles' changing segregation patterns, measuring racial composition in increasing scale aggregates around individual locations, performing factor analysis of these multiple measurements and clustering blocks in 20 categories, depending on homogeneity and ethnicity. These approaches provide rich substantial descriptions of urban phenomena, but their capacity to identify geographical patterns depends heavily on the spatial autocorrelation of the variables under study, and they do not guarantee the definition of a consistent geographical partition.

Regionalization algorithms differ from the aforementioned methods by their capacity to produce a complete spatial partitioning with geographically continuous clusters. This goal is attained through neighborhood constraints over the aggregation process (Openshaw, 1977). Considering strictly contiguous entities, rook neighbors are the ones that share one edge and queen neighbors include the former plus pairs that only share one point of their perimeters (Perruchet, 1983; Mu & Wang, 2008). More flexible neighbor definitions can be implemented through distance thresholds (Perruchet, 1983; Sander et al, 1998). Two main neighborhood-constrained approaches have been developed: partitioning and hierarchical regionalization (Berkhin, 2006; Guo, 2003).

Partitioning regionalization algorithms extend methods akin to k-means clustering (Hartigan & Wong, 1979), aiming to divide a data set into a predefined number of groups, while optimizing an objective function (Openshaw & Rao, 1995; Duque et al, 2012). Initial feasible solutions can be elaborated through random zoning or from a set of seeds, to which neighboring areas are reallocated or added until a predefined criterion is satisfied (Nagel, 1965; Openshaw, 1977). As checking all possible aggregate combinations is computationally infeasible in large datasets, these methods rely on exact optimization approaches or on a variety of heuristics - such as local search, simulated annealing and tabu search - in order to find an optimal solution (Duque et al, 2007; Guo & Wang, 2011).

A great diversity of algorithms[1] have been proposed for partitioning regionalization, progressively improving accuracy and computational efficiency (Duque, 2004; Nagel, 1965; Openshaw, 1977; Openshaw & Rao, 1995; Vickrey, 1961). Duque et al (2012) have proposed an interesting alternative to the arbitrary definition of a number of clusters, substituting this parameter with a population threshold, thus circumventing the optimal scale definition problem rather than resolving this issue. An extension of this approach has also proven to be a useful procedure to aggregate regions in order to improve the accuracy of survey data estimates (Spielman & Folch, 2015). However, as partitioning methods rely on arbitrarily predefined numbers of regions or population thresholds, this approach does not allow to efficiently address the question of determining an optimal scale or number of clusters[2].

Hierarchical regionalization algorithms generate a nested chain of spatially contiguous clusters - which can be represented as a tree or a dendrogram -, while optimizing an objective global function akin to Ward's (1963) method, or following local optimization criteria based on different measures of similarity (Carvalho, 2009; Lankford, 1969). These methods can either adopt a bottom-up strategy, aggregating units towards an all-encompassing region, or a top down approach, subdividing one area into smaller subsets (Monmonier, 1973). Bottom-up aggregation is most commonly used, joining the two contiguous units that either minimize the total heterogeneity increase, other objective functions (Openshaw, 1973), or which are the most similar neighbors (Lankford, 1969).

[1] Duque et al (2007) provide an exhaustive review of partitioning regionalization methods.

[2] Theoretically, this could be done through repeated partitioning tests at every aggregation level, but the computational cost would be enormous with large datasets, compared to the nested multiscalar structure that can be produced by a single run of hierarchical algorithms.

Several local similarity criteria have been described (Carvalho, 2009; Guo, 2008). Single linkage joins the clusters that contain the most similar pair of basic units, tending to produce heterogeneous groups which are linked by a series of close pairs. Complete linkage is focused on the most different units between two clusters, generating aggregates where all observations are similar to each other, while being strongly affected by outliers. Average linkage considers the average dissimilarity of all cross-cluster pairs of units, being less biased by outliers and having better performance than single and complete linkage (Carvalho, 2009).

We have focused on hierarchical regionalization, because it produces nested solutions at different scales. However, this approach has two important drawbacks (Berkhin, 2006). First, there are no clear rules to determine an optimal number of clusters, which is precisely the problem we aim to resolve from a scalar perspective. Second, solutions at higher scales are dependent on the mergers which have been performed in previous steps, which can lead to suboptimal configurations. Mu & Wang (2008) have developed a regionalization algorithm that can attenuate this problem, as it works by parallel aggregation from a set of local seeds defined by a local minima criterion. When all of the units have been assigned to a cluster they are merged in order to form a new layer, iterating this process until it converges in one unit. In such a way, dependence on prior decisions is limited to the lineage of each cluster and is independent from distant local aggregates. Moreover, Mu & Wang introduce a variant of average linkage, using factor analysis to synthetize multiple attributes in score that defines dissimilarity among units. This procedure and other PCA-based variants are particularly useful to calculate dissimilarities with spatially correlated variables, because they are designed to control for multicollinearity (Abdi & Williams, 2010; Spielman & Folch, 2015; White et al, 1991).

Hybrid hierarchical and partitioning regionalization algorithms follow a connect-and-divide strategy, generating a contiguity-constrained hierarchical clustering graph and then performing a top-down partitioning of this structure (Guo, 2008; Guo & Wang, 2011). The hierarchical step allows to efficiently integrate a contiguity constraint, reducing the computational complexity of the following procedures. Then the partitioning process optimizes an objective function, such as total sum of squared differences, and can introduce additional constraints, such as a minimum population. This combination improves the efficiency and accuracy of the regionalization process (Guo & Wang, 2011), but it does not resolve the question of determining an optimal number of clusters.

In sum, considerable progress has been made on improving regionalization methods, particularly for optimizing space partitioning into a given number of regions or regions of a given size, addressing the MAUP zoning problem. However, the question of determining the best scale of analysis remains unsolved, so there is no clear strategy to cope with the MAUP scale effect. This is a general problem of all clustering methods, statistical and spatial, and we suggest a neat solution for the latter cases. Relevant non-spatial approaches to this question will be discussed in the next section.

2.3. Geographic self-organized maps, graphs and scale-space clustering algorithms.

Spatial clustering in homogeneous regions is a challenging task in geographic data analysis, due to the issues discussed above and to the difficulty of combining two different kinds of proximities: geographic coordinates of geographic space and other variables in an attribute space (Henriques et al, 2012). Three kinds of methods have been proposed to address this question (Duque et al, 2012).

Firstly, two-stage approaches identify attribute clusters and then aggregate the geographically contiguous areas which belong to the same clusters. This method is heavily dependent on the attributes' spatial patterns and does not ensure the spatial contiguity of the resulting regions (Openshaw & Rao, 1995). Secondly, several methods consider the coordinates of the centroids of the units along with other attributes. In this case the spatial contiguity depends on the weight of the spatial information, if it is too low it can produce discontinuities, if it is too high heterogeneous circular regions may result (Perruchet, 1983). Thirdly, several algorithms perform spatially constrained homogeneous clustering, by limiting the possible unions to units that have specific neighborhood relationships, most frequently shared borders (Lankford, 1969; Mu & Wang, 2008; Duque et al, 2012).

An extensive discussion of the variants within this last type goes beyond the scope of this article, but we consider this kind of strategy to be the most adequate to our objectives, mainly because it ensures spatial contiguity among the resulting clusters. Instead, we will discuss three recent approaches which are particularly interesting.

The Geographic Self-Organizing Maps (Geo-SOM) incorporate spatial constraints to the Kohonen (1982) neural maps (Bacao et al., 2005; Henriques et al., 2012). This involves training an artificial neural network, which consists of an input space containing individual observations and their attributes, and of an output space which is a grid of neurons. The cartographic coordinates and other variables are treated separately, training the neurons with geographical proximities before mapping the attribute patterns. By iteration, data is mapped into the output space, weighting the observations by their evolving proximity to each neuron, leading to a flat representation of the multidimensional topology of the input space.

This method seems promising for big data analysis but until now it has been solved with rather low spatial resolution and the determination of the appropriate clusters depends critically in a posteriori treatments of data in the output space (Henriques et al., 2012; Feng et al., 2014). Moreover, neural network training is an opaque process, where is very difficult to recover the information produced at different stages, and the single layer output does not allow for interscalar analysis.

A different approach relies on graph theory clustering (Calinski & Harabaz, 1974). It follows a connect-and-divide strategy, where a set of spatial units is first linked by a tree structure, building links that optimize a similarity criterion among units that satisfy spatial neighborhood constraints (Duque et al, 2012). Thus, an exhaustive structure is obtained, which is then subdivided into subgraphs, with an algorithm that minimizes attribute

heterogeneity while maximizing the number of groups that satisfy a general constraint, such as a minimum population in each cluster.

This procedure offers interesting insights into the problem of efficiently linking spatially contiguous units while respecting homogeneity controls. Moreover, even if Duque et al (2012) rely on exogenous criteria for determining the optimal number of clusters, an endogenous determination of the best number of partitions seems feasible (Calinski & Harabaz, 1974). However, the graph approach has one fundamental limitation regarding scalar nonlinearities of spatial effects. In fact, as the subdivision order is unknown during the linking stage, all of the connections must be determined by close-range interactions, so relevant large-scale effects might be overlooked.

A third method for spatial clustering, which is closer to our objectives, has been developed by Mu & Wang (2008). This algorithm aggregates spatial units to their most similar neighbors, creating a new layer of clusters, which is used as the substrate for a new round of aggregation and so forth, until all units are melted in one cluster. In this way, major steps of the process and its associated information can be preserved for further analysis.

In Mu & Wang's method, the similarity among adjacent units is defined by an attribute score produced by multidimensional factor analysis of several socioeconomic variables. This allows finding local minima, the units that most closely resemble their set of neighbors, and local maxima. Then, local minima are linked to their most similar neighbor, which is linked to its most similar one, and so on until a local maxima is reached (Mu & Wang, 2008). Thus spatially contiguous clusters are produced, while providing a multiscalar set of embedded clusters. Nevertheless, we consider that two important improvements can be done with a related algorithm design.

Firstly, the strategy of linearly linking local minima and maxima seems inappropriate for geographical purposes, as it can induce clustering among dissimilar units. Conversely, a more flexible grouping procedure can be a better way to capture fractal frontiers between adjoining clusters. Secondly, defining similarity by an average score of factors, with fixed weights at different scales, risks overlooking relevant nonlinear distance effects and might be an inaccurate weighting schema at higher levels of aggregation. As opposed to the graph approach, this limitation can be solved with a related strategy.

These issues and the proposed improvements will be detailed in the next section, which develops a novel algorithm for multiscalar spatial clustering.

2.4. A local-hierarchical regionalization algorithm

Building on Mu & Wang's (2008) approach, we develop a simpler local-hierarchical regionalization algorithm with two relevant modifications: a more flexible neighbors' definition and a recalculation of orthogonal scores at each scale. These adjustments will be explained within a brief general description of the clustering process (Figure 2).

Starting at block level, a set of neighbors is defined for each unit $i \in I$ - I being the set of entities at any level -, generating a binary matrix that defines the aggregation constraint. Blocks are the smallest urban areas separated by streets and their perimeters are irregular

shapes, so it is unfeasible to use shared-borders procedures. Thus, we define as neighbors all the entities which have any pair of points of their perimeters under a distance threshold (Perruchet, 1983), which starts at 20 meters[3]. This distance reaches over standard GS streets but remains under the span of the smallest blocks, thus preventing to assign non-immediate neighbors. This threshold is proportionally increased towards higher scales, attaining a maximum span of 72 meters, reaching over the widest avenues, rivers and other topographical barriers. Thus, a realistic neighborhood constraint is implemented, as adjacency criteria evolves with scale.

The attributes of each unit (Table 2) are normalized and processed by principal component analysis (PCA), obtaining a set of K partial scores (s_{ik}) for each entity i. These orthogonal vectors preserve information while controlling for multicollinearity, allowing for an optimal differentiation among units (Abdi & Williams, 2010; Cutter et al, 2003). The eigenvalue of each score k accounts for a proportion p_k of the total variance among units. As we have selected a set of positively correlated variables (Table 1), each unit can be characterized by an aggregated social distress score (SDS_i) which is an eigenvalue-weighted sum of partial scores:

$$SDS_i = \sum_{k=1}^{K} p_k \, s_{ki}$$

As opposed to Mu-Wang's method, which uses a similar data treatment strategy but collapse three factors in a weighted average, we treat the scores independently for each pair of units. This allows capturing the local variations of different partial scores, because the vectors accounting for most overall variance can be homogeneous in small areas. In such a case, their scores will be mutually annulated and the corresponding units will be differentiated by differences in lower-ranked scores.

Likewise, the dissimilarity among 2 neighbors i and j can be measured as a multidimensional distance of scores which can be calculated either as a sum of absolute ($absDS_{ij}$) or squared ($sqrDS_{ij}$) score differences (SDS is written as S, for simplicity):

$$absDS_{ij} = \sum_{k=1}^{K} p_k \, abs(s_{ik} - s_{jk}) \qquad sqrDS_{ij} = \sum_{k=1}^{K} p_k (s_{ik} - s_{jk})^2$$

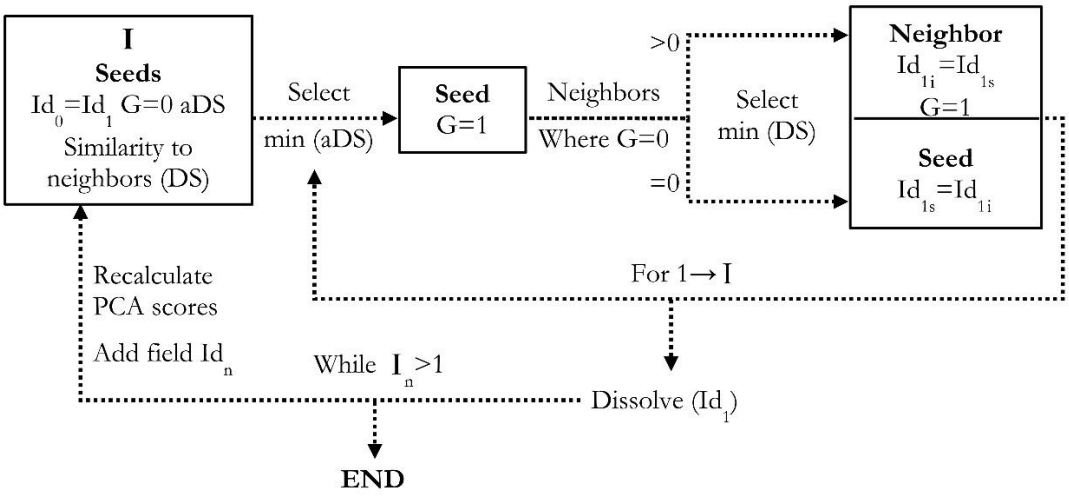
Both difference definitions have been tested for the GS, and each produces a different aggregation behavior, as will be detailed in section 4.1. The multidimensional distances between a unit and each of its neighbors are averaged in order to obtain a local similarity index (aDS_i), which allows ranking all units from the most locally similar to the most locally dissimilar:

[3] This flexible neighborhood definition functions in a similar way to the queen adjacency criterion, and allows working with discontinuous entities or imperfectly drawn shapefiles. Mu & Wang (2008) used census tracts, which are designed as a continuous lattice, and implemented a more constrained rook-neighbor definition.

$$aDS_i = \frac{\sum_{j=1}^{I} DS_{ij}}{I}$$

This set of 'seeds' provides a baseline for the clustering algorithm to proceed (Figure 2). Each unit is given identification variables ($Id0_i$ and $Id1_i$), a grouping marker (G_i), an arbitrary number of attribute variables and their corresponding score (SDS_i). Attribute distances to each neighbor $\left(absDS_{ij} \,\&\, sqrDS_{ij}\right)$ and a local similarity index (aDS_i) are computed at each round. In a round 'n' each unit becomes a 'seed' only once, giving priority to local minima. Each 'seed' selects the most similar neighbor among unmarked ones ($G_i = 0$), marks it ($G_i = 1$) and alters its secondary Id (Idn_i=Idn_{seed}). If the current 'seed' has been previously grouped, it will transfer the Id of the first 'seed' in the cluster. If no unmarked neighbors are available, the 'seed' will adopt the Id of its most similar one, thus avoiding orphan units.

Figure 2: Logical model of a local-hierarchical regionalization algorithm



Source: authors. The main improvement over previous algorithms is the recalculating PCA step that assumes that magnitudes in multiple correlations are scale-dependent

Performing as many iterations as there are spatial entities, each unit becomes a 'seed' only once, in an ordered way starting from the lowest mean score differences. This provides a flexible[4] seeding criterion that gives priority to local minima and ensures that the results can be replicated[5]. Each 'seed' selects the most similar neighbor among unmarked ones (G=0), marks it (G=1) and alters its secondary Id (Id1n=Id1s). If the current 'seed' has been previously grouped, it will transfer the Id of the first 'seed' in the cluster. If no unmarked neighbors are available, the 'seed' will adopt the Id of its most similar neighbor (Id1s=Id1v). In such a way, no orphan units are left at the end of each round.

[4] An entity which has been previously marked by another can also be selected and the aggregation process does not necessarily follow a local minima-maxima direction.

[5] As opposed to random seed selection.

Then, the base units are dissolved into clusters that share the same secondary Id, a new Id field is added to accommodate the grouping information for the next level and a new clustering round is started. Next, units are merged by secondary Id, attribute variables are combined as weighted averages[6], a new set of PCA scores is computed at the following scale. In contrast to Mu-Wang's method, this procedure is performed at each level, in order to account for the most relevant spatial interactions among variables at different scales. For example, crime rates' spatial autocorrelation can be particularly strong at block level, but agglomeration economies can be sharper at a metropolitan span (Andresen & Linning, 2012; Fujita & Thisse, 2013).

This process iterates until all units are merged into one cluster. The information of the scores, averaged variables and the geometries of the units are preserved in tables and shapefiles after each round. The phylogeny of the clustering process is registered in a final table containing the original Id and the secondary Id of each level for all the original units. All this data can be used for subsequent data mining. We don't know how the probability distribution of the different variables might affect the results, but this issue will be addressed in further research

The updating of PCA scores at each aggregation round is a key improvement of the proposed algorithm, allowing capturing nonlinear and unpredictable effects of the spatial behavior of attributes at different scales. However, this procedure is computationally demanding and excludes the implementation of smoother step-by-step hierarchical clustering strategies more akin to Ward's (1963) method. Even if this involves a rather coarse graining between successive scales of analysis, it is an adequate trade-off as we are more interested in studying spatial effects than in the refinement of grouping techniques. In the next chapters two case studies with applications of this algorithm are presented where the ability of the method to recover spatial information is put to test.

[6] By population, area, perimeter or any other appropriate parameter.

3. Case Study: Identifying the best level of analysis for social distress[7]

In general, cluster analysis aims to classify large sets of observations into groups that are internally homogeneous while maximizing the differences among groups (Calinsky & Harabasz, 1974; Krzanowski & Lai, 1988). However, from this intuitive definition it is rather difficult to implement an objective stopping rule - understood as a definition of the optimal number of partitions - and a great variety of procedures have been proposed (Milligan & Cooper, 1985).

3.1. Determination of an optimal scale or number of clusters

Typically, in hierarchical clustering the average of any intra-group dispersion measure decreases as the number of groups increases (Tibshirani et al, 2001). Plotting this ratio usually leads to a curve with two different sections: a steep slope for small numbers of clusters and a rather flat descent for higher numbers (Salvador & Chan, 2004). The transition between slopes is called the 'knee', which is vaguely considered as an indicator of the best number of clusters (Thorndike, 1953), because a higher number of partitions would divide homogeneous clusters, while a lower number of divisions would join heterogeneous groups. However, this 'knee' is not always apparent and several methods aim to identify it, such as comparing differences, ratios or second derivatives of heterogeneity gains between successive aggregations, intersecting fitted lines or identifying distant points from fitted curves (Krzanowski & Lai, 1988; Milligan & Cooper, 1985; Salvador & Chan, 2004). Nevertheless, these methods are solely based in the internal homogeneity of the clusters, while other relevant parameters should be considered.

In a broad Monte Carlo evaluation of 30 stopping rules, Milligan & Cooper (1985) identified a procedure developed by Calinsky & Harabasz (1974) as the best performer. These authors identify an optimal number of clusters through the highest value of the ratio

$$\left[{trace\ B}/{k-1}\right] \Big/ \left[{trace\ W}/{n-k}\right]$$

where B represents between groups heterogeneity, W is the total within groups heterogeneity, k is the number of partitions and n is the total number of items. Thus, this index searches a balance between a maximum of isolation among clusters and their minimum internal heterogeneity. Furthermore, Tibshirani et al (2001) show that, even for simulated data with no group structure, clustering processes are able to generate spurious groups. Thus, they develop a "gap statistic", identifying the optimal number of partitions by the maximum reduction of the observed within group heterogeneity compared to its expected value with a null distribution (Tibshirani et al, 2001).

[7] This chapter was published in collaboration with Dr. Matias Garretón (Garreton & Sanchez, 2016)

These issues have not been properly researched in the context of spatial clustering, although MAUP effects certainly have a strong impact on aggregation measures. A proper method to identify an optimal scale of regionalization should consider between and within group heterogeneity, while controlling for spurious correlations and aggregations. This question will be addressed in the fourth section of this article, after describing the regionalization algorithm and the dataset which will be used in the corresponding experiments.

In particular, our main objective is to define the best level of analysis for hierarchical regionalization methods, comparing the aggregation behaviors of empirical and random datasets. In fact, the increase of correlation coefficients with scale which is observed in spatial clustering with random data is a spurious effect, which can be discounted from observations with empirical data in analogous settings. This allows singling out the best level of analysis, defined by a maximum of authentic spatial self-organization, leading to an accurate diagnostic of socially distressed zones in GS. Thus, a second goal of this work is to develop a cartographic and statistical description of the most critical areas in this city, at the most appropriate analytical scale.

A case study that allows exploring these effects and the main question of determining the best level of analysis - in a real world setting - will be briefly outlined in the next section.

3.2. Social distress indicators

Combining 2012 Chilean Census data, available at person and household levels, six variables were calculated for each one of 47,414 blocks of GS. Three of these variables correspond to individuals' characteristics and three to housing conditions (Table 2). By definition, all the variables take values between 0 and 1. In addition, local crime densities for 2012 were calculated from data of the Interior Ministry of Chile[8], selecting four categories which concern urban violence (author et al, forthcoming). These variables were also normalized between 0 and 1. More attributes could be introduced, but an easily interpretable dataset will be used for this case.

Income data and socioeconomic level indicators have not been included, in order to have independent diagnostic criteria to ascertain the spatial accuracy of the clustering method. In fact, the comparison of the following results with other segregation studies shows a remarkable geographic coincidence of the most critical areas, identified with different methods and datasets.

[8] In the context of a research agreement with the Centre for Territorial Intelligence of the Adolfo Ibañez University.

Table 2: Selected indicators for social diagnosis

Variable	Description	Formula
Unemployment	Percentage of population willing to work but without employment	Unemployed/ Employed or willing to work
Dependence	Percentage of inactive or unemployed population	1- (Employed/Total population)
Uneducated	Inverse of education years for population older than 24 years	Population >24/ Sum of education years (>24)
Overcrowding	Average number of rooms for each inhabitant, calculated at household level	Mean (Rooms in residence/ Residents)
Precariousness	Percentage of shanty housing	Precarious accommodations/ Total accommodations
Insalubrity	Percentage of housing without formal sanitation systems	Insalubrious accommodations/ Total accommodations
High violence	Density of homicides, rapes and gravest injuries	High violence reports/ Area
Insurgence	Density of weapons-related offenses and aggressions to officers	Insurgence reports/ Area
Drugs	Density of drug-related crimes and offenses	Drug-related reports/ Area
Aggressions	Density of offenses against the person	Aggression reports/ Area

Source: Authors' elaboration with data from Census 2012 and the Interior Ministry of Chile.

The selected indicators have been constructed in order to assign higher values to the conditions which have negative social connotations (Table 2). This adjustment ensures that all of the variables are positively correlated with the attribute score at all clustering levels, allowing consistently differentiating and ranking the units by critical social conditions. This hierarchy is based on eigenvalue-weighted sums of PCA partial scores, a method that resolves the weighting problem which has been signaled by Cutter et al (2003) in a similar approach to diagnosing social vulnerability. In this GS' case study, the scalar specific attribute score thus calculated will be interpreted as a 'social distress score' (SDS).

This methodology has been applied to the identification of the best level of analysis, leading to an accurate diagnostic of socially critical areas in GS, as detailed in the following section.

3.3. Choice of an optimal analysis level

We will define an 'optimal' scale of multivariate spatial clustering as the level that represents the strongest coproduction of a set of attributes within and throughout the corresponding units. This is both related to evolving multicollinearity in the attribute set and to the consistency of the clustering process, which can be measured by intra-group compacity and inter-group isolation. Multicollinearity is quantified as the average of the absolute values of correlation coefficients[9] among the 10 variables which have been used to elaborate the SDS. In order to avoid variance biases of different variables, a Fisher Z transformation of the coefficients was performed before computing the mean, which was back transformed to a correlation (Alexander, 1990). Intra-group heterogeneity - the inverse of compacity - is measured as the Within-group sum of Squared Differences (WSD) between each elementary[10] unit's SDS and the average SDS of the cluster. Inter-group isolation is measured as the Between-group sum of Squared Differences (BSD) between each cluster's SDS and the average SDS of the clusters. SDS squared differences, calculated from partial PCA scores, have been chosen over other multivariate heterogeneity measures in order to control for multicollinear effects.

Nevertheless, as MAUP scale effects influence the local-hierarchical clustering process, particularly by generating a spurious increase of correlation coefficients towards higher levels of aggregation (Figure 3), the aforementioned measures must be adequately controlled. Accordingly, 120 spatial Monte Carlo datasets with empirical distributions have been elaborated, shuffling the selected variables (Table 2) among blocks, thus generating independent random spatial patterns while preserving the statistical distribution of each attribute. These datasets were processed by the regionalization algorithm, using two attribute distance measures: absolute ($absDS$) and squared ($sqrDS$) differences of partial PCA scores (see section 3.1). For each measure, 60 runs of the regionalization algorithm were performed, allowing calculating two adjusted indicators of the clustering process:

[9] Considering a total of 45 unique values for this case, excluding the diagonal (self-coefficients) of the correlation matrix. Absolute values are considered in order to avoid the annulation of positive and negative coefficients.
[10] Blocks, in this case.

First, an Adjusted Fischer Average of Correlation coefficients (AFAC):

$$\mathrm{AFAC} = \mathrm{EFAC} - \mathrm{RFAC}$$

Where EFAC is the Empirical Fischer Average of Correlations, obtained from a single run of the regionalization algorithm with real data, and RFAC is the Random Fischer Average of Correlations, calculated as the mean value of the 60 runs with shuffled data, for each set.

Second, an Adjusted Heterogeneity Ratio (AHR):

$$\mathrm{AHR} = \frac{\mathrm{EBSD} / \mathrm{RBSD}}{\mathrm{EWSD} / \mathrm{RWSD}}$$

Where EBSD and EWSD are respectively the between group and within group sums of squared differences of SDS, obtained with real data, and RBSD and RWSD are the corresponding indicators averaged from the 60 random tests.

Remarkably, considering only real data, the averaged coefficients regularly increase towards higher levels while the between-within ratios markedly decrease, but after controlling for random effects, both indicators reach maximum values at the same intermediate levels, for both dissimilarity definitions (Table 2). These indicators show that the optimal level of analysis for the selected variables in GS is roughly situated at a clustering level around 219 and 299 zones, depending on the dissimilarity measure. The regionalization algorithm used for this evaluation evolves at discrete scales, and a more precise definition could be obtained with single-step aggregation procedures. However, for a first approach these results will serve as a proof of principle for the proposed strategy to define an optimal scale of analysis.

The first question that must be solved is the choice between the absolute and squared distance algorithms, as the first produces higher values of AHR while the second performs best in AFAC (Table 3). As our main concern is to cope with MAUP effects, which are directly associated with correlation measures, it is suitable to decide upon adjusted correlations. Moreover, these measures reflect real spatial interactions among observations, and can be unequivocally interpreted in terms of the set of selected variables (Table 2). On the contrary, AHR is a ratio of ratios, which in turn stem from a series of calculations over PCA orthogonal transformations. Thus, AHR is a highly sensitive parameter that may be strongly affected by any of the involved factors, and should not be used to compare one model to another. Hence, we have calculated the area contained by linear interpolation among observations of empirical and random series[11], obtaining a value of 2.37 units of Fischer Averaged Correlation coefficients per logarithm of Units for the absolute distance algorithm versus 2.87 for the squared distances version (Figure 3), leading to choose the latter.

[11] With a fitted curve or a continuous hierarchical algorithm this could be calculated as an integral difference, but in this case no simple formula had a satisfying fit to the real data series.

Table 3: Clustering indicators, empirical and random-adjusted

Regionalization with absolute PCA partial scores differences

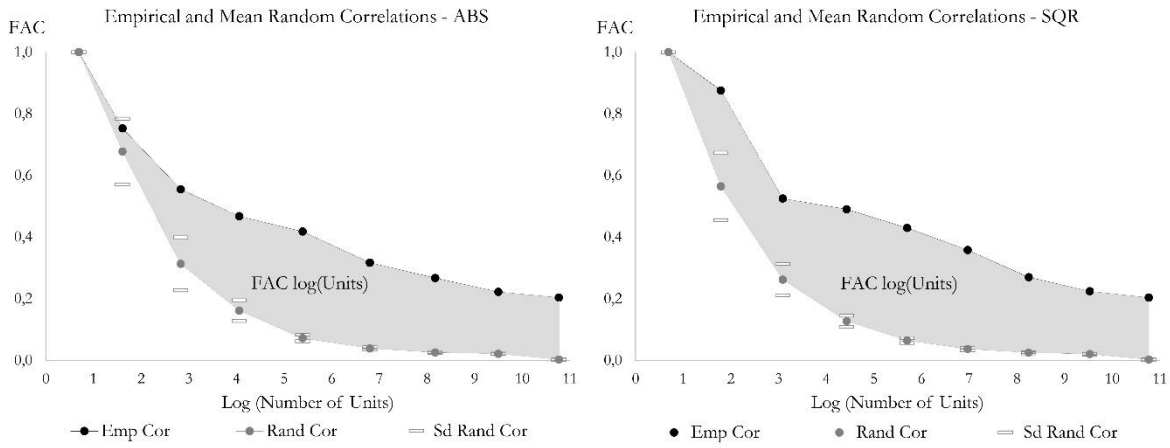
Scale	N° zones	EFAC	AFAC	EBSD	EWSD	EBW rate	AHR
1	47414	0.204	0.201	953	0		
2	13332	0.222	0.200	503	291	1.728	0.395
3	3540	0.267	0.241	288	505	0.571	0.504
4	899	0.317	0.278	131	635	0.206	0.553
5	219	0.418	**0.345**	68	712	0.095	**0.656**
6	58	0.468	0.306	26	773	0.034	0.470
7	17	0.555	0.241	13	827	0.016	0.325
8	5	0.753	0.075	8	874	0.009	0.216
9	1 (2)	1.000	0.000	0	953		

Regionalization with squared PCA partial scores differences

Scale	N° zones	EFAC	AFAC	EBSD	EWSD	EBW rate	AHR
1	47414	0.204	0.201	953	0		
2	13773	0.225	0.204	521	289	1.804	0.347
3	3827	0.270	0.245	238	493	0.482	0.370
4	1062	0.358	0.321	119	608	0.196	0.448
5	299	0.430	**0.365**	61	690	0.089	**0.463**
6	84	0.491	0.363	31	766	0.040	0.401
7	22	0.525	0.263	12	824	0.015	0.259
8	6	0.876	0.310	8	874	0.009	0.198
9	1 (2)	1.000	0.000	0	953		

Source: Authors' calculations. Notes: Maximum values of the optimality indicators are underlined and in bold case. For correlation averages (EFAC and CFAC), the reported values at scale 9 correspond to (2) zones, as displayed in the corresponding column.

Figure 3: Adjusted Fischer Averaged Correlations

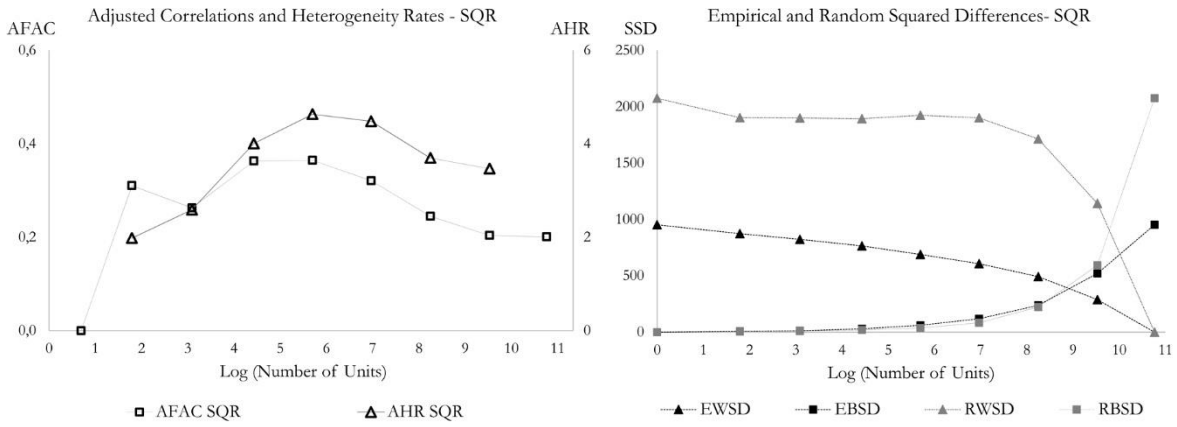


Source: authors. ABS and SQR respectively denote the results obtained with the regionalization algorithm using absolute and squared differences of attribute partial PCA scores. The FAC axis represents Fischer Averaged Correlation coefficients. Emp and Rand Cor respectively correspond to empirical values and the averages of 60 random data sets, Sd Rand Cor being the standard deviation of the latter. FAC log(Units) indicates the area contained by linear interpolation curves among observations of empirical and random series.

The second question is to determine the optimal scale of analysis and the corresponding number of clusters for the selected method. In the case of the squared attribute distances algorithm and considering AFAC as primary criterion, two very similar levels can be identified, level 5 with 299 units and an adjusted coefficient of 0.365 and level 6 with 84 and 0.363, respectively. However, AHR allows to clearly differentiating both levels, leading to select the fifth one (Table 3). At this stage, the high sensitivity of this double ratio is useful to differentiate among observations, while any systematic effects that may be produced by algorithm settings will similarly affect all the results of the same series.

Regarding correlations and heterogeneity ratios, major differences are observed between empirical and random datasets. In the case of Fischer averaged coefficients, there are important correlations of real data even at block level but they are absent in the random datasets, indicating that the selected variables are actually coproduced in GS' territory. These initial differences are first amplified by the regionalization process and then decrease, as the correlations converge to a theoretical maximum of 1, attained when only two units remain (Figure 3). Concerning the sums of squared differences of SDS, the structure of real data can be seen as a much lower between-units heterogeneity (BSD) at block level, compared with the random datasets, and a similar difference in internal heterogeneity (WSD) at the final stage of only one metropolitan aggregate (Figure 4).

Figure 4: Adjusted Correlations and Heterogeneity, SQR algorithm



Source: authors. All values correspond to regionalization with squared (SQR) differences of PCA partial scores. AFAC and AHR are Adjusted Fischer Averaged Correlations and Heterogeneity Ratios. WSD and BSD correspond to Within-group and Between-group sums of Squared Differences, differentiated for Empirical and Random datasets. As they represent total distances, the slope of BSD curves may be misleading, as they increase with higher numbers of units. However, when considering mean values, the distances between clusters actually increase at higher levels of aggregation.

The fact that the highest scores of both measures single out the same optimal level[12] - at least with the algorithm variants which have been tested here - highlights the close relationship between AFAC and AHR. In fact, as hierarchical regionalization algorithms simultaneously increase within-group homogeneity and between-group heterogeneity, an improvement in correlation consistency is expected, due to noise reduction inside the clusters and to a better differentiation among them (Mu & Wang, 2008:97). This opens a way to directly evaluate regionalization algorithms with real-world data, rather than with pre-designed or simulated spatial patterns. Furthermore, the results obtained so far support the argument to use AFAC and AHR both to rank different algorithms based on their performance with a specific set of data, and to single out the best level of analysis within the chosen model. Thus, hierarchical dendrograms should be cut at the level that maximizes AFAC while AHR should be used to differentiate among close ties. However, this conjecture is based on the comparison of two closely related algorithms and it should be thoroughly tested with a wider array of hierarchical regionalization methods, a task that will be performed in forthcoming research.

[12] Other usual but rather informal optimal level indicators were tested with the same data, such as within-group heterogeneity ratios between successive levels and diverse variants of the elbow criterion, which also singled out level 5. However, the discussion of these results would be excessively lengthy without adding relevant insights to this argument.

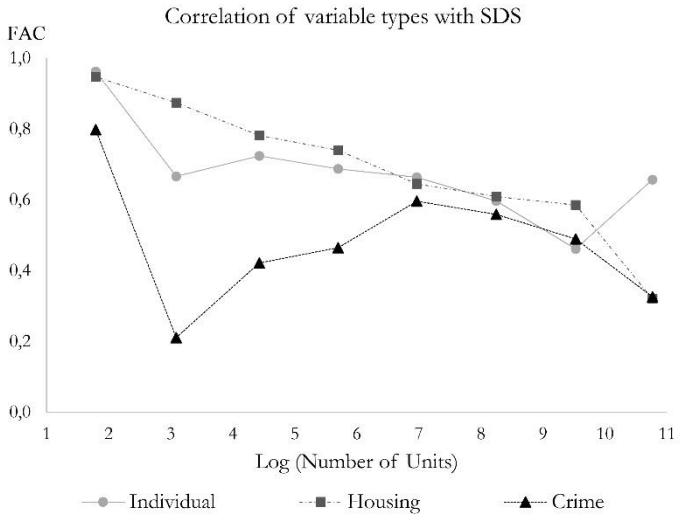
Considering the above, yet in order to ascertain if the social diagnostic at the level that has been singled out by the highest AFAC and AHR indicators sustains the inference of its optimality, it is worthwhile to develop the following cartographic analysis.

3.4. Socially critical zones in Greater Santiago at multiple scales

Greater Santiago is the main urban system of Chile, having an approximate population of 6 million inhabitants. It is a strongly segregated city, with high income disparities and severe urban inequalities, concerning health, education, transport, public spaces and service deficiencies in poor neighborhoods (De Mattos, 2002; Hidalgo, 2007; Sabatini & Brain, 2008). Thus, the variables which have been selected for this study offer a relevant but restricted perspective.

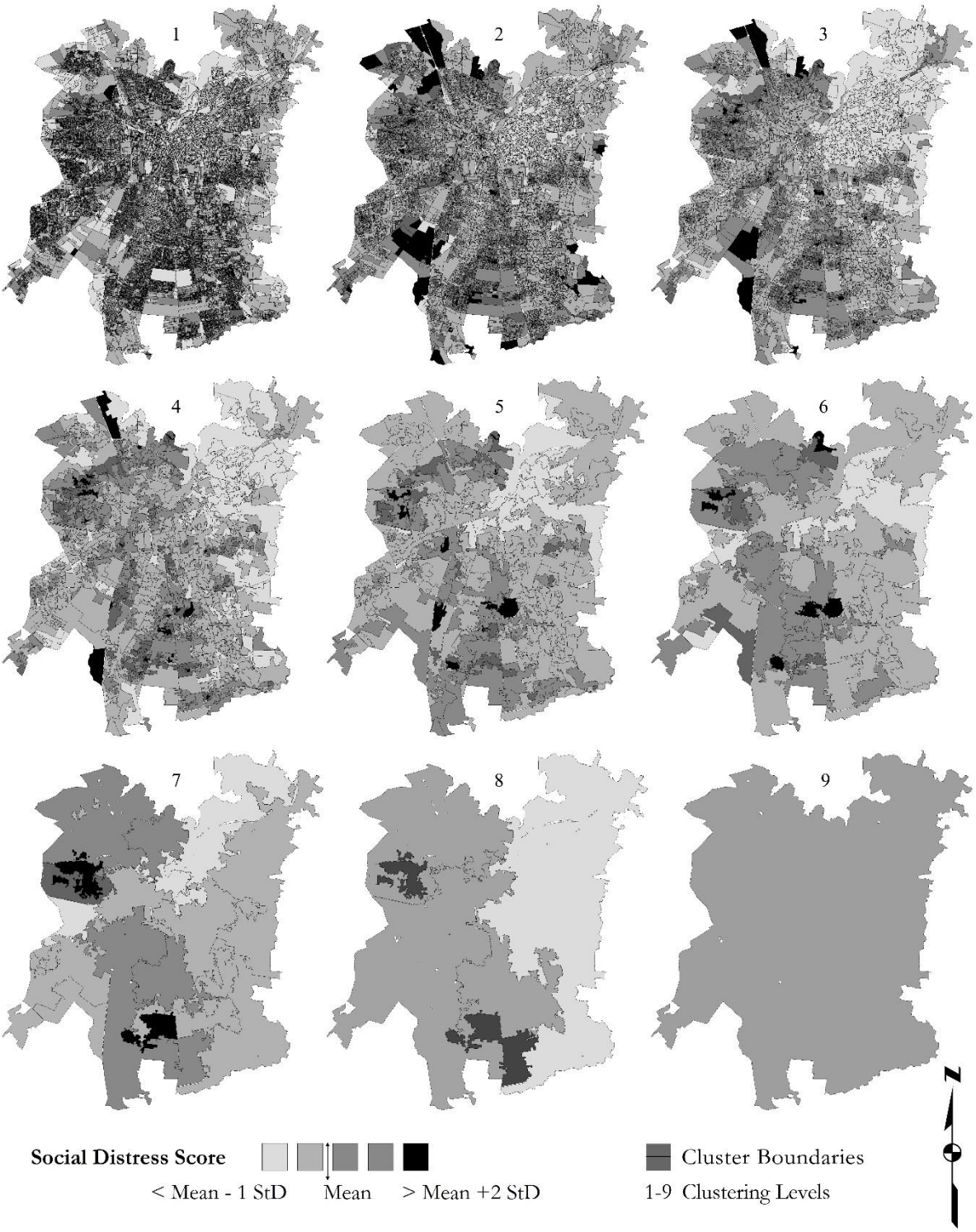
Remarkably, the relative contribution of the selected variables (Table 2) to the SDS shows important scalar variations (Figure 5). At small scales of aggregation, individual, housing and crime variables are almost equally correlated to the eigenvalue weighted PCA score. However, crime variables' contribution sharply decreases at higher scales, which is consistent with research that shows small-scale spatial correlations for this kind of data (Andresen & Linning, 2012). Overall, housing variables exert the strongest influence over SDS scores, reflecting a relevant spatial specialization of GS' housing market al all scales. These variations show the importance of recalculating PCA scores at several levels of aggregation, as multiple correlation patterns may change at different scales.

Figure 5: Scalar variations of social distress score composition



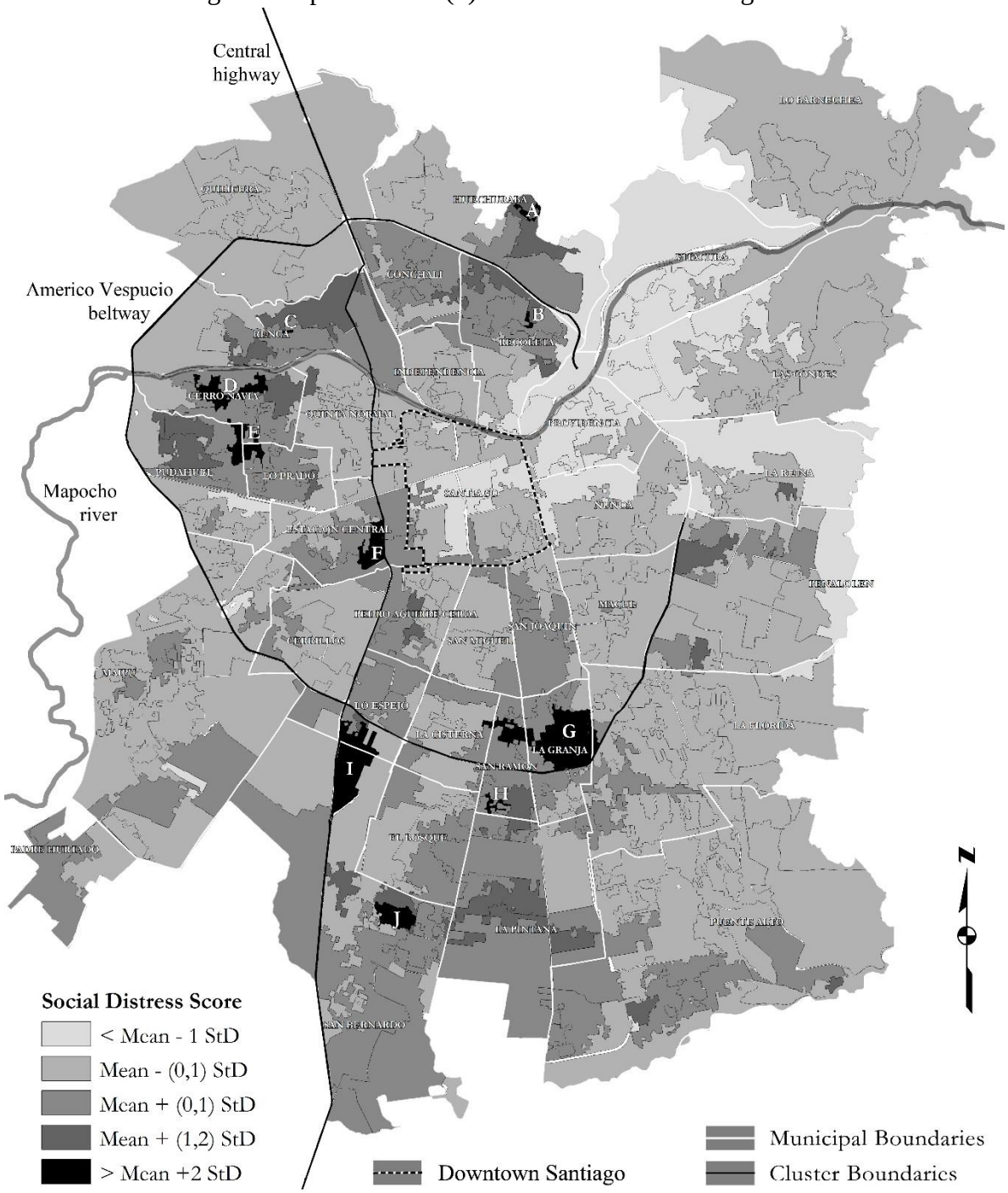
Source: authors. The FAC axis represents the Fischer Averaged Correlation coefficients, aggregated by variable type, between each of the selected variables (Table 2) and SDS at different scales.

Figure 6: Clustering levels of GS by social distress score



Source: authors' analysis with Chile's 2012 census data.

Figure 7: Optimal level (5) for social distress zoning in GS



Source: authors' analysis with Chile's 2012 census data.

The multiscalar cartography produced by our algorithm with the selected data is consistent with previous studies of GS's socio-spatial divides, housing inequalities and urban violence (De Mattos, 2002; Hidalgo, 2007; author et al, forthcoming). In general, the characteristic segregation pattern of GS is more or less conspicuous in levels one to eight (Figure 6). The high-income quadrant, from downtown towards the north-east, is particularly clear in the third scale, as multiple clusters of low SDS represented in light gray., Darker areas towards the northern, western and southern peripheries are visible from the second to the fourth level, corresponding to poor and excluded areas, severed by a clearer radial pattern of middle class housing, developed around highways and main public transport corridors. At intermediate levels, the darkest areas reveal the combination of discriminatory housing policies and multiple phenomena, such as poverty concentration and urban violence. Level eight clearly reveals the sharp socio-spatial divide that reflects the severe income and life quality inequalities between high-income groups, a majority of the Chilean population and cast-out territories.

The fifth scale of clustering, singled out as the optimal level of analysis (see section 4.1) is a rich source of information for the analysis of social distress in GS (Figures 6, 7). Critical zones are defined as those having a SDS above two standard deviations from the mean. From the 299 units at the fifth scale, thirteen clusters were thus selected, with a mean of 9,002 inhabitants, slightly under the mean population of census districts[13] in Chile. For the ten indicators used to build the SDS (Table 2), this subset has mean values which are significantly higher[14] than the other units' average, with insalubrity and precariousness rates which are over six times higher, while more than doubling overcrowded housing rates, high violence and insurgency densities. The detailed analysis of this data would be excessively long, but we will describe the most salient features of the critical units (Figure 6).

Sector 'A' is situated in the notorious settlement of 'La Pincoya', founded in 1969 from illegal land takeovers. This area presents the highest violent crime and the second drugs and insurgency densities, also having high rates of precarious and overcrowded housing. Zone 'B' roughly corresponds to 'Santa Ana' neighborhood, which has the highest overcrowding rate, also being the territory where several members of a band that executed the greatest robbery in Chilean history have been arrested. Cluster 'C' is a small area in 'Cerro Colorado' neighborhood, having the highest precariousness, insalubrity and dependence rates, and the second worst education and employment levels. Zone 'D' partially matches the 'Montijo-Resbalon' areas, located in the southern banks of the 'Mapocho' river, which has received rural immigrants since the late XIX century in formerly illegal settlements that have been gradually urbanized since Allende's government. This area shows the highest insurgency density, and very low levels of education and employment. Sector 'E' approximately contains the 'Pudahuel-Norteamérica' settlements, with a similar history to sector 'D', and presenting the highest unemployment ratio and interpersonal violence density. Zone 'F' corresponds to the 'Araucania-Nogales' settlements, closed at the west by

[13] This subdivision is immediately below municipalities, while containing census tracts and blocks.

[14] T test with over 99% certainty for all the variables.

the ‘Central’ highway. This area corresponds to the first regularization of a land takeover in GS, where 90 families were assigned small parcels in 1947, and presenting nowadays the highest drug offenses density, and very high insalubrity and overcrowding ratios. Zone ‘G’ contains the ‘San Gregorio-Malaquías Concha’ settlement, the first extensive social housing developments in GS, built since 1959 in order to accommodate the earliest massive eradications in Chile, in rather precarious conditions. A half century later, this area still presents deficient housing conditions, while developing high levels of urban violence. Cluster ‘H’ corresponds to ‘La Bandera’ settlement, founded as a massive illegal takeover in 1969 and formalized by Allende’s government in 1971. This neighborhood presents the lowest education levels, severe dependence, precariousness and overcrowding rates, and high crime densities. Sector ‘I’ is a mixture of ‘Nueva Espejo’ settlements with industrial zones, where the spatial proximity of low-skilled jobs contrasts with low education levels, high unemployment and dependency rates, and adverse housing conditions. Sector ‘J’ partially matches the ‘Olivo’ and ‘Portada’ settlements, founded in the sixties around the satellite town of ‘San Bernardo’, expanded afterwards in order to accommodate families eradicated by Pinochet’s dictatorship. This area shows rather high levels for all of the selected indicators, with the exception of insalubrity rates.

In sum, most of the highest SDS units correspond to well-known critical neighborhoods. A thorough discussion of their local identities and substantive characteristics is beyond the scope of this article, but the technical approach developed so far has been certainly useful to distinguish them in a metropolitan context. In these places, poor households have been concentrated by rural immigration, the first housing policies, forceful eradications during Pinochet’s dictatorship, or by more recent massive developments of social housing. Acknowledging the incompleteness of the selected indicators and having probably overlooked some relevant cases, this shows that critical social conditions are historically produced by urban policies and geo-economic trends, while being expressed as different and complex combinations of socio-spatial handicaps.

It should be noted that GS’ case presents several historic peculiarities, mostly related to deregulation of urban development through neoliberal policies implemented in Pinochet’s dictatorship, which have intensified socioeconomic segregation processes. Thus, it is unclear if the kind of analysis which has been performed here would lead to similar results in other contexts. For example, the contrast of urban inequalities between GS and Greater Paris, which have very different historical and regulatory conditions, has shown remarkable similarities and sharp differences between both cities (Garreton, 2013). However, the aggregation behavior of similar sets of variables should present related properties in different contexts, so diagnostics based on AFAC and AHR or similar indicators could help to accurately identify common and particular characteristics in international comparisons.

Finally, the results obtained so far demonstrate the usefulness of the proposed regionalization diagnostic strategy and its statistical robustness, suggesting new approaches to compare different contexts through differences on the scale and characteristics of their optimal analysis levels. To conclude, the main findings of this work and relevant lines for further research will be highlighted in the last section of this article.

3.5. Discussion

In this work, we have underscored the theoretical and empirical relationships between MAUP and regionalization approaches, thus developing a strategy to cope with scale effects which allows determining the best level of analysis. With this objective, an improvement of existing hierarchical regionalization algorithms (Mu & Wang, 2008) has been implemented, recalculating PCA scores - which are used to calculate dissimilarity among units - at several steps of aggregation, thus capturing scalar variations of multicollinearity. Particularly, at higher scales a marked decrease of the influence of crime variables on spatial interactions has been observed in GS, which is consistent with previous research (Andresen & Linning, 2012).

The main contribution of this research is to propose a strategy to determine the best hierarchical regionalization algorithm for a real dataset and then to select its optimal level of analysis (section 4.1). This is based on two adjusted indicators for the aggregation process, calculated with the results of one real and 60 spatial Monte Carlo generated datasets, allowing controlling for spurious MAUP effects. The best algorithm is considered to be the one producing a maximum aggregated AFAC, calculated as an integral difference between Fischer averaged correlations of real and shuffled data at every aggregation step, or by a suitable approximation. As a stopping rule to cut dendrograms, the optimal scale or number of clusters can be determined by the maximum AFAC as primary criterion, while close ties can be differentiated by AHR, which is a double ratio of between and within cluster heterogeneity of empirical and random datasets. Remarkably, both indicators single out the same levels for algorithms with two different dissimilarity definitions. These endogenous criteria for a stopping rule could contribute to focus hybrid regionalization methods (Guo & Wang, 2011), defining an optimal partitioning scale with results obtained at the preceding hierarchical structuration.

A statistical and cartographic analysis of GS' socially distressed areas at the optimal scale thus defined confirms the accuracy of this methodology, allowing identifying notorious neighborhoods with consistent identitary, historical and socioeconomic local handicaps. Some of these characteristics have been only briefly described, and the important question of what a cluster means in an urban setting has not been addressed yet. As recent research clearly shows, spatial clustering can provide rich frameworks to understand socio-spatial phenomena and to identify neighborhoods in more objective ways (Clark et al, 2015; Spielman & Logan, 2013). We believe that the proposed methodology opens interesting research perspectives on these subjects, clearly identifying aggregation scales that could lead to relevant substantive analysis of the places thus identified. For instance, the critical areas highlighted in this work can be useful for policy design and for further statistical and qualitative research.

It should be noted that this work has compared two closely related regionalization methods and further research is needed - involving different cases and a wider array of algorithms and dissimilarity measures - in order to confirm the general performance of the proposed

stopping rule. Nevertheless, the results obtained so far support the proposed strategy to identify an optimal scale of analysis, which has solid foundations on MAUP and clustering theory, thus contributing to the theoretical and empirical understanding of the spatial self-organization of interdependent real-world phenomena.

4. Case Study: Hierarchical Systems for Hedonic Appraisal[15]

Generating hedonic pricing models in a mass real estate appraisal context is a highly complex task, since its performance in terms of predictive power must be comparable with human valuation which is the prevailing worldwide practice (Lenk et al, 1997). These systems aim to replicate the decision making process of hedonic prices, which is a balance between supply and demand of a given good (Rosen, 1974). These decisions involve a large number of variables, and since these processes take place in the mind of the traders, we don't know its formal structure.

Despite this, there is a great development of models that allow for approximations of this structure to be made with good levels of determination (for a full review of methods reviewing McCluskey et al 2013, Zurada et al, 2011). It has been observed that there is a high interdependence among housing attributes (Basu & Thibodeau, 1998), generating non-linear behavior. Added to this, there is a high dynamism in the real estate markets, for example for the case of Santiago, Chile (Lozano, 2015; Parrado et al, 2009; Sagner, 2011). Because of this it is necessary to resort to intelligent methods that capture market behavior facing different scenarios. In trying to overcome these difficulties, the literature has seen a fertile development of work in recent years, in what has been called Computer-Assisted Mass Appraissal (CAMA) (McCluskey & Anand, 1999; Kilpatrick, 2011; Zurada et al, 2011).

The first econometric works on hedonic prices focused on understanding the determinants of housing prices, but the addition of more complex architectures has improved the predictive power of these models (see Tay & Ho, 1992). The development of hedonic pricing models has advanced greatly since Rosen (1974) who created the conceptual framework for estimating hedonic price based on linear regression analysis, which was proposed some years earlier by Ridker & Henning (1967). Hedonic pricing models, as proposed by Rosen, functionally relate the price of a property with its inherent attributes, such as floor area, land area or number of rooms; neighborhood attributes such as income or density; location or macro zone attributes; and attributes of regulation, as land use or construction in height permit. Since the availability of massive information and development of computing power, this topic has gained great interest in the community of expert systems, whom based on the design of intelligent systems have reduced the gap between human valuers and CAMA (McCluskey & Anand, 1999).

Several variables have been used to calibrate hedonic models, but since the development of GIS and Big Data, increasing levels of data have been available in the form of indices and metrics that have importantly contributed to price prediction (Geoghegan et al, 1997). Sirmans et al (2005) review the most used attributes present in the literature. They showed that most variables correspond to intrinsic attributes of houses. Since Ridker & Henning (1967), it is known that the value of a home will not just depend on these attributes, but

[15] This chapter was partially published, in collaboration with Dr. Marcelo Villena (Sanchez & Villena, 2016)

also on a number of variables that have to do with the neighborhood, and accessibility, which contribute significantly to the predictions (Zurada et al, 2011). Basu & Thibodeau (1998) found evidence of spatial autocorrelation in housing prices, putting further evidence regarding the importance of the geographical dimension of hedonic estimates. With the development of GIS, the inclusion of diverse spatial variables has proliferated. Some examples of spatial variables used in the literature are: the ecological landscape (Geoghegan et al, 1997; Brasington & Hite, 2005), access to green areas, (Bastian et al, 2002; Kong et al, 2007); air quality (Kim et al, 2003; Anselin & Le Gallo, 2006; Anselin & Lozano Gracia-2008); noise pollution (Cohen & Coughlin, 2008); access to education (Sedgley et al, 2008); etc. The availability of large databases is becoming increasingly common; hence the variables to consider are still under constant development, although all of them can be classified according to four categories of attributes defined by Rosen.

The functional relationship between price and input variables is unknown since the valuation is a process that occurs in the mind of the traders. This relationship will depend on the perception of any given agent and therefore will vary from one to another. The search for an approach to this function has taken several ways. Ridker & Henning (1967) approached the valuation function with a linear relationship, while Palmquist (1984) incorporated nonlinearities through log-linear models that significantly improved the predictive power of models. Later Hornik et al (1989) proved that neural networks are universal approximators of nonlinear functions; since then its use spread to the hedonic models, which have been used with remarkable results (Tay & Ho, 1992; Do & Grudnitski, 1992; Limsombunchai, 2004; Peteresnon & Flanagan, 2009). Given the subjective nature of the valuations, it is necessary to have systems that mimic the operation of the human mind as closely as possible.

Many studies label artificial neural networks as black boxes as they do not provide a method for directly analyzing the effect of the input variables in the final prediction, in contrast to the econometric models (Ge et al, 2003; Limsombunchai, 2004). Several studies compare the predictive power of both approaches. Some of them claim that hedonic regressions models have greater accuracy (Worzala et al., 1995), but the vast majority leans toward the superiority of neural network models (Selim, 2009; Tay & Ho, 1992; Do & Grudnitski, 1992; Zurada et al 2011). Their results show that the method of artificial neural networks has an advantage in its ability to map the nonlinear relationship between variables (Nghiep & Cripps, 2001). Given this, researchers have used several network configurations for mass assessment of properties that are developed in parallel with the appearance of relevant variables for the prediction (Tay & Ho, 1992; Peteresnon & Flanagan, 2009; Ge et al, 2003; Wilson et al, 2002). Other techniques from artificial intelligence have been proposed for modeling hedonic price (Landajo et al, 2012; Kauko, 2003), but the one-hidden-layer perceptrons are the most used in mass appraisal.

The focus on the predictive power of valuation systems has also led to the construction of intelligent models that combine techniques from various methodological currents (Tay & Ho, 1992). Systems that combine forecasts have been tested in various fields, and there is evidence that even the simple average of forecasts in an expert system improves the

prediction (Diebold, 1988, Diebold & Lopez, 1995; Genre et al, 2013; Aiolfi, 2010; Kristjanpoller et al, 2014). Goonatilake & Khebbal (1995) define the types of hybrid intelligent system where the combination forecasting models can be classified as inter communicated systems. Such systems have been applied in hedonic price forecasts, obtaining better results than individual models (Quigley, 1995; Kilpatrick, 2011; McCluskey & Anand, 1999). Another kind of system as defined Goonatilake & Khebbal consists of polymorphic systems, which adapt their structure according to the object under analysis.

Since Adam Smith that has been presumed that *Homo economicus*, is a rational agent who makes decisions based on complete information, so that maximizes their own benefit. This implies that all individuals in the population should behave the same way when the empirical evidence shows that in reality this does not happen. The economic agent actually has biases that make it behave irrationally when viewed from this perspective. The magnitude of this bias will influence the heuristics used by the agent in their decisions, which may differ between a submarket of the population to another. This is manifested in the form of the utility functions of each agent, which depend on the value that the agent assigns to each attribute. An interesting approach to this type of system is developed by Goodman & Thibodeau, 2004, where hierarchical clustering methods are applied to divide markets, significantly improving the prediction errors in the multiple regression models.

Despite the abundance of works in computer-assisted mass appraisal, the potential of implementing hierarchical structures to more sophisticated models as neural networks models has not been explored, even though the evidence suggests that this generates better predictions. The purpose of this work is to fill this gap in the literature by studying the impact of incorporating complex architectures to other predictive models, such as: econometrics models, artificial neural networks and hybrid models of combined forecasts.

4.1. Methods

4.1.1. The hedonic model

The origin of the hedonic price method goes back to the work of Hass (1922) and Court (1939), while its theoretical basis was then provided by Lancaster (1966) and Rosen (1974). From the work of Ridker & Henning (1967), hedonic pricing models have been widely used in understanding the determinants of prices and attribute valuation in real estate (comprehensive reviews can be found in the work of Follain & Jimenez, 1985; Sheppard, 1999; Malpezzi, 2003; and Sirmans, Macpherson, & Zietz, 2005).

The hedonic price method is derived from the theory of consumer behavior, where goods are valued based on the value added of its attributes, which vary among space. The price of a house P, can be expressed as

$$P = f(T, V, U, Z) + \varepsilon \qquad (1)$$

Where T represents the inherent characteristics of the house, V are the neighborhood attributes, U are macrozone attributes, Z regulatory attributes (Figueroa & Lever, 1992), and ε is a stochastic error term with zero expected value.

4.1.2. The data

Access to information is the main constraint to defining the variables to be used. In this study, mostly freely available public databases are used, plus some indexes developed by the Center for Territorial Intelligence of University Adolfo Ibáñez, as is detailed in Table 4. They are considered inherent variables such as surface and ground floor area; service accessibility variables, such as distance to schools, kindergartens, workplaces and subway stations; Socio-economic variables of each block, like density, average price per square meter, proportion of each income bracket, floating population; macrozone variables, in this case the municipality used, the unit that manages the regulatory plans; and environmental variables. Besides the housing variables, we included the year of the transaction, which will capture the temporal trend in prices.

The dataset was built based in 65,239 housing transactions in Greater Santiago (Chile) between the years 2010 and 2013. The descriptive statistics of the database can be seen in Table 5. There are some variables with normal distribution as floor area or travel times, and others that are Pareto distribution, such as price, age or density. This means that there are many cases in one end of the distribution and very few cases at the other end, generating difference of orders of magnitude between the 3rd quartile and maximum values in these variables.

A characteristic of complex systems is the interdependency, which implies that the variables are interrelated, which may violate the assumptions of some modeling approaches, especially linear ones. The relationship between variables can be seen in the correlation matrix presented in Figure 8. It can be inferred that there are variables with strong positive correlation, like between price and the proportion of high income, or between the distances between workplaces; and other variables with a strong negative correlation, like the floor area and distance work centers, or between the price of the property and the density of the block. Given the clear interdependence observed in the data, models that allow incorporating them into a functional form should have better predictive power.

Table 4: Definition of variables

Variable name	Description	Source
ANO	Transaction year	CBR
UF_TRANS	Transaction amount	CBR
SUP_CONSTR	Lot Size (m2)	SII
SUP_TERR	Floor area (m2)	SII
Con_patio	Courtyard (1 = yes, 0 = no)	SII
Edad	Age	SII
JARDIN_5M	Kindergarten 5 min walking (1 = Yes, 0 = No)	CIT
JARDIN_10M	Kindergarten to 10 min walking (1 = Yes, 0 = No)	CIT
JARDIN_15M	Kindergarten to 15 min walking (1 = Yes, 0 = No)	CIT
COLE_10MIN	College ranking top50 - 10 min walking (1 = Yes, 0 = No)	CIT
COLE_15MIN	College ranking top50 - 15 min walking (1 = Yes, 0 = No)	CIT
COLE_LEJOS	College ranking 50 to more than 20 minutes drive peak time (1 = yes)	CIT
SC_15MIN	Sub center 15 min (1 = yes, 0 = no)	CIT
SC_TPRIV	Private time to closest sub center	CIT
Ten_Metro	Metro is 18 minutes walk (1 = Yes, 0 = No)	CIT
Ten_MetroC	Metro is 10 minutes walk (1 = Yes, 0 = No)	CIT
TPR_CBD	Private time to work center CBD peak hour	CIT
TPR_ElGolf	Private time to work center El Golf peak hour	CIT
TPR_NvaLCo	Private time to work center Nueva Las Condes peak hour	CIT
TPR_Provi	Private time to work center Providence peak hour	CIT
ABC1_12P	Proportion of very high income in that block (%)	CENSO
C2_12P	Proportion of high income in that block (%)	CENSO
C3_12P	Proportion of median income in that block (%)	CENSO
D_12P	Proportion of low income in that block (%)	CENSO
E_12P	Proportion of very low income in that block (%)	CENSO
Casas_UFPr	Average price floor of that block	SII
DEN_MZ	Population density; pre-census population per hectare per acre.	CENSO
GSE12_NUM	Total inhabitants per acre	CENSO
POB_FLOT	Block floating population	CIT
HA_MZ	Block surface in hectares	CIT
IMORANCASA	Local spatial autocorrelation lot price	CIT
COM_CAS	Macrozone ranking of house prices	CIT
COM_CIT	Macrozone ranking of properties prices	CIT
COM_PR_CIT	$ / m2 average per commune	CIT
Z_UFM2_CAS	Floor price quartile	CIT
Zona_MA	Enviromental cluster based on surface temperature and greenery	CIT
ZonaMA_NEG	Zona_MA cluster with negative index	CIT
ZonaMA_POS	Zona_MA cluster with positive index	CIT
Por_veg_pr	Percentage of vegetation per acre.	CIT Landsat 8
Pro_predio	Average farm size per block	CIT
AV_15_MIN	Green area at 15 min walking (1 = Yes, 0 = No)	CIT
Amp_tst	Average amplitude of surface temperature between warm and cold month	CIT Landsat 8

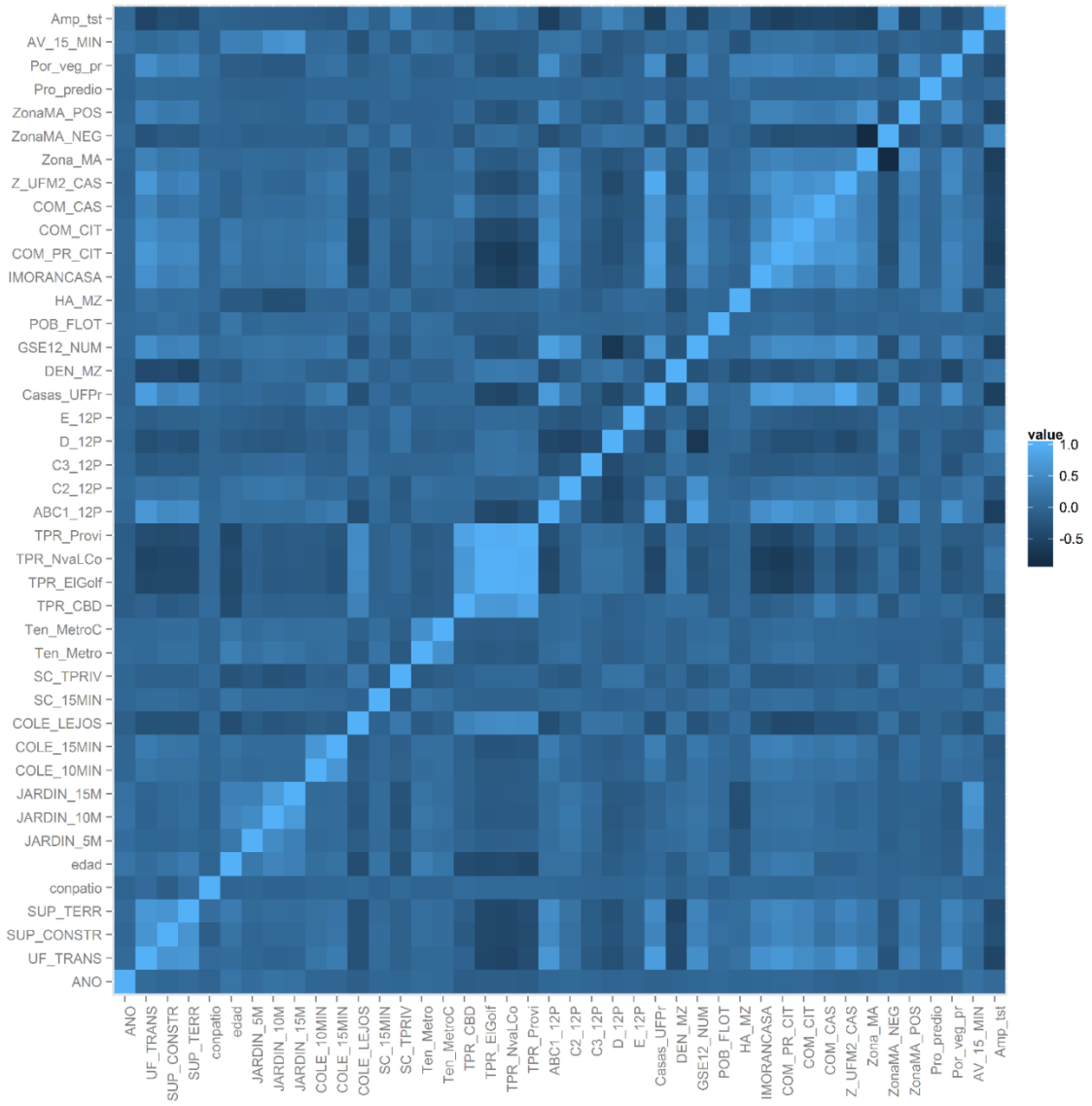
Source: authors

Table 5: Descriptive statistics

Variable name	Min	1st Quartile	Median	Mean	3rd Quartile	Max
ANO	2010	2011	2011	2011	2012	2013
UF_TRANS	357	885	1272	1838	2072	12000
SUP_CONSTR	36	51	69	78,97	98	189
SUP_TERR	62	112	149	175,2	207	600
Con_patio	0	1	1	0,9832	1	1
EDAD	0	4	14	20	29	324
JARDIN_5M	0	0	0	0,4257	1	1
JARDIN_10M	0	0	1	0,747	1	1
JARDIN_15M	0	1	1	0,8578	1	1
COLE_10MIN	0	0	0	0,01269	0	1
COLE_15MIN	0	0	0	0,02905	0	1
COLE_LEJOS	0	0	1	0,5163	1	1
SC_15MIN	0	0	0	0,02276	0	1
SC_TPRIV	0	616,7	923,3	1010,6	1256,2	4075,6
Ten_Metro	0	0	0	0,1671	0	1
Ten_MetroC	0	0	0	0,06084	0	1
TPR_CBD	0	1829	2403	2521	3234	5502
TPR_ElGolf	0	2382	3057	3064	4034	5464
TPR_NvaLCo	0	2353	3135	3113	4119	5696
TPR_Provi	0	2137	2822	2887	3721	5451
ABC1_12P	0	0	0,01695	0,12593	0,14815	1
C2_12P	0	0	0,1717	0,2159	0,3361	1
C3_12P	0	0,06667	0,24419	0,27712	0,375	1
D_12P	0	0,03571	0,24	0,3194	0,48649	1
E_12P	0	0	0	0,06167	0,05714	1
Casas_UFPr	0	13,32	16,9	18,68	20,95	66,83
DEN_MZ	0	91,01	162,77	171,69	229,45	1305,97
GSE12_NUM	0	1	2	2,093	3	4
POB_FLOT	0	0	0	50,4	0	11469,9
HA_MZ	0	0,4473	0,7605	5,7845	2,3135	407,9732
IMORANCASA	-0,401704	0,008412	0,075754	0,379027	0,295252	10,213109
COM_PR_CIT	0	15,37	18,45	20,8	21,35	50,17
COM_CIT	0	15	23	22,01	28	41
COM_CAS	0	19	28	24,69	31	41
Z_UFM2_CAS	0	1	2	2,061	2	4
Zona_MA	-1	-1	0	-0,3996	0	1
ZonaMA_NEG	0	0	0	0,4728	1	1
ZonaMA_POS	0	0	0	0,07321	0	1
Pro_predio	0	120,1	168,6	480,7	287,1	58141,7
Por_veg_pr	0	0	0	12,06	15,17	100
AV_15_MIN	0	1	1	0,7878	1	1
Amp_tst	0	20,58	22,18	21,68	23,26	31,17

Source: authors

Figure 8: Correlation matrix



Source: authors

4.1.3. Multiple Regression Analysis

Hedonic regression models are the most widespread in the housing appraisal literature. Taking advantage of the rich database of housing prices and their attributes, we built 11 models of regression to represent equation (1). The first model is the most basic and considers only the surface of the house and its land area.

For each additional model we added variables corresponding to an attribute type that can potentially affect the price of housing as seen in Table 6. Thus, to the basic model we add the dummy variable courtyard, then the age of the property and transaction. In model 4 we add the socioeconomic group, gse12_num. Later, location variables, including time and distance to Providencia (Business district) and to the nearest metro station was added.

Finally, variables that account for vegetation, distance to schools, the ranking of location-based rates, density and surface of blocks, and the surface temperature are added. The tenth model considers all available attributes and all the statistically significant variables found.

This approach resembles the one of Hendry (1995), going from the general to the specific. The results of all econometric models are presented in Table 6. In particular, we see that all the variables considered are statistically significant at 1%. Similarly, it is clear that each set of attributes improved the adjusted R^2, in fact between the first base model and the tenth, the overall predictive power of the regression is improved by more than 18pp.

The purpose of these preliminary regressions reported in Table 6 is to define the specifications to be used in the various models to be tested in this study. These were calculated on the entire database, without separate training and test set, and without segmentation.

As a control method an eleventh model that considers all the variables of the database is generated.

Table 6: Preliminary regressions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	uf_trans	uf_trans	uf_trans	uf_trans	uf_trans	uf_trans	uf_trans	uf_trans	uf_trans	uf_trans
sup_terr	7.401***	7.099***	7.837***	6.438***	6.139***	5.810***	5.819***	5.554***	5.318***	4.975***
	(0.102)	(0.104)	(0.108)	(0.0955)	(0.0921)	(0.0905)	(0.0899)	(0.0871)	(0.0928)	(0.0925)
sup_constr	17.12***	18.02***	17.79***	14.94***	12.70***	12.08***	11.78***	12.02***	11.86***	11.10***
	(0.240)	(0.250)	(0.245)	(0.216)	(0.213)	(0.206)	(0.204)	(0.195)	(0.195)	(0.195)
conpatio		742.4***	631.4***	470.0***	458.0***	392.5***	376.5***	298.9***	293.3***	296.5***
		(29.11)	(29.85)	(27.05)	(25.50)	(24.66)	(24.63)	(24.74)	(24.49)	(23.83)
edad			-10.98***	-10.80***	-14.82***	-11.46***	-11.82***	-9.938***	-8.934***	-10.09***
			(0.293)	(0.269)	(0.289)	(0.291)	(0.292)	(0.288)	(0.295)	(0.297)
year			72.22***	76.57***	60.35***	77.72***	84.10***	88.68***	84.21***	91.98***
			(4.659)	(4.292)	(4.194)	(4.113)	(4.072)	(3.958)	(3.951)	(3.923)
gse12_num				463.0***	471.4***	415.2***	388.0***	284.4***	275.1***	230.6***
				(5.093)	(4.964)	(4.796)	(4.815)	(4.633)	(4.576)	(4.583)
tpr_provi					-0.285***	-0.284***	-0.252***	-0.288***	-0.282***	-0.338***
					(0.00488)	(0.00487)	(0.00489)	(0.00482)	(0.00482)	(0.00536)
ten_metro					-182.2***	-110.7***	-137.6***	-168.0***	-159.8***	-149.5***
					(12.97)	(12.63)	(12.53)	(12.20)	(12.17)	(12.01)
zona_ma						683.7***	680.7***	570.4***	584.5***	306.4***
						(27.50)	(27.18)	(26.62)	(26.77)	(27.83)
zonama_neg						454.4***	471.0***	426.6***	459.1***	295.3***
						(30.19)	(29.82)	(29.01)	(29.22)	(29.37)
av_15_min						-369.8***	-362.3***	-412.3***	-341.0***	-341.0***
						(9.820)	(9.688)	(9.445)	(9.382)	(9.414)
cole_15min							1,024***	810.2***	820.3***	708.0***
							(44.53)	(44.61)	(44.85)	(44.92)
com_cas								29.19***	30.90***	26.82***
								(0.424)	(0.434)	(0.422)
ha_mz									6.078***	6.804***
									(0.421)	(0.483)
den_mz									-0.113**	-0.147***
									(0.0450)	(0.0458)
amp_tst										-100.4***
										(3.094)
Constant	-810.2***	-1,558***	-146,598***	-155,695***	-121,913***	-156,317***	-169,177***	-178,684***	-169,774***	-182,767***
	(15.61)	(34.77)	(9,373)	(8,636)	(8,438)	(8,274)	(8,192)	(7,964)	(7,950)	(7,883)
Observations	63,054	63,054	63,054	63,054	63,054	63,054	63,054	63,054	63,054	63,054
R-squared	0.494	0.497	0.512	0.586	0.607	0.630	0.640	0.664	0.668	0.678

Robust standard errors in parentheses
***p<0.01, **p<0.05, *p<0.1

Source: authors

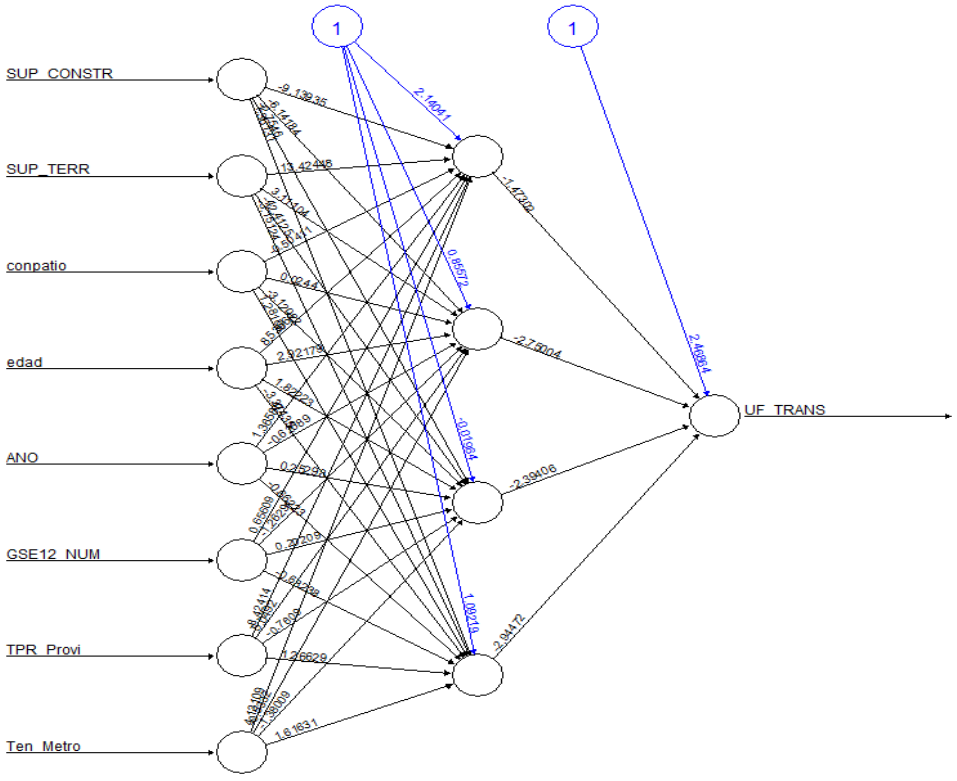
4.1.4. Artificial Neural Networks

Artificial neural networks are models inspired in the performance of biological neurons, but ultimately they are mathematical models that can provide a functional relationship between input variables and an output variable, as shown in Figure 9. Hornik (1989) shows that neural networks with one hidden layer with sigmoidal transformation function are universal approximators of any non-linear function, so its use in hedonic price modeling has been extended since. Multilayer perceptrons with one hidden layer have a function for equation (1) of the form

$$P_i = \phi\left(w_o + \sum_{j\in J} w_j \cdot \phi\left(w_{oj} + \sum_{k\in K} w_{kj} \cdot X_{ki}\right)\right) \qquad (2)$$

Where ϕ corresponds to the transformation function, in this case sigmoid, J is the number of neurons in the hidden layer, and K is the number of variables in the input layer. The optimal number of neurons in the hidden layer is an open issue, but Blum (1992) suggests that this number must be between the size of the input and output layers, so the midpoint is taken, i.e. $(K+1)/2$. With regard to the input variables 11 sets will be tested, corresponding to the specifications of the 11 different linear regression models. To train the neural networks the input variables must be normalized between -1 and 1, and outputs between 0 and 1, but this linear transformation does not affect the variables probability distribution.

Figure 9: Artificial Neural Network Architecture



Source: authors

4.1.5. Combining forecasts

The combination of forecasts is widespread in housing appraisal systems (Quigley, 1995; McCluskey & Anand, 1999), where expert systems combine the results of different base models to make a final prediction. Three methods of combining forecasts are tested: Simple average, regression weighted average and neural weighted average. To remove noise from poorly performing models, forecast combinations only consider the 6 models with lower training-sample MAPE (mean absolute percentage error).

4.1.6. Hierarchical architectures

The spatial disaggregation of markets can significantly improve the predictive power of the models of hedonic regression (Goodman & Thibodeau, 2003). Using different algorithms allowed them to model the boundaries between the various markets rather than arbitrarily defining them. These boundaries can be physical or abstract depending on the approach (regionalization or clustering). The purpose of segmenting markets not only has to do with improving the accuracy of the models, but with improving interpretability, so a regionalization algorithm was used to create contiguous spatial units.

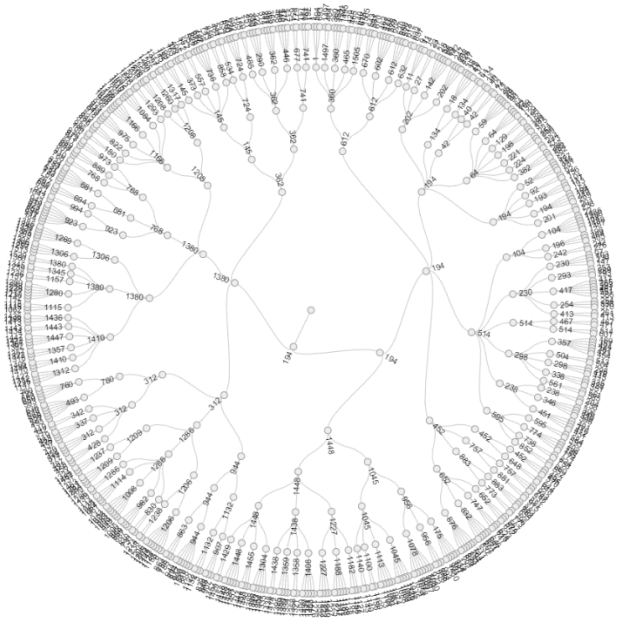
Assembling spatial units into meaningful groups is a difficult task because they must deal with high computational complexity while controlling the modifiable area unit problem (MAUP), spatial autocorrelation and multicollinearity attributes. Garretón & Sánchez (2016) shows that the magnitude of these problems varies among the scale at which it is analyzed, so their proposed method is used to incorporate emerging effects of interaction between variables. This creates a hierarchical structure of spatially contiguous regions, based on 55.000 blocks of the Great Santiago, as seen in Figure 10.

As Goodman & Thibodeau, a sub market is defined if it has at least 200 housing transactions. Starting at block level, if that block has less than 200 transactions, then the submarket is defined in the upper levels, going up the hierarchical tree until an upper region meets the restriction of 200 transactions. The sub model associated with this submarket is used for valuations of properties of that block. Those properties that were part of more than one sub-market, are valued with the most specific submarket available. A total of 322 submarkets were generated as can be seen in Figure 11, and therefore the hedonic price model for the submarket i is

$$P_i = f(T_i, V_i, U_i)$$

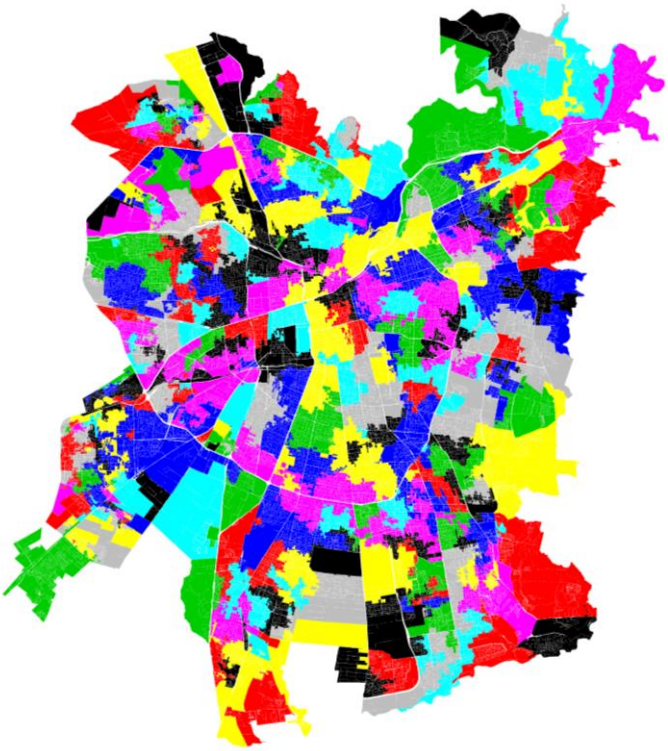
To make a full comparison, the analysis of Goodman & Thibodeau will be extended by considering other forms of the function $f(T_i, V_i, U_i)$, besides the linear form. Models that have proven to have better predictive power, such as neural networks and combined forecast systems will also be tested. Finally, hybrid models and individual models with and without market segmentation will be compared to see the impact of segmentation in different predictive performance.

Figure 10: Hierarchical dendrogram



Source: authors. The lower lever correspond to each block in Great Santiago

Figure 11: Submarket map of Great Santiago.



Source: authors

4.1.7. Sensitivity

Estimating of the relationship between a dependent variable and its predictors is a central problem in empirical research. The statistical methods such as correlation analysis and linear regression analysis are widely used in various disciplines in order to answer this question, as in medicine, social sciences, earth sciences, etc. Other disciplines, such as engineering in all its forms, have generated simulation methods to evaluate more complex systems, which unlike statistical models, not necessarily consist of differentiable functions. Having methods to address this problem can answer many questions in the most diverse fields, so the development of these tools is an issue that remains an active area of research (Xingli & Olden 2015, Fischer 2015).

Using partial derivatives of the prediction with respect to input variables has been used for both linear and neuronal models to understand their elasticity (Deif, 2012; Davis, 1989). The first attempt on neural networks was conducted by Davis 1989, who using a bootstrap approach analyzes the behavior of derivatives under disturbances of the input variables. Another interesting work is developed by Dimopoulos (Dimopoulos et al 1995, Dimopoulos et al 1999), who generates various sensitivity metric from the partial derivative function obtained for each variable, allowing to analyze the model directly after the training the neural network. Other works where partial derivatives are applied are Intrator & Intrator (2001), Reyjol et al., 2001, and recently (Gevrey & Dimopoulos 2015) who estimates the cross derivatives between two predictors to know their interdependence.

There have been other approaches to understand the sensitivity of the networks, in addition to partial derivatives. Olden et al 2004 make a thorough review, reporting methods that heuristically analyze the sensitivity of the networks (Lek et al., 1996; Scardi., 1996; Recknagel et al, 1997), or visually, with interpretive diagrams (Özesmi & Özesmi, 1999), but most of these studies are limited to developing a ranking of importance of the variables, without giving further information on the sign of the interaction or statistical significance of sensitivity metrics.

4.1.8. Multilayer perceptron partial derivative

Several authors even refers to training artificial neural networks as regressions (Intrator & Intrator 2001, Specht). Artificial neural networks, like regression models, functionally link a dependent variable with independent variables. This is often a differentiable function, especially considering the network with one hidden layer and sigmoid activation function, which is the most widely used network architecture in several fields (Chen & Ware, 1999; Nghiep & Al, 2001), and has the property of being a universal approximator of nonlinear functions (Hornik et al, 1989). Due to the functional nature of artificial neural networks, all statistics techniques are implementable in order to understand its sensitivity (Cheng & Titterington, 1994).

The multilayer perceptron which is used has one hidden layer, and sigmoid activation function, as shown in Figure 9. This architecture has the output function

$$f(X_i) = \phi\left(w_o + \sum_{j\in J} w_j \cdot \phi\left(w_{oj} + \sum_{k\in K} w_{kj} \cdot X_{ik}\right)\right)$$

Where ϕ corresponds to the transformation function, in this case sigmoid, J is the number of neurons in the hidden layer, and K is the number of variables in the input layer. The optimal number of neurons in the hidden layer is an open issue, but Blum (1992) suggests that this number must be between the size of the input and output layers, so the midpoint is taken, i.e. $(K+1)/2$ (Sánchez & Villena, 2015). The gradient vector of f with respect to X_k is $d_i = [d_{i1}, \dots, d_{iK}]^T$(Dimopuoulos et al 1995), and

$$d_{ik} = s_i \sum_{j=1}^{J} w_{0j} I_{ji} (1 - I_{ji}) w_{kj}$$

Where s_i is the derivative of the output layer with respect to its input, I_{ji} is the output of hidden node j for the input X_i, and scalars w_{kj} y w_j are the weights between the input k and hidden neuron j, and between hidden neuron j and the output.

Instead of directly consider the partial derivative, elasticity is calculated, which has the same units for all variables, generating indicators that are directly comparable across variables. Elasticity is calculated as

$$E_{X_{ki}} = \frac{X_{ki}}{\bar{f}(X_i)} d_{ik}$$

Since the objective is to find a method comparable to other models, the expected value of the elasticity of each variable will be obtained, and its respective standard deviation. The SSD_k index (Dimopoulos et al 1999) is additionally generated, which is calculated as the sum of the square of the derivative of the variable k.

4.2. Results

To generate the results in each submarket, 80% of the data available was taken for training and 20% for testing. This data was used to train the 25 models described earlier in each submarket. 50 partitions were made randomly, and the testing-set MAPE was measured in each simulation as a directly interpretable measure of error that is comparable between all methods.

4.2.1. Predictive power

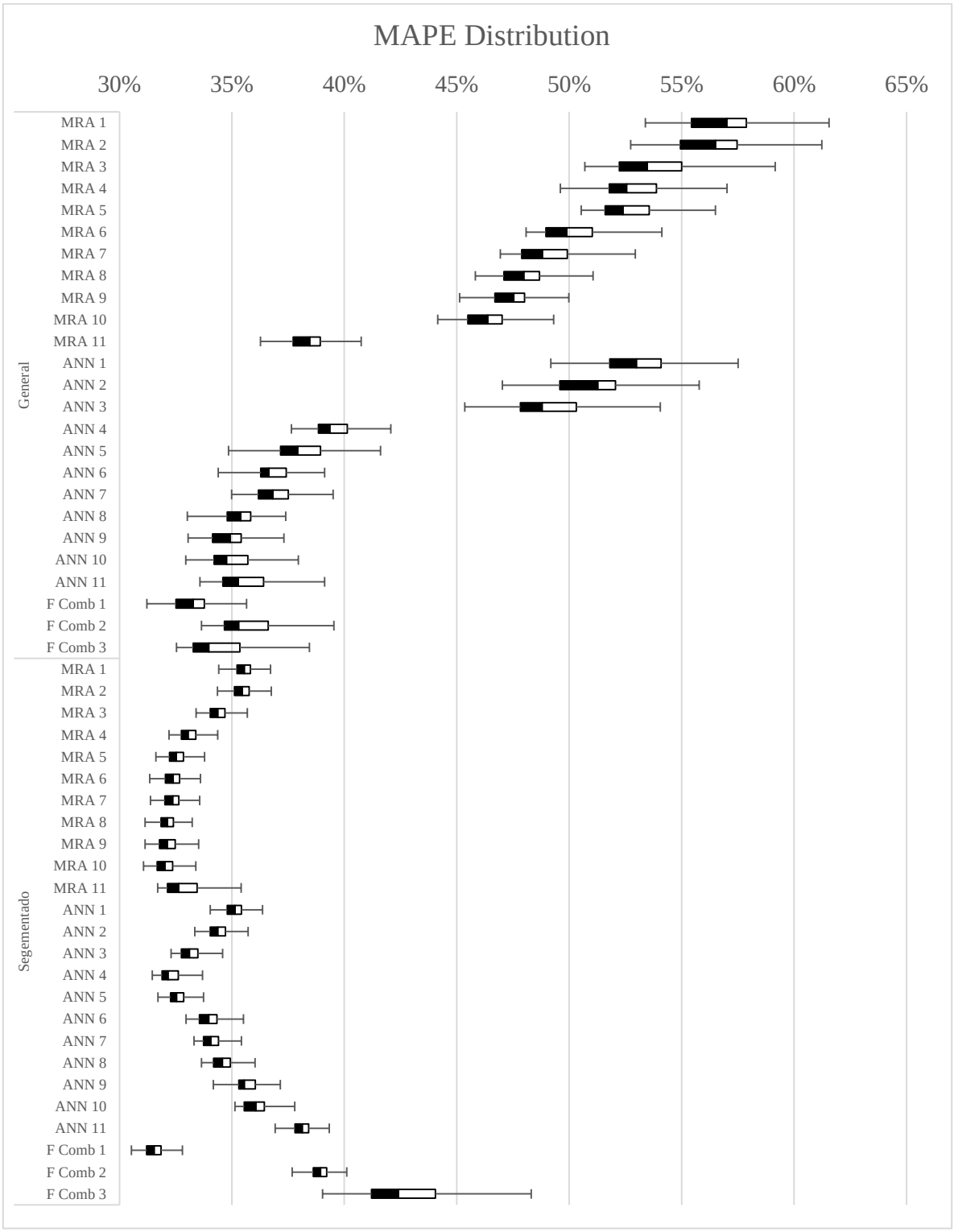
The distribution of the errors between samples can be seen in Figure 12. As can be seen, segmenting markets improves the predictive power of all models, whether regression models, neural networks or combination of forecasts, although the greatest reduction in MAPE occurs in the regression models. This is because the neural network models and combined forecasts already incorporate nonlinear effects in their training, so the added value of market segmentation is lower compared to simpler models. It can also be seen that demand segmentation helps reducing the variability of predictions, moving in a much more limited range than the general models.

The general models validate the performance of neural networks, these being superior to all regression models, as several authors have suggested, and the same with systems combined forecast, exceeding the performance of other basic models. Interestingly, this behavior is reversed for segmented models, where in many cases regression models outperform neural networks and hybrid models. This is related to the phenomenon mentioned above, the

segmentation itself incorporates effects of spatial autocorrelation and multicollinearity attributes, so the simplest models improve performance without modifying the valuation function. It also happens because the market segmentation reduces the variance of territorial variables, so models with large number of these variables are exposed to overtraining given the high number of parameters and the low number of significant variables. This can be corroborated seeing that simple neural networks and hybrid models, have less error than more complex models. This result is very sensitive to the architecture of the system, so to validate this fact is necessary to sensitize this method to different submarkets thresholds and for different neural networks configurations.

Another interesting result is that the hybrid model which makes the simple average of the other predictions is the best performer, in both scenarios, which is previously pointed out in the literature (Genre et al, 2013), indicating that not always the more complex model will make the best forecast.

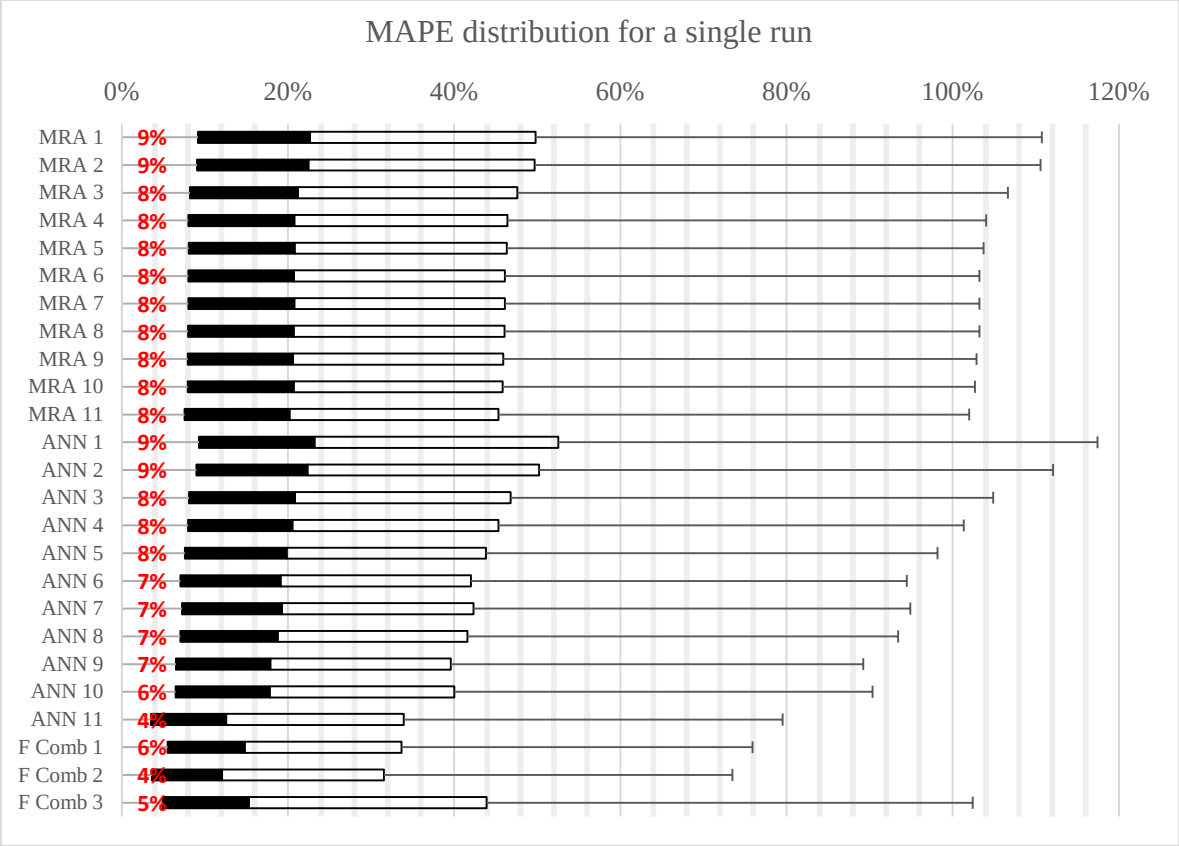
Figure 12: Distribution of MAPE over all runs



Source: authors

These results are robust, but they aggregate all the valuations performed by a model in each one of the 50 iterations, and this hides information about how the MAPE is distributed within each simulation. Added to this, the high level of error, almost always above 30%, requires a thinner analysis in order to build a reliable system of mass appraisal, and therefore solving the original problem. The errors obtained by the best of the 50 simulations are shown in Figure 13, and can be seen that when opening the MAPE by transaction, the dispersion increases dramatically. All models have at least 25% of valuations with MAPE under 9%, improving this indicator as the model increases in complexity. However, they have another 25% of valuations by over 40%, with differences depending on the area where the valuations are done.

Figure 13: Distribution of MAPE in a single run



Source: authors. The red percentage represents the 1st quartile of MAPE in this single run, meaning that a 25% of the observations have that MAPE or less

Anyway a high percentage of assessments that falls within acceptable ranges. By selecting model with the lowest MAPE in each submarket, we can see that all models have some share, and this proportion of cases where a model is the top performer can be seen in Table 7. In no case the best performance exceeds a MAPE of 9%, and the linear models that are best performers have the lowest errors. Note that the overall MAPE of the system, with each sub-market appraised with the best performance model, is of 5%, which is

significantly lower than the results shown in Figure 10, which consider only a single model for all transactions.

Table 7: Best performer distribution

Model	MAPE	Relative Freq
MRA 1	6%	1,71%
MRA 2	5%	1,41%
MRA 3	4%	2,13%
MRA 4	4%	1,57%
MRA 5	3%	1,50%
MRA 6	2%	1,38%
MRA 7	2%	0,16%
MRA 8	3%	0,49%
MRA 9	3%	1,45%
MRA 10	4%	1,81%
MRA 11	6%	4,28%
ANN 1	5%	2,54%
ANN 2	5%	2,96%
ANN 3	5%	3,20%
ANN 4	5%	3,31%
ANN 5	6%	3,70%
ANN 6	7%	4,22%
ANN 7	6%	4,23%
ANN 8	6%	4,72%
ANN 9	7%	5,30%
ANN 10	7%	5,94%
ANN 11	2%	13,76%
F Comb 1	6%	9,80%
F Comb 2	9%	14,79%
F Comb 3	1%	3,63%
Total general	5%	100,00%

Source: authors

If we see the spatial distribution of errors in Figure 14, it is seen that there are some clusters with high predictive power, and some areas in red, where all the efforts of improving the predictive power of the system should be focused.

Figure 14: Spatial distribution of the best MAPE

Source: authors

4.2.2. Sensitivity

With respect to sensitivity, local partial derivative for each best model was calculated with respect to its input variables. The SSD indicator is then measured, and the variable that means the greatest impact on the prediction of the model is recorded. The results of the sensitivity can be seen in Table 8, which says that the variables representing density of the block, surface of the block, and surface area of the property have the highest elasticity. These variables are interrelated, as they represent the quality of life that can be achieved in that location. The following variables have to do with green areas (the surface temperature is a proxy of it), and then the municipality where the property is located. It is noteworthy that large part of the variables appears as the most relevant, depending on the geographical location, which indicates that the spatial disaggregation of the analysis gives very relevant information for decision-making around the property market.

Table 8: Sensitivity distribution

Variable	Relative Freq
DEN_MZ	10,20%
HA_MZ	9,10%
SUP_TERR	8,30%
Amp_tst	8,20%
COM_CAS	8,10%
conpatio	7,00%
AV_15_MIN	6,40%
COLE_15MIN	6,10%
ANO	5,80%
EDAD	5,70%
TPR_Provi	5,30%
ZonaMA_NEG	5,20%
GSE12_NUM	5,20%
Ten_Metro	5,00%
Zona_MA	4,20%

Source: authors

4.3. Discussion

This study extends the knowledge about the application of hedonic models in segmented markets. We show that housing forecasts improves for all models when done at a submarket level. This might occur due to what authors as Kahneman & Tversky (1978) called bounded rationality, which implies that the decision-making differs between subjects from different contexts. In this context, sub-markets models should be better to capture the behavior of individuals, which holds for all models.

Results confirm that neural network models exceed the predictive capability of the regression models when applied in a general framework; however, this situation changes when hierarchical clustering methods are used, allowing simpler models as regression analysis to outperform them. These results have several implications, since it allows to build high accuracy predictive systems that can be analyzed like any other econometric model. This happens given the black box nature of artificial neural networks and hybrid models, which do not provide information on the impact of each variable in the prediction.

This finding gives a new edge to the debate on the use of regression or neural networks for house appraisal. The major argument in favor of neural networks is their power of generalization and its ability to incorporate nonlinear behavior, but in this case study we show that when market segmentation incorporates these phenomena then the added value of neural networks is much lower.

What can be inferred from this is that there is not a right model, all of them have limitations and assumptions, and the choice of the right method will depend on the context in which it is sought to be implemented. Given these limitations, the idea of combining techniques from different methodological currents to create systems that have better predictive power has gained ground. It is noteworthy that the simple average is the one with the best performance, both in the segmented model, as in the general model, as Genre et al (2013) anticipated. This is explained as the different models that are combined are determined independently, so it is assumed that there is no relation between residues of the different models, and there should not be a functional relationship between their forecasts.

This work shows strong evidence of the impact of market segmentation on different predictive systems, but this was tested with a limited number of models, and therefore results may vary if other methods are considered. Another factor to consider is that the segmentation method used considers nonlinear phenomena and spatial autocorrelation, and thus helps reduce the complexity of the problem, but this is not true for all segmentation methods.

The results confirm that the partial derivative method provides reliable information, allowing doing more complex analysis than other heuristics that only generate variable rankings without providing information on local elasticity (Olden et al 2004). This has several implications since having the partial derivative function allows gaining more insights into the system' behavior by analyzing for example, the elasticity in different ranges of the predictor variable

It is noteworthy that artificial neural networks, despite having better predictive power, are very sensitive to its architecture, and therefore the results may differ under different network configurations, which is a challenge to be addressed in future research. In addition there are other scenarios where regressions outperform artificial neural networks (Sanchez & Villena, 2015), so these different models are complementary tools that should support the decision making processes altogether.

This work opens the way for the development of hybrid systems of computer-assisted mass appraisal, because it combines approaches that had been developing separately, market segmentation, and building systems of forecasts combination. Care must be taken as it was shown that increasing the structural complexity of the systems will not necessarily improve the predictive power, so it is key to carefully design the architecture of intelligent systems. Creating segmented models allows us to understand the dynamic of local market, which allows industry players to design more specific strategies, and therefore is a tool that adds real value to current solutions.

5. Conclusion

Diversity is an inherent part of nature, so to study such a complex phenomenon as the Self-organization of space, one must begin by recognizing this diversity and deal with it within the proposed analysis methods, and must be tackled with tools with multidisciplinary approach.

In this work, it has been presented a spatial clustering algorithm which is able to identify relevant geographic patterns without external intervention, other than selecting the initial dataset. The fundamental novelty of this method is its capacity to exploit spatial interactions as useful sources of information about self-organizing topological phenomena, rather than considering them as spurious effects. In particular, the reevaluation of multicollinearity through recalculation of PCA scores at different scales allows capturing the similarities induced by spatial correlations at attribute-specific extents.

Nevertheless, two key improvements can be conceived for this algorithm. Firstly, develop a computationally efficient strategy for smoothing the aggregation process, making it more similar to pairwise matching schemas such as Ward's (1963) method. Secondly, introduce endogenous controls to generate more uniformly-sized units at any scale of the process, in order to improve the accuracy of the scale-specific weighting method and the usefulness of the clusters for practical purposes.

Its application to social distress analysis in GS shows a remarkable capacity to identify notorious neighborhoods, in agreement with the identitary, historical and socioeconomic context of the highlighted areas. The areas thus identified can be useful for policy design and for further statistical and qualitative research.

In any case, the results obtained so far consistently show the existence of mathematical connections among correlation matrices and statistical dispersion in the observed multiscalar data aggregations. Remarkably, this leads to a consistent identification of the best level of analysis through the 'elbow' criterion, SS ratios, mean correlation differences between real and random datasets and RR compacity-isolation ratios. As all these measures are based on different relationships between individual and mean values, further research should clarify the arithmetical principles of this convergence.

And above all, the different behaviors of aggregation between real and random datasets clearly show that the emerging patterns in spatial clustering are partially a spurious MAUP effect, yet they reveal a dominant influence of the real coproduction of socio-spatial phenomena. Herein, we have developed a methodology which opens concrete ways to systematically analyze these interactions.

The generation of tools to understand this phenomenon, which is present in most societies, allow you to have greater knowledge of the impact of different policies on the evolution of a city, which will define the optimal policies that improve our welfare. The development of similar models that overcome the limitations of hierarchical clustering represent a natural

extension of this work, as well as finding new cases of application of this method in the design of our cities.

In sum, we are providing useful tools that will contribute to a more rigorous exploration of the black box of spatial interdependence and multiscalar self-organizing phenomena, while linking these questions to relevant real world issues.

6. Bibliography

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. http://doi.org/10.1002/wics.101

Aiolfi, M., Capistrán, C., & Timmermann, A. G. (2010). *Forecast combinations*. CREATES research paper, (2010-21).

Alexander, R. A. (1990). A note on averaging correlations. *Bulletin of the Psychonomic Society*, *28*(4), 335-336.

Amaral, P. V., & Anselin, L. (2014). Finite sample properties of Moran's I test for spatial autocorrelation in tobit models. *Papers in Regional Science*,*93*(4), 773-781.

Andresen, M. A., & Linning, S. J. (2012). The (in) appropriateness of aggregating across crime types. *Applied Geography*, *35*(1), 275-282.

Anselin, L., & Le Gallo, J. (2006). Interpolation of air quality measures in hedonic house price models: spatial aspects. *Spatial Economic Analysis*,*1*(1), 31-52.

Anselin, L., & Lozano-Gracia, N. (2008). Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empirical economics*, *34*(1), 5-34.

Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical analysis*, *27*(2), 93-115.

Bação, F., Lobo, V., & Painho, M. (2005). Self-organizing maps as substitutes for k-means clustering. In *Computational Science–ICCS 2005*(pp. 476-483). Springer Berlin Heidelberg.

Bastian, C. T., McLeod, D. M., Germino, M. J., Reiners, W. A., & Blasko, B. J. (2002). Environmental amenities and agricultural land values: a hedonic model using geographic information systems data. *Ecological Economics*,*40*(3), 337-349.

Basu, S., & Thibodeau, T. G. (1998). Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics*, *17*(1), 61-85.

Berkhin, P. (2006). A survey of clustering data mining techniques. In*Grouping multidimensional data* (pp. 25-71). Springer Berlin Heidelberg.

Berry, B. J. (1961). *A method for deriving multi-factor uniform regions*. Bobbs-Merrill.

Blum, A. (1992). Neural networks in C++. *NY: Wiley*, 60.

Brasington, D. M., & Hite, D. (2005). Demand for environmental quality: a spatial hedonic analysis. *Regional science and urban economics*, *35*(1), 57-82.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis.*Communications in Statistics-theory and Methods*, *3*(1), 1-27.

Carvalho, A. X., Albuquerque, P. H., Almeida, G., & Guimaraes, R. (2009). Spatial Hierarchical Clustering. *Revista Brasileira de Biometria*, 27(3), 411–442.

Chen, D. G., & Ware, D. M. (1999). A neural network model for forecasting fish stock recruitment. *Canadian Journal of Fisheries and Aquatic Sciences*,*56*(12), 2385-2396.

Cheng, B., & Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical science*, 2-30.

Cohen, J. P., & Coughlin, C. C. (2008). Spatial hedonic models of airport noise, proximity, and housing prices*. *Journal of Regional Science*, *48*(5), 859-878.

Court, A. (1939). Hedonic price indexes with automobile examples. *The dynamics of the automobile demand*.

Cutter, S. L., Boruff, B. J., & Shirley, W. L. (2003). Social vulnerability to environmental hazards*. *Social science quarterly*, *84*(2), 242-261.

Davis Jr, G. W. (1989). Sensitivity analysis in neural net solutions. *Systems, Man and Cybernetics, IEEE Transactions on*, *19*(5), 1078-1082.

De Mattos, C. A. (2002). Mercado metropolitano de trabajo y desigualdades sociales en el Gran Santiago:¿ una ciudad dual?. *EURE (Santiago)*, *28*(85), 51-70.

Deif, A. (2012). *Sensitivity analysis in linear systems*. Springer Science & Business Media.

Diebold, F. X., & Lopez, J. A. (1996). Forecast evaluation and combination.

Diebold, F. X. (1988). Serial correlation and the combination of forecasts.*Journal of Business & Economic Statistics*, *6*(1), 105-111.

Dimopoulos, Y., Bourret, P., & Lek, S. (1995). Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Processing Letters*, *2*(6), 1-4.

Dimopoulos, I., Chronopoulos, J., Chronopoulou-Sereli, A., & Lek, S. (1999). Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in Athens city (Greece). *Ecological modelling*, *120*(2), 157-165.

Do, A. Q., & Grudnitski, G. (1992). A neural network approach to residential property appraisal. *The Real Estate Appraiser*, *58*(3), 38-45.

Duque Cardona, J. C., & Artís Ortuño, M. (2004). *Design of homogeneous territorial units: A methodological proposal and applications* (Doctoral dissertation, PhD Dissertation, Departamento de Econometria Estadistica y Economica Espanola, University of Barcelona, Barcelona).

Duque, J. C., Anselin, L., & Rey, S. J. (2012). THE MAX-P-REGIONS PROBLEM*. *Journal of Regional Science*, *52*(3), 397-419.

Duque, J. C., Ramos, R., & Suriñach, J. (2007). Supervised regionalization methods: A survey. *International Regional Science Review*, *30*(3), 195-220.

Figueroa B, E., & Lever D, G. (1992). Determinantes del Precio de Mercado de los Terrenos en el área Urbana de Santiago. *Cuadernos de Economía*, 99-113.

Fischer, A. (2015). How to determine the unique contributions of input-variables to the nonlinear regression function of a multilayer perceptron.*ECOLOGICAL MODELLING*, *309*, 60-63.

Fischer, M. M. (1980). Regional taxonomy: A comparison of some hierarchic and non-hierarchic strategies. *Regional Science and Urban Economics*,*10*(4), 503-537.

Follain, J. R., & Jimenez, E. (1985). Estimating the demand for housing characteristics: a survey and critique. *Regional science and urban economics*, *15*(1), 77-107.

Fujita, M., & Thisse, J. F. (2013). *Economics of agglomeration: cities, industrial location, and globalization*. Cambridge university press.

Galster, G. C. (2012). The mechanism (s) of neighbourhood effects: Theory, evidence, and policy implications. In *Neighbourhood effects research: New perspectives* (pp. 23-56). Springer Netherlands.

Garreton, M., & Sánchez, R. (2016). Identifying an optimal analysis level in multiscalar regionalization: A study case of social distress in Greater Santiago. *Computers, Environment and Urban Systems*, *56*, 14-24.

Garreton, M. (2013). Mobility inequalities in Greater Santiago and the Ile-de-France region : housing and transport policies in metropolitan governance. *PhD Thesis. Université Paris-Est.*

Ge, J. X., Runeson, G., & Lam, K. C. (2003). Forecasting Hong Kong housing prices: An artificial neural network approach. In *International conference on methodologies in housing research, Stockholm, Sweden*.

Gehlke, C. E., & Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, *29*(185A), 169-170.

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average?. *International Journal of Forecasting*, *29*(1), 108-121.

Geoghegan, J., Wainger, L. A., & Bockstael, N. E. (1997). Spatial landscape indices in a hedonic framework: an ecological economics analysis using GIS. *Ecological economics*, *23*(3), 251-264.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis*, *24*(3), 189-206.

Gevrey, M., Dimopoulos, I., & Lek, S. (2006). Two-way interaction of input variables in the sensitivity analysis of neural network models. *Ecological modelling*, *195*(1), 43-50.

Giam, X., & Olden, J. D. (2015). A new R 2-based metric to shed greater insight on variable importance in artificial neural networks. *Ecological Modelling*, *313*, 307-313.

Goodchild, M. F. (1986). *Spatial autocorrelation* (Vol. 47). Geo Books.

Goodman, A. C., & Thibodeau, T. G. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics*, *12*(3), 181-201.

Goonatilake, S., & Khebbal, S. (1995). Intelligent hybrid systems: issues, classifications and future directions. *Intelligent Hybrid Systems, John Wiley & Sons*, 1-20.

Guo, D., Peuquet, D. J., & Gahegan, M. (2003). ICEAGE: Interactive clustering and exploration of large and high-dimensional geodata.*GeoInformatica*, *7*(3), 229-253.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, *22*(7), 801-823.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied statistics*, 100-108.

Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc..

Haas, G. C. (1922). A statistical analysis of farm sales in blue earth county, Minnesota, as a basis for farm land appraisal.

Hendry, D. F. (1995). *Dynamic econometrics*. Oxford University Press.

Henriques, R., Bacao, F., & Lobo, V. (2012). Exploratory geospatial data analysis using the GeoSOM suite. *Computers, Environment and Urban Systems*, *36*(3), 218-232.

Hidalgo Dattwyler, R. (2007). ¿ Se acabó el suelo en la gran ciudad?: Las nuevas periferias metropolitanas de la vivienda social en Santiago de Chile.*EURE (Santiago)*, *33*(98), 57-75.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359-366.

Intrator, O., & Intrator, N. (2001). Interpreting neural-network results: a simulation study. *Computational statistics & data analysis*, *37*(3), 373-393.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263-291.

Kauko, T. (2003). On current neural network applications involving spatial modelling of property prices. *Journal of housing and the built environment*,*18*(2), 159-181.

Kilpatrick, J. (2011). Expert systems and mass appraisal. *Journal of Property Investment & Finance*, *29*(4/5), 529-550.

Kim, C. W., Phipps, T. T., & Anselin, L. (2003). Measuring the benefits of air quality improvement: a spatial hedonic approach. *Journal of environmental economics and management*, *45*(1), 24-39.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, *43*(1), 59-69.

Kong, F., Yin, H., & Nakagoshi, N. (2007). Using GIS and landscape metrics in the hedonic price modeling of the amenity value of urban green space: A case study in Jinan City, China. *Landscape and Urban Planning*, *79*(3), 240-252.

Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(3), 231-240.

Kristjanpoller, W., Fadic, A., & Minutolo, M. C. (2014). Volatility forecast using hybrid Neural Network models. *Expert Systems with Applications*,*41*(5), 2437-2442.

Krupka, D. J. (2007). Are big cities more segregated? Neighbourhood scale and the measurement of segregation. *Urban Studies*, *44*(1), 187-197.

Krzanowski, W. J., & Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 23-34.

Lancaster, K. J. (1966). A new approach to consumer theory. *The journal of political economy*, 132-157.

Landajo, M., Bilbao, C., & Bilbao, A. (2012). Nonparametric neural network modeling of hedonic prices in the housing market. *Empirical Economics*,*42*(3), 987-1009.

Lankford, P. M. (1969). Regionalization: theory and alternative algorithms.*Geographical Analysis*, *1*(2), 196-212.

Lauridsen, J., & Mur, J. (2006). Multicollinearity in cross-sectional regressions. *Journal of geographical systems*, *8*(4), 317-333.

Lefebvre, H. (1974). La production de l'espace. *L Homme et la société*, *31*(1), 15-32.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., & Aulagnier, S. (1996). Application of neural networks to modelling nonlinear relationships in ecology. *Ecological modelling*, *90*(1), 39-52.

Lenk, M. M., Worzala, E. M., & Silva, A. (1997). High-tech valuation: should artificial neural networks bypass the human valuer?. *Journal of Property Valuation and Investment*, *15*(1), 8-26.

Limsombunchai, V. (2004, June). House price prediction: Hedonic price model vs. artificial neural network. In *New Zealand Agricultural and Resource Economics Society Conference* (pp. 25-26).

Lozano Navarro, F. J. (2015). *Elasticidad precio de la oferta inmobiliaria en el Gran Santiago [Housing supply elasticity in Greater Santiago]* (No. 65012). University Library of Munich, Germany.

Malpezzi, S. (2003). Hedonic pricing models: a selective and applied review.*Section in Housing Economics and Public Policy: Essays in Honor of Duncan Maclennan*.

Massey, D. S., & Denton, N. A. (1988). The dimensions of residential segregation. *Social forces*, *67*(2), 281-315.

McCluskey, W., & Anand, S. (1999). The application of intelligent hybrid techniques for the mass appraisal of residential properties. *Journal of Property Investment & Finance*, *17*(3), 218-239.

McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., & McIlhatton, D. (2013). Prediction accuracy in mass appraisal: a comparison of modern approaches. *Journal of Property Research*, *30*(4), 239-265.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159-179.

Monmonier, M. S. (1973). Maximum-Difference Barriers: An Alternative Numerical Regionalization Method*. *Geographical analysis*, *5*(3), 245-261.

Mu, L., & Wang, F. (2008). A scale-space clustering method: Mitigating the effect of scale in the analysis of zone-based data. *Annals of the Association of American Geographers*, *98*(1), 85-101.

Mur, J., López, F., & Herrera, M. (2010). Testing for spatial effects in seemingly unrelated regressions. *Spatial Economic Analysis*, *5*(4), 399-440.

Murray, A. T., & Shyy, T. K. (2000). Integrating attribute and space characteristics in choropleth display and spatial data mining. *International Journal of Geographical Information Science*, *14*(7), 649-667.

Nagel, S. S. (1965). Simplified bipartisan computer redistricting. *Stanford Law Review*, 863-899.

Nghiep, N., & Al, C. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. *Journal of real estate research*, *22*(3), 313-336.

Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, *178*(3), 389-397.

Openshaw, S., & Rao, L. (1995). Algorithms for reengineering 1991 Census geography. *Environment and planning A*, *27*(3), 425-446.

Openshaw, S., & Taylor, P. J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *Statistical applications in the spatial sciences*, *21*, 127-144.

Openshaw, S. (1973). A regionalisation program for large data sets.*Computer Applications*, *3*(4), 136-147.

Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the institute of british geographers*, 459-472.

Özesmi, S. L., & Özesmi, U. (1999). An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological modelling*,*116*(1), 15-31.

Palmquist, R. B. (1984). Estimating the Demand for the Characteristics of Housing. *The Review of Economics and Statistics*, 394-404.

Parrado, E., Cox, P., & Fuenzalida, M. (2009). Evolución de los Precios de Viviendas en Chile. *Economia chilena*, *12*(1), 51-68.

Perruchet, C. (1983). Constrained agglomerative hierarchical classification.*Pattern Recognition*, *16*(2), 213-217.

Peterson, S., & Flanagan, A. (2009). Neural network hedonic pricing models in mass real estate appraisal. *Journal of Real Estate Research*, *31*(2), 147-164.

Pilevar, A. H., & Sukumar, M. (2005). GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases. *Pattern recognition letters*,*26*(7), 999-1010.

Quigley, J. M. (1995). A simple hybrid model for estimating real estate price indexes. *Journal of Housing Economics*, *4*(1), 1-12.

Recknagel, F., French, M., Harkonen, P., & Yabunaka, K. I. (1997). Artificial neural network approach for modelling and prediction of algal blooms.*Ecological Modelling*, *96*(1), 11-28.

Reyjol, Y., Lim, P., Belaud, A., & Lek, S. (2001). Modelling of microhabitat used by fish in natural and regulated flows in the river Garonne (France).*Ecological Modelling*, *146*(1), 131-142.

Ridker, R. G., & Henning, J. A. (1967). The determinants of residential property values with special reference to air pollution. *The Review of Economics and Statistics*, 246-257.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *The journal of political economy*, 34-55.

Sabatini, F., & Brain, I. (2008). La segregación, los guetos y la integración social urbana: mitos y claves. *EURE (Santiago)*, *34*(103), 5-26.

Sagner, A. (2011). Determinantes del precio de viviendas en la región metropolitana de Chile. *El Trimestre Económico*, *78*(312), 813-839.

Salvador, S., & Chan, P. (2004, November). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In*Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on* (pp. 576-584). IEEE.

Sánchez, R. & Villena, M (2016). Hierarchical Systems for Hedonic Mass Appraisal. *Applied Soft Computing (submitted)*

Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm gdbscan and its applications.*Data mining and knowledge discovery*, *2*(2), 169-194.

Scardi, M. (1996). Artificial neural networks as empirical models for estimating phytoplankton production. *Marine ecology progress series. Oldendorf*, *139*(1), 289-299.

Sedgley, N. H., Williams, N. A., & Derrick, F. W. (2008). The effect of educational test scores on house prices in a model with spatial dependence.*Journal of Housing Economics*, *17*(2), 191-200.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, *36*(2), 2843-2852.

Sheppard, S. (1999). Hedonic analysis of housing markets. *Handbook of regional and urban economics*, *3*, 1595-1635.

Sirmans, S., Macpherson, D., & Zietz, E. (2005). The composition of hedonic pricing models. *Journal of real estate literature*, *13*(1), 1-44.

Smith, N. (2002). New globalism, new urbanism: gentrification as global urban strategy. *Antipode*, *34*(3), 427-450.

Specht, D. F. (1991). A general regression neural network. *Neural Networks, IEEE Transactions on*, *2*(6), 568-576.

Spielman, S. E., & Folch, D. C. (2014). Reducing Uncertainty in the American Community Survey through Data-Driven Regionalization. *PloS one*,*10*(2), e0115626-e0115626.

Spielman, S. E., & Logan, J. R. (2013). Using high-resolution population data to identify neighborhoods and establish their boundaries. *Annals of the Association of American Geographers*, *103*(1), 67-84.

Tay, D. P., & Ho, D. K. (1992). Artificial intelligence and the mass appraisal of residential apartments. *Journal of Property Valuation and Investment*,*10*(2), 525-540.

Thorndike, R. L. (1953). Who belongs in the family?. *Psychometrika*, *18*(4), 267-276.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411-423.

Vickrey, W. (1961). On the prevention of gerrymandering. *Political Science Quarterly*, 105-110.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function.*Journal of the American statistical association*, *58*(301), 236-244.

Webster, R., & Burrough, P. A. (1972). COMPUTER-BASED SOIL MAPPING OF SMALL AREAS FROM SAMPLE DATA. *Journal of Soil Science*, *23*(2), 222-234.

White, D., Richman, M., & Yarnal, B. (1991). Climate regionalization and rotation of principal components. *International Journal of Climatology*, *11*(1), 1-25.

Wilson, I. D., Paris, S. D., Ware, J. A., & Jenkins, D. H. (2002). Residential property price time series forecasting with neural networks. *Knowledge-Based Systems*, *15*(5), 335-341.

Wong, D. W. (2004). The modifiable areal unit problem (MAUP). In*Worldminds: Geographical Perspectives on 100 Problems* (pp. 571-575). Springer Netherlands.

Worzala, E., Lenk, M., & Silva, A. (1995). An exploration of neural networks and its application to real estate valuation. *Journal of Real Estate Research*,*10*(2), 185-201.

Zurada, J., Levitan, A., & Guan, J. (2011). A comparison of regression and artificial intelligence methods in a mass appraisal context. *Journal of Real Estate Research*, *33*(3), 349-387.