euro geo European Journal of Geography

Volume 13, Issue 5, pp. 001 - 014

Regular Issue
published quarterly

Editor: Dr Panos Manetos, pmanetos@uth.gr

Article Info
Received: 12/05/2022 · Accepted: 04/10/2022
Corresponding Author: (*) thomasx@survey.ntua.gr
DOI: https://doi.org/10.48088/ejg.t.cha.13.5.001.014

Research Article

Development of a Geodemographic System for Attica, Greece

Thomas **Hatzichristos**[1*], Athanasia **Darra**[1], & Anastasia **Kostellou**[2]

[1] *National Technical University of Athens, Greece*
[2] *Geoinformatics Ltd, Greece*

Keywords
Geodemographics,
GIS,
AI,
Fuzzy clustering,
Fuzzy C-Means

Abstract

Recent years have witnessed fundamental changes in the financial and retail service sector, with competition among companies intensifying as a result of changing markets. The need to address the specific requirements of the customer groups has become the guiding principle behind the business strategies adopted by companies. These customer needs and, subsequently, the provision of the appropriate products and services to the customer, are dependent on where they live, their personal characteristics (e.g., age, education, income, households etc.). Geodemographic systems take advantage of information technology to analyze these types of data for a better understanding of the consumer characteristics and improved performing of marketing strategies. In the present study, a geodemographic system is implemented utilizing geographic information systems (GIS) technology and artificial intelligence (AI). GIS technology offers a powerful set of tools for the input, management, and visualization of data, while AI provides advanced analytical tools such as the unsupervised fuzzy classification through the Fuzzy C-Means algorithm. The proposed methodology is applied to the Attica region in Greece. It uses the official socioeconomic data of the Hellenic Statistical Authority. The relevant database uses 78 socio-economic variables. The study area consists of 2500 area units, each one with more than 1000 inhabitants. The results, ten socio-economic classes, are analyzed and discussed. The project is funded by the Athens Chamber of Commerce and Industry (ACCI) and the results of the analysis are open to all users by the official site of ACCI.

Highlights:

- *State of the art multivariable geographical analysis for enterprises*
- *Implementation of AI and Fuzzy C means to analyze and understanding the underlying consumer characteristics*
- *Methodological framework for improved performing of marketing strategies for enterprises*

1. INTRODUCTION

Recent years have witnessed fundamental changes in the financial and retail service sector, with competition among companies intensifying as a result of economic crisis and changing markets. The need to address the specific requirements of the customer groups has become the guiding principle behind the business strategies adopted by companies. These customer needs and, subsequently, the provision of the appropriate products and services to the customer, are dependent on where they live, their personal characteristics (e.g., age, education, income, households etc.). As Jansen and colleagues describe location impacts on almost all aspects of life “Difference in location might influence various aspects, such as social status, the consumption of private goods, the availability of public goods, jobs, education, and other desired aspects. Location is, therefore, an important determinant of household welfare” (Jansen et al. 2011).

Geodemographic systems take advantage of information technology to analyze these types of data for a better understanding of the consumer characteristics and improved performing of marketing strategies. This research proposes the development of a geodemographic system as a main component of any Spatial Decision Support System concerning the companies based in Attica, Greece. The system is implemented utilizing geographic information systems (GIS) technology and artificial intelligence (AI). GIS technology offers a powerful set of tools for the input, management, and visualization of data, while AI provides the appropriate advanced analytical tools for clustering.

The concept of geodemographics is relatively new, having a history of almost 40 years (Harris et al. 2005). Geodemographics is depicted as “the analysis of social and economic data in a geographical context for commercial purposes related to marketing, site selection, advertising and sales forecasting” (Johnston et al. 2000). Also, Sleight’s definition is “the analysis of people by where they live” (Sleight, 1997). Birkin (1995) and Harris et al. (2005) both emphasize the broad range of indicator variables used in such systems; housing, socioeconomic and demographic characteristics, which, when supplemented by geography, reaffirm the term and its composition, hence geography plus demography and its name of “geo-demo-graphics” (Birkin, 1995, Harris et al. 2005). The addition of lifestyle variables, although expensive, has revolutionized geodemographics and lessened its overreliance on census data (Burns, 2014).

The ACORN and Mosaic systems, now-named CACI and Experían, are two of the geodemographic classifications available commercially in the UK. Tapestry is the most widely used commercial geodemographic system in the US. However, there is no such geodemographic system existing in the private or public sector in Greece.

Clustering techniques constitute the core of geodemographic systems and the methods used are critical for their effectiveness. There is an extensive bibliography on clustering techniques in geodemographic analysis, that is discussed in section 2. Our main scope is to present the advantages of fuzzy clustering using a geographic database with the appropriate variables for customer segmentation analysis in general and to support the marketing strategies of the companies at the metropolitan Athens area.

The proposed methodology is applied to the Attica region in Greece. It uses the official socioeconomic data of the Hellenic Statistical Authority. The relevant database uses 80 socio-economic variables. The study area consists of 2500 area units, each one with more than 1000 inhabitants. The resulted ten socio-economic classes are analyzed and discussed with focus on the real estate market.

The rest of the paper is structured as follows. In section 2, the methodological framework is analyzed and in section 3, the case study in Athens, Greece is presented. Finally, section 4 concludes the paper.

2. METHODOLOGICAL FRAMEWORK

The generation of homogenous demographic groups is a process having several stages (Fig. 1). Among them the most objective, from a methodological point of view, is the selection of the appropriate classification algorithm (Fig. 1-D). This paper describes a methodology for the delineation of homogenous demographic groups utilizing tools of artificial intelligence, more specifically fuzzy clustering. In addition, GIS technology is necessary, to provide a powerful set of tools for the management and visualization of the information, while it is the ideal environment for its analysis. The integration of GIS with AI is a valuable addition in GIS toolbox. GIS can increase their analytical abilities through the incorporation of fuzzy clustering techniques.

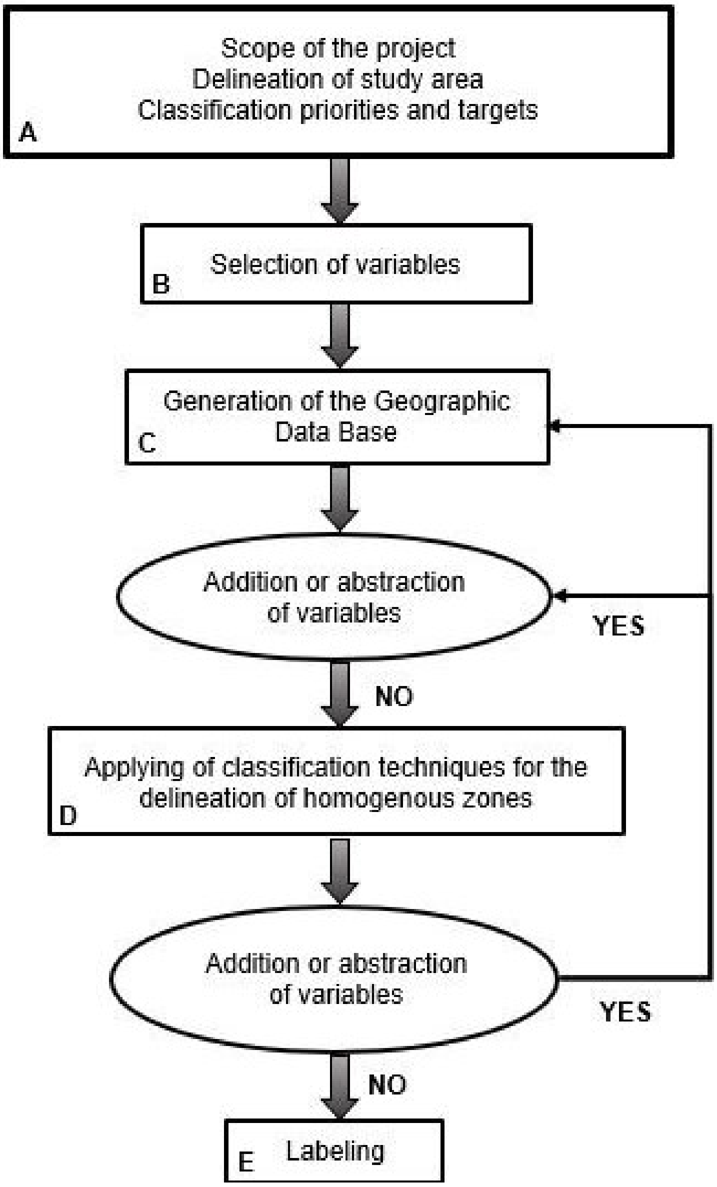

Figure 1. Methodological Framework

Even today, most of the known systems use a conventional statistical classifier. This method and its drawbacks have been well analyzed by S. Openshaw (Openshaw, 1992). Openshaw suggested an unsupervised neural network like Kohonen's Self-Organizing map (SOM), or alternatively Adaptive Resonance Theory (ART) (Openshaw, 1996) for addressing these disadvantages. Both methods have the drawback that they are deterministic, following the “black or white” nature of Boolean logic and thus they do not consider, as they should, the fuzziness of geographic characteristics.

Classic Boolean logic is binary, that is a certain element is true or false, an object belongs to a set or it does not. Fuzzy logic, introduced by Zadeh in 1965 permits the notion of nuance (Zadeh, 1965). Apart from being true, a proposition may be anything from al-most true to hardly true (Kosko, 1994). In comparison with the Boolean sets, a fuzzy set does not have sharply defined boundaries. The notion of a fuzzy set provides a convenient way of dealing with problems in which the source of imprecision is the absence of sharply de-fined criteria of class membership rather than the presence of random variables. Figure 2 illustrates the difference between the ordinary Boolean sets and Fuzzy sets.

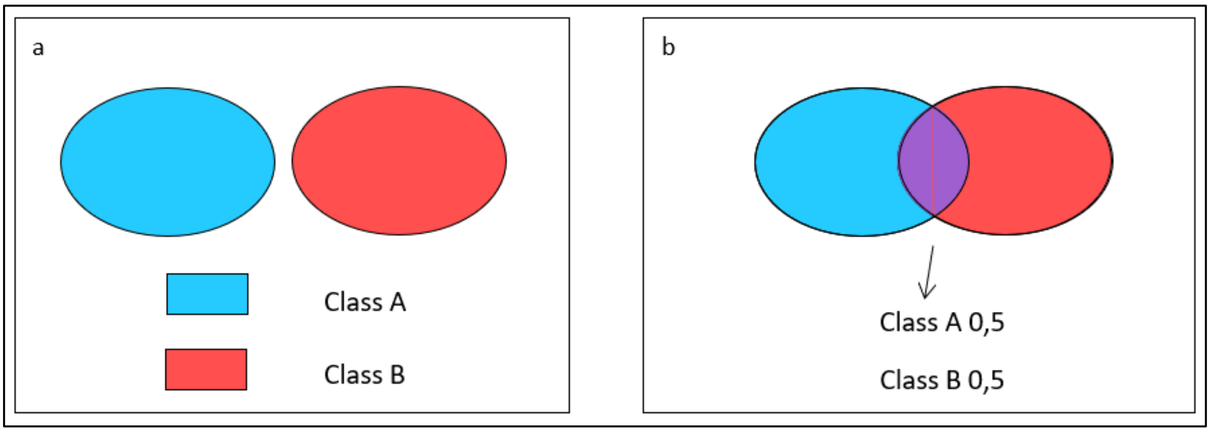

Figure 2. (a) Boolean Sets (b) Fuzzy Sets

A significant defect in any binary classification is that each spatial unit to be classified, is unequivocally grouped with other units of its group, and thus bears no similarity to members of other groups which is not true. One way to characterize an individual areal unit’s similarity to all the groups was introduced in 1965 by Zadeh (Zadeh, 1965). The key to Zadeh’s idea is to represent the similarity a unit shares with each group with a function (termed the membership function) whose values (called memberships) are between 0 and 1. Each areal unit will have a membership in every group, memberships close to unity signify a high degree of similarity between the unit and the group, while memberships close to zero imply little similarity between the unit and that group. Additionally, the sum of the memberships for each areal unit must be unity. There are several algorithms available to perform the task previously mentioned. Among them the Fuzzy C-Means (FCM) is the best known and widely used algorithm designed to perform fuzzy clustering (Bezdek, 1981 & 1984). The FCM algorithm assumes that the number of classes, and the amount of fuzziness are known in advance and minimizes an objective function to find the best set of classes. The clustering criterion used by the FCM algorithm is associated with the generalized least-squared errors function. To check the quality of the clustering obtained, the validity criteria that are usually applied are the partition coefficient F, the proportion exponent, and the partition entropy H, while other indices have been proposed (Grekousis, 2013).

The use of the FCM algorithm for geodemographic analysis has been pointed out by many researchers (Burrough, 1992, Hatzichristos, 2004, See et al. 2001). Other fuzzy algorithms that improve FCM have also been used for geodemographics, such as Gath Geva and Gustafson Kessel (Grekousis et al. 2013). The FKN algorithm that combines SOM with FCM is an interesting proposition (Bezdek et al. 1994, Hatzichristos, 2004). Also, a geographically weighted clustering algorithm is also useful in cases that continuity is desired (Grekousis, 2021). In this paper the FCM algorithm is suggested because it is easier to be implemented by the potential users. There are many software packages that utilize the Fuzzy C-Means algorithm in comparison with the other alternatives. The scope of the paper is to point out the advantages of fuzzy clustering compared to the binary classifiers and not selecting the best available fuzzy clustering algorithm. The Fuzzy C-Means algorithm requires two parameters as input. The number of classes and the amount of fuzziness defined by the parameter m. To establish whether the optimal number of classes have been found, FCM must be run several times with various numbers of classes in the desired range. To select the optimal number of classes, we can plot the three validity criteria over the number of classes. The optimal number of classes c is the number for which the values of each validity criterion is the lowest of the curve at the transition from c-1 to c.

The overlapping of the classes is defined by the m-parameter. The procedure of the definition of the m parameter is subjective. We propose two criteria to support the selection of its optimal value:

a. The exploratory analysis of the membership values and

b. The spatial homogeneity of the classes.

The final phase is the explanation and the labelling (Fig, 1-E) of the classes, based on the statistics for each class, derived by the algorithm. The proposed methodology is applied to Athens, Greece and the results are analyzed in the following section.

3. CASE STUDY

The scope of this project is to support the marketing strategies of the companies and entrepreneurs operating in Attica, Greece. (Figure 1-A). The proposed approach is applied to the area covered by the Athens Chamber of Commerce and Industry, which consists of the Attica Region, except the Piraeus area in Greece. This region has a population of approximately 4 million people. Attica is marked by a high diversity in the spatial distribution of ages, professions and housing types and can be used as a paradigm for other urban areas.

The spatial analysis, for reasons of confidentiality, was carried out at the spatial unit of the "Spatial Analysis Units of Cities" (SAUC), which are small units that correspond approximately to the Inventory Sectors of Hellenic Statistical Authority (HSA) and are only available from the HSA (Figure 3). Attica consists of 2500 SAUCs with about 1000 inhabitants each. Piraeus region is excluded because it is covered by another Chamber.

Although there is a great diversity between the variables used in different geodemographic systems, there is an absence of a suitable theory to guide the variables' selection (Spielman et Thill 2008). The final choice depends to a great degree on the problem at hand (Spielman et Thill 2008), on the system used and from the desired configuration of the socioeconomic clusters (Boyer & Burgaud 2000). According to Openshaw [18] the basic variables are child and adult ages, the number of persons in a household, migration, religion, ethnicity, affluence, occupation, and suburban lifestyle.

The complexity of Attica's social landscape makes it ideal for geodemographic exploration. The variables are selected based on our experience taking into account the relevant literature previously described, from the wider dataset of HSA. Highly correlated variables were excluded. The variables used are illustrated in Table 1 (Fig. 1-B).

The final geographic database (Fig. 1-C) is consisted by the attributes (78 variables) and by the geometrical boundaries of the SAUCs. The following step is fuzzy clustering (Fig. 1-D). The Fuzzy C-Means algorithm is used. To utilize the previously described geographic database with the FCM algorithm the values of the data base are normalized to a scale from 0 to 100. The algorithm requires two parameters as input, the number of classes and the parameter m, defining the amount of fuzziness. Usually, m takes values between $1 < m < 3$. In this case, the initial parameter of m was set to be 1.4. A critical issue is the selection of the optimal number of classes, in a desired predefined range. This can be accomplished with the help of the cluster validity parameters, provided by the software. Several runs need to be carried out with a different number of clusters being specified for each run (in this example between 5 and 12, where m = 1.4), to establish the optimum number of clusters. The cluster validity criteria are shown in Fig. 4 (a), (b) & (c) below. According to these indices, the best partitioning is achieved with 10 clusters.

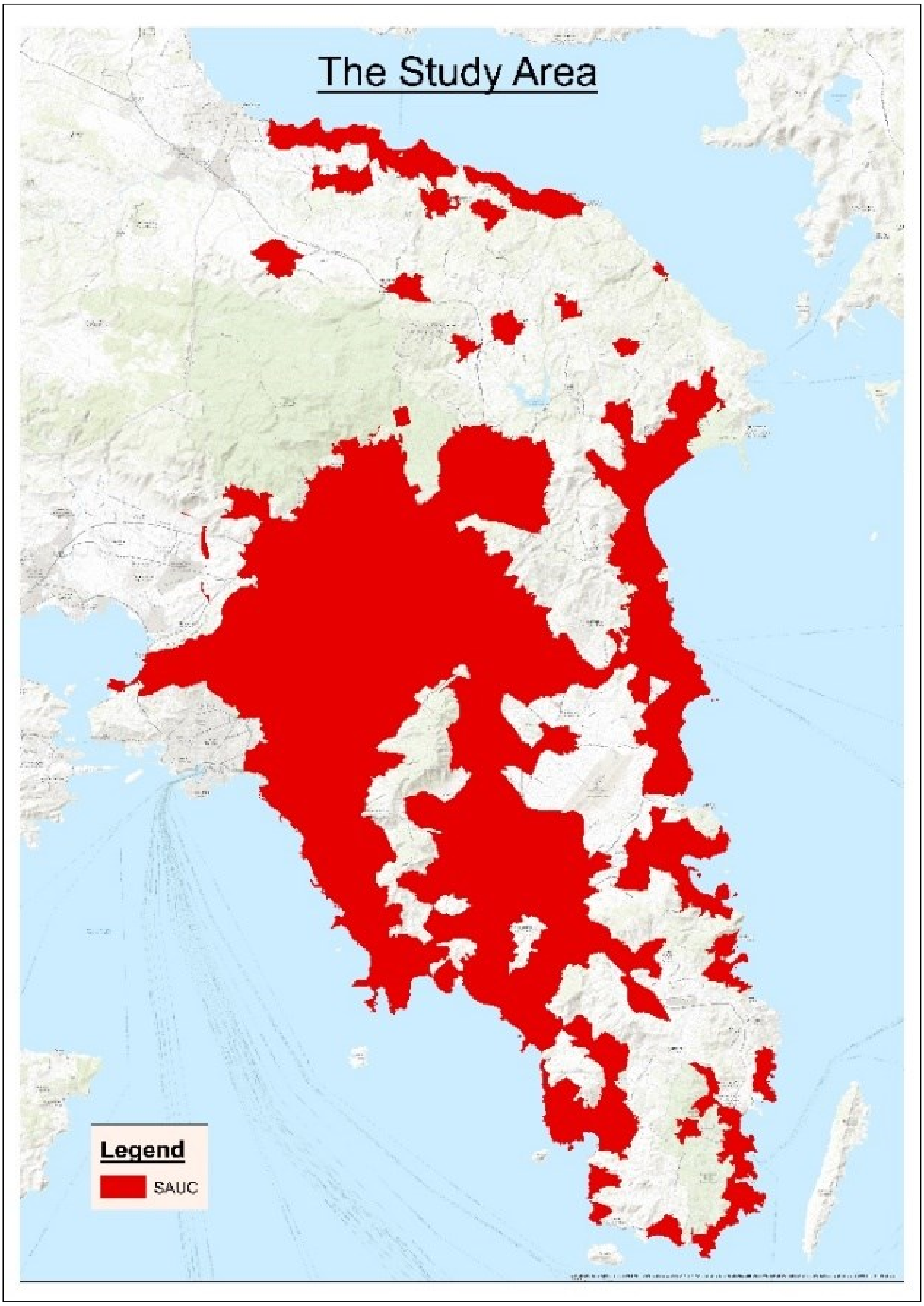

Figure 1. The Study Area – Attica, Greece

Table 1. The variables per spatial unit (SAUC)

#	Variable	#	Variable
1	Set of households	40	Building construction between 1981-2000
2	Females	41	Building construction from 2001 onwards
3	Males	42	Detached houses
4	Ages 0-14 (infants and children)	43	Duplexes
5	Ages 15-24 (students)	44	Apartment building
6	Ages 25-34 (young adults)	45	Main use not residence
7	Ages 35-44 (adults)	46	Without parking spot
8	Ages 45-54 (middle aged)	47	With 1 parking spot
9	Ages 55-64 (less productive)	48	With 2 parking spots
10	Ages 65-74 (recently retired)	49	With ≥ 3 parking spots
11	Ages ≥ 75 (elderly)	50	With internet access
12	Foreigners	51	Without internet access
13	Married population	52	Houses < 30 m^2
14	Single population	53	Houses 31-50 m^2
15	Unmarried population	54	Houses 51-70 m^2
16	Households without children	55	Houses 71-90 m^2
17	Households with 1 child	56	Houses 91-120 m^2
18	Households with 2 children	57	Houses 121-200 m^2
19	Households with 3+ children	58	Houses ≥ 200 m^2
20	Compulsory education	59	Employed
21	Higher education	60	Unemployed
22	Highest education	61	Employers
23	With PhD	62	Self-employed
24	Households with 1 member	63	Employees
25	Households with 2 members	64	Pensioners
26	Households with 3 members	65	Pupils / Students
27	Households with 4 members	66	Dealing with household chores
28	Households with ≥ 5 members	67	Incomers
29	Houses with 1 room	68	Other case of income
30	Houses with 2 rooms	69	Managers
31	Houses with 3 rooms	70	Professionals
32	Houses with 4 rooms	71	Technologists
33	Houses with ≥ 5 rooms	72	White collar workers
34	Homeowners	73	Farmers / Livestock breeders etc.
35	Residential tenants	74	Craftsmen
36	Associations	75	Unskilled
37	Building construction before 1919	76	Services
38	Building construction between 1919-1960	77	Population density
39	Building construction between 1961-1980	78	Housing density

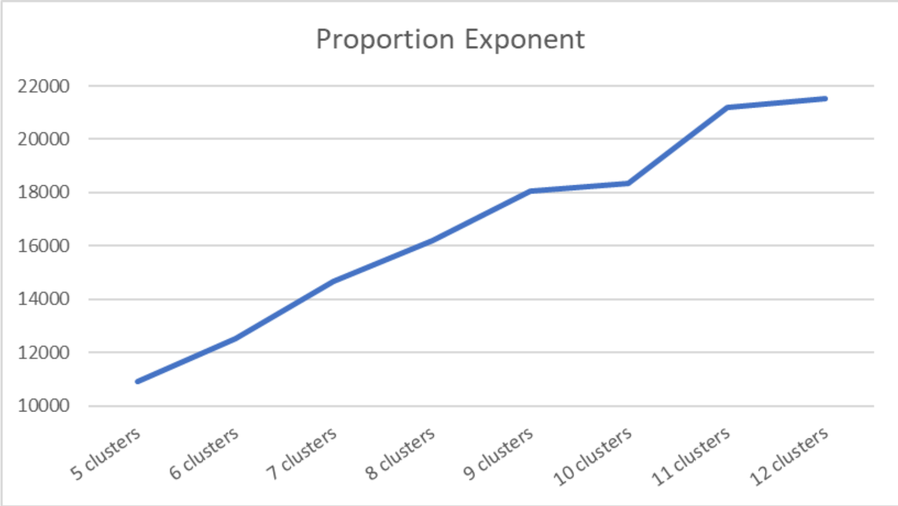

Figure 4 (a). Partition Coefficient

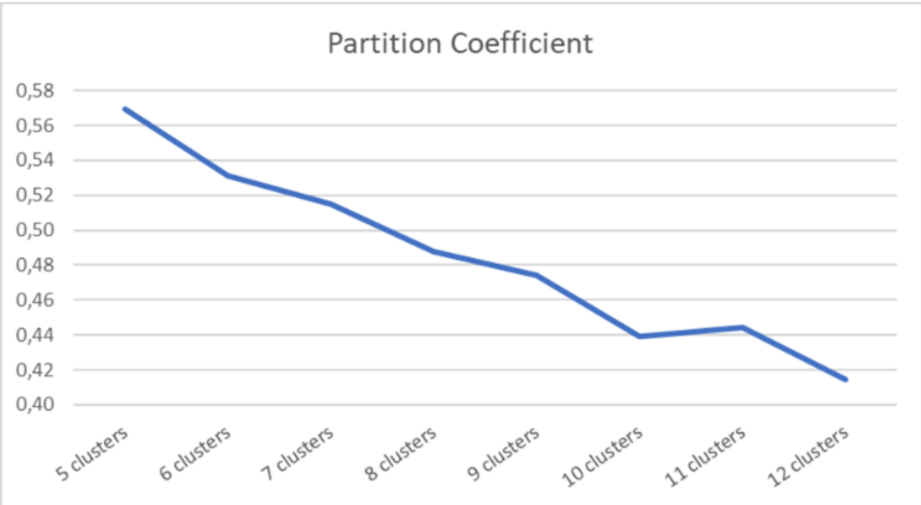

Figure 2 (b). Partition Exponent

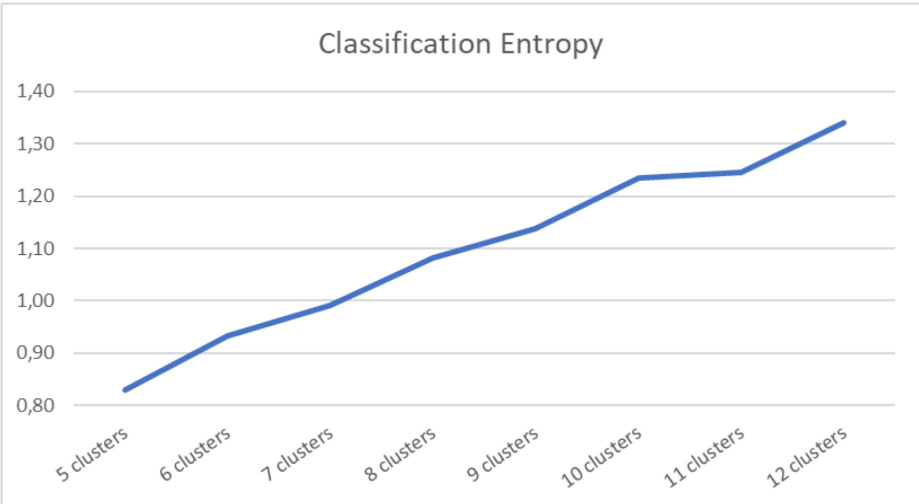

Figure 4 (c). Classification Entropy

Following the selection of the optimal number of clusters the next step is the selection of the optimal value of the parameter m. The overlapping of the clusters is de-fined by the m-parameter. The selection of the optimal value of the m parameter is based on two criteria:

1. The exploratory analysis of the membership values
2. The spatial homogeneity of the clusters (in this case for m=1.4, m=1.3, m=1.2)

The value 1.2 is selected for the m parameter and the classification process is carried out for c=10 and m=1.2. Two arrays are derived from this process. The first one (Table 2) displays each variable's cluster center for every cluster and the second one (Table 3) the membership values of each SAUC for all clusters.

Table 1. Clusters centers (part - 6 out of 78 variables)

Clusters\Variables	Ages 0-14	Females	Singles	Households with 3+ children	Higher education	Households with 3 members
A	25	30	17	25	27	29
B	16	30	24	19	32	29
C	34	43	26	28	37	43
D	25	34	23	27	34	36
E	24	50	44	29	51	46
F	28	37	22	23	36	40
G	46	73	54	55	72	72
H	20	32	24	18	29	33
I	39	45	26	47	41	42
J	18	29	22	20	30	30

Table 2. Fuzzy classification memberships (part – 10 out of 2500 SAUCs)

SAUCs\Clusters	A	B	C	D	E	F	G	H	I	J
1	0%	6%	0%	62%	0%	20%	0%	1%	0%	10%
2	1%	0%	0%	9%	0%	80%	0%	3%	0%	7%
3	1%	0%	0%	48%	0%	7%	0%	2%	0%	42%
4	1%	0%	0%	4%	0%	0%	0%	1%	0%	95%
5	0%	0%	0%	7%	26%	1%	0%	0%	64%	1%
6	0%	1%	0%	0%	0%	0%	0%	0%	0%	99%
7	2%	0%	0%	2%	0%	0%	0%	0%	0%	96%
8	0%	1%	0%	1%	0%	0%	0%	0%	0%	97%
9	0%	0%	0%	0%	98%	0%	0%	0%	1%	0%
10	0%	0%	0%	1%	0%	0%	0%	0%	0%	99%

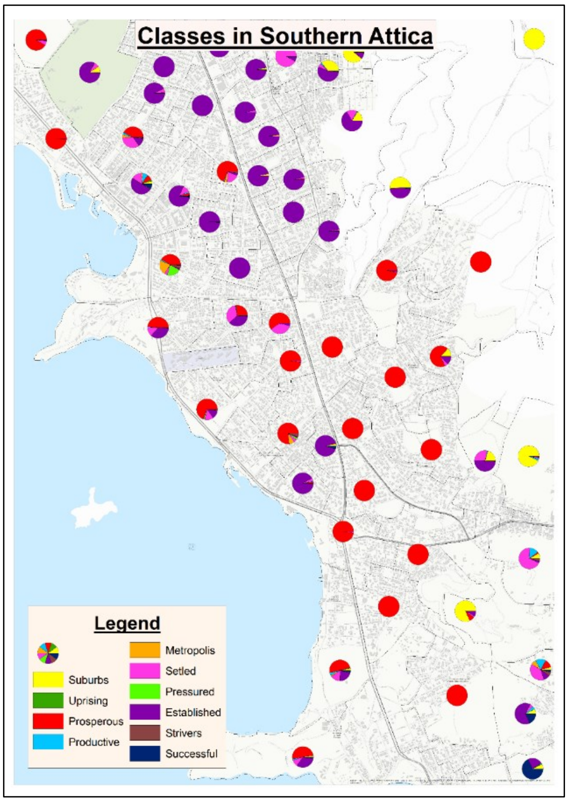

Figure 5. Fuzzy Classification results in Southern Attica

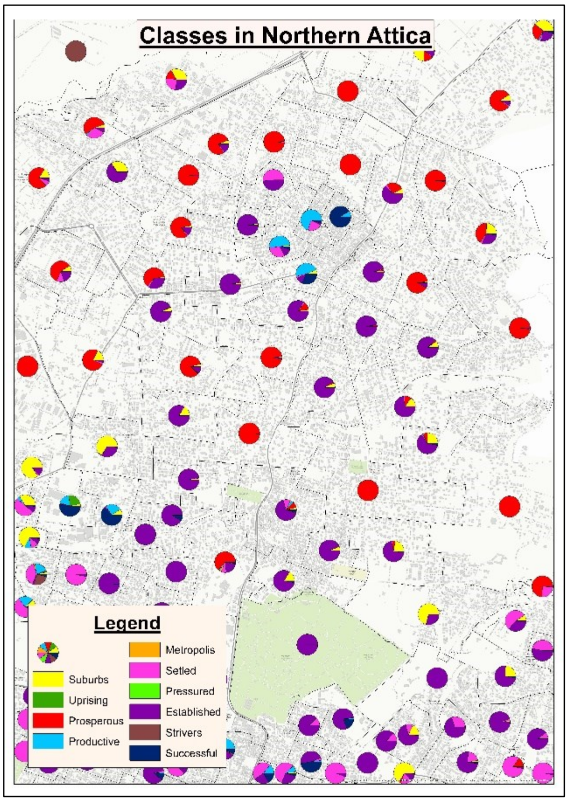

Figure 6. Fuzzy Classification results in Northern Attica

The segmentation of the area by means of fuzzy clustering and GIS is now complete. The demographic data are classified into 10 clusters using the validity criteria. Part of the results are illustrated in Figures 5 & 6. Each SAUC is not assigned exclusively to one class only, but rather has a membership value for each class. The next step in the delineation of the homogenous classes is the interpretation and description of the classes. This is also a subjective step. This procedure is supported by the estimation of the class's centers for each variable (Table 2). Each cluster is described according to its dominant characteristics. For example, the synoptic presentation of the class "Prosperous" (Figures 7 & 8) follows:

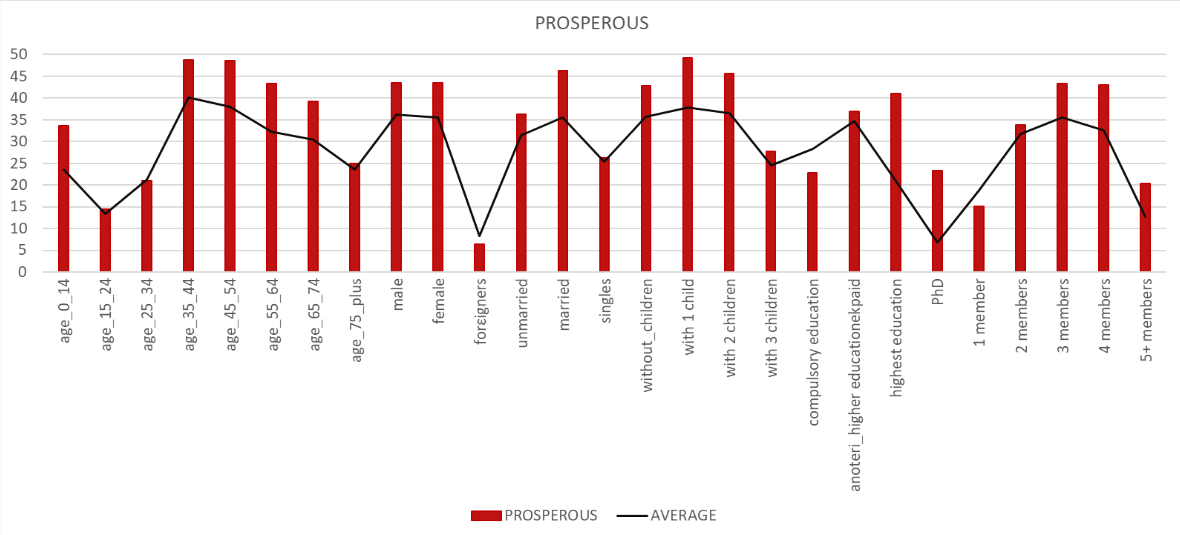

Figure 7. Average values per variable for the class "Prosperous" (part - 27 out of 78 variables)

This class includes households with several children and very high percentages of people aged 35-55 in relation to the average (AVG). However, the areas are sparsely populated, as both population density and housing density are quite low. Foreigners are few. In terms of marital status, married people have a higher value than the AVG. There are more families with one child, followed by those with two children. In terms of the size of the families, families with three and four members are the most often observed and in fact with a value that far exceeds the AVG. In terms of education level, people with higher education and holders of doctoral degrees have the highest percentages, compared to all other groups. Regarding the size of the houses, this group mainly includes houses with four rooms, while houses with more than five rooms have the highest frequency of all the other groups. Also, mainly owners rather than tenants live in this area. Most of the houses are built after 1981, so this is a group with mainly new constructions, with the basic type of building being a two-storey building and with the majority having one or more parking spaces. The area size of most houses is over 121 sq.m., while the houses with an area size over 201 sq.m. is 5 times above the AVG. Regarding the working status of the residents, the percentage of employees is above the AVG, while the percentage of the unemployed is below the AVG. Employees and retirees are close to the AVG. The self-employed, employers, executives and earners have the largest percentage compared to the other groups. The proposed geodemographic system is visualized as interactive maps through a Web Map Application, hosted in the website of Athens Chamber of Commerce and Industry (in Greek at the moment).

Typically, geodemographic systems are being used as demand indicators for service providers. A geodemographic system can help marketeers to better understand the target group to which they should look for potential clients. For example, in the areas of South and North Attica (as shown in Figures 5 & 6), where the socio-economic profile of the residents corresponds to the upper class, (Prosperous) marketing departments should focus on products or services who have, or would like to have, common characteristics with those described above. The same logic applies for the rest of the classes. Another aspect for further research concerning the spatial patterns' is the high population density and the increase in rental price that characterizes the greater urban area (Tsoutsos & Photis 2019). Another factor for consideration is the spatial competition in retailing as argued by Wieland (Wieland 2018).

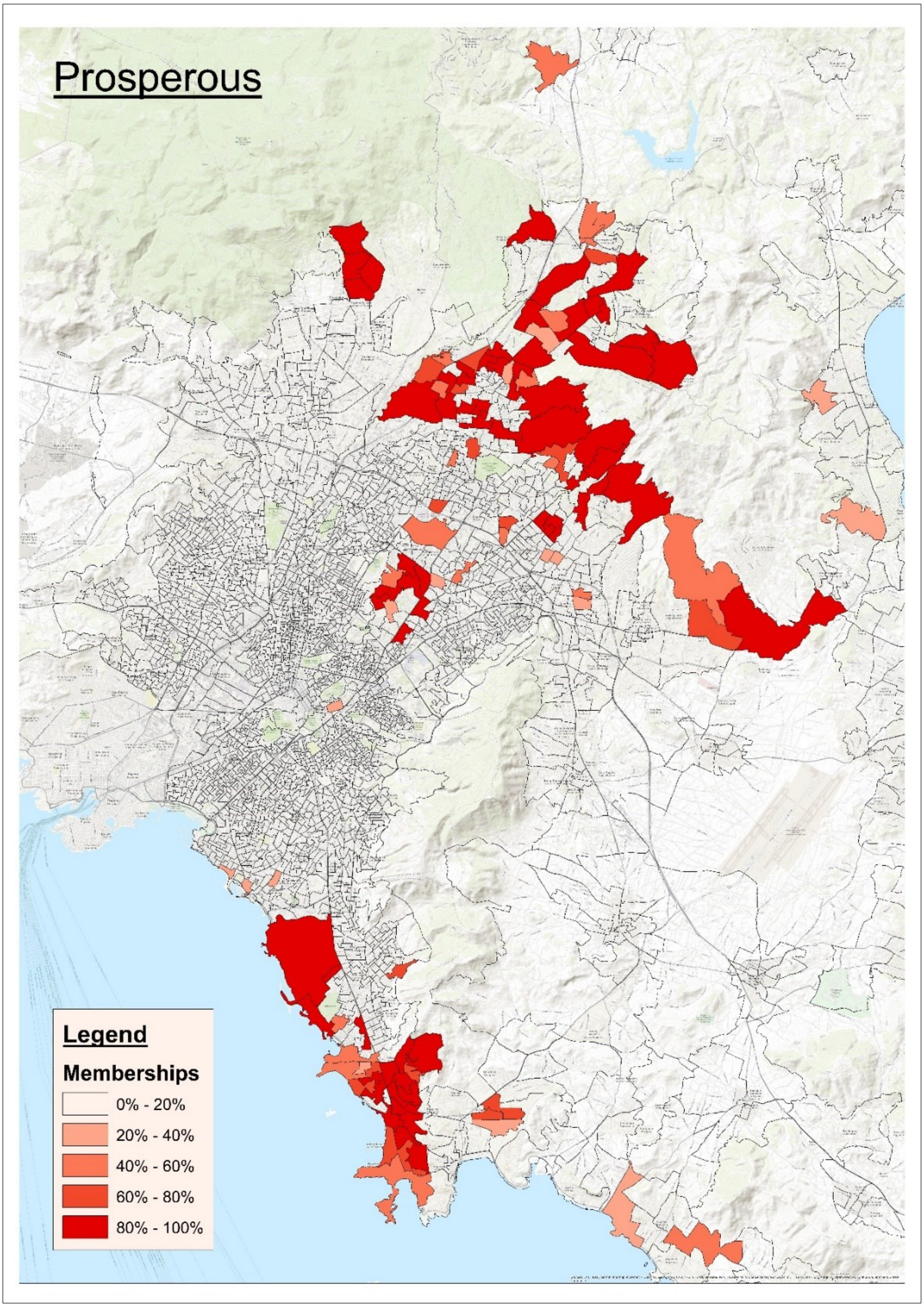

Figure 8. Class "Prosperous"

4. CONCLUSIONS

In this paper a contemporary methodology for supporting marketing strategies for the companies and entrepreneurs of Attica region, Greece is developed. GIS technology, which most researchers view as the obvious environment for the handling of geographic information is used. This methodological approach is enhanced by appropriate AI techniques for the analysis of the data. The methodological framework as well as the technological tools used, can be applied to other regions with the same characteristics too, as part of a Spatial Decision Support System for Marketing purposes.

In the case of geodemographic analysis, where the interest in tracing even small differences of the individual habits of the general trends is focused, fuzzy clustering provides a detailed description.

The fuzzy logic approach to clustering provides a set of advantages, among which the most important are:

- the method is not influenced by noise,
- the resulting information is maximized, and
- the membership values can be further used as an input to regionalization algorithms.

It should be noted that conventional wisdom about geographic analysis is based on the binary model of space. This is a formidable barrier given that traditional techniques have existed for a prolonged period and are widely accepted. Also, recent studies propose classification analysis at the level of the Individual [6]. However, this realization constitutes a strong alternative solution at a time characterized by the revolution of Artificial Intelligence tools and by a proliferation of data.

ACKNOWLEDGEMENTS

The development of the geodemographic system was funded by the Athens Chamber of Commerce and Industry (ACCI).

REFERENCES

Bezdek, C.J. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

Bezdek C. J., Ehrlich R, Full W. (1984). FCM: the Fuzzy C-Means Clustering Algorithm, In Computers & Geosciences, Vol. 10 No 2-3, 191- 203.

Bezdek J., Tsao E., Pal N. (1994). Fuzzy Kohonen clustering networks, Pattern Recognition, Volume 27, pp. 757-764

Birkin M. (1995). Customer Targeting, Geodemographics and Lifestyle Approaches., GIS for Business and Service Planning, In Longley, P. and Clarke, G., Eds; Cambridge: London.

Boyer L., & Burgaud D. (2000). Le marketing avance. Du one-to-one au e-business, Organisation, Paris.

Burns P. (2014). Geodemographics: Creating a Classification at the Level of the Individual, PhD Thesis, University of Leeds, England.

Burrough P. (1992). Fuzzy Classification Methods for Determining Land Suitability from Soil profile Observation and Topography, Journal of Soil Science Volume 43, 193-210.

Grekousis G. (2021). Local fuzzy geographically weighted clustering: a new method for geodemographic segmentation, IJGIS, Volume 35, 152-174.

Grekousis G, (2013). Giving fuzziness to spatial clusters: A simple index for choosing the optimal number of clusters. International Journal on Artificial Intelligence Tools, Volume 22, 1350009.

Grekousis G., Hatzichristos T. (2013). Fuzzy clustering analysis in geomarketing research. Environment and Planning B: Planning and Design, Volume 40, 95-116.

Harris, R., Sleight, R., & Webber, R. (2005). Geodemographics, GIS and Neighbourhood Targeting, Wiley Publishing, Chichester.

Hatzichristos T. (1999). Delineation of Demographic Regions using GIS and Fuzzy Classification, GIS Planet Conference, Lisbon, Portugal

Hatzichristos T. (2004). Delineation of Demographic Regions using GIS and Computational Intelligence, Environment and Planning B, 31 (1), 39-49

Jansen S., Coolen H., and Goetgeluk R. (2011). The Measurement and Analysis of Housing Preference and Choice, Dordrecht: Springer.

Johnston RJ., Gregory D., Pratt G., & Watts, M. (2000). The Dictionary of Human Geography, 4th ed.; Blackwell Publishers: Oxford.

Kosko B. (1994). Fuzzy Thinking; Flamingo Press: New York.

Openshaw S. (1996). Census Users Guide, Geoinformation International: Cambridge, 239-268.

Openshaw S. (1995). Census user's handbook, Geoinformation International: Cambridge.

Openshaw S. (1992). Developing Appropriate Spatial Analysis Methods for GIS. In Geographical Information Systems: Principles and Applications, David J. Maquire, Michael F. Goodchild and David W. Rhind Eds.; Longman: London, 389-402.

See L., Openshaw S. (2001). Fuzzy Geodemographic targeting. In Regional Science in Business, G. Clarke, M. Masson Eds.; Springer: Germany, 269-282.

Sleight P. (1997). Targeting Customers: How to use Geodemographic and Lifestyle Data in your Business, NTC Publications Henley-on-Thames: UK.

Spielman S. E., & Thill J. C. (2008). Social area analysis, data mining and GIS. Computers, Environment and Urban Systems, Volume 32, 110-122.

Tsoutsos M.C. & Photis Y.N. (2020). Spatial point pattern analysis of urban retail stores: the case of twelve large- and medium-sized Greek cities, European Journal of Geography, 11(4), 36-63, https://doi.org/10.48088/ejg.m.mar.11.4.36.63

Wieland T. (2022). Spatial patterns of excess mortality in the first year of the COVID19 pandemic in Germany, European Journal of Geography, 13(4), 18-33, https://doi.org/10.48088/ejg.t.wie.13.4.018.033

Zadeh L.A. (1965). Fuzzy sets, Information and Control, Volume 8, 338-353.