Representing Semantic Similarity of Socioeconomic Units with Cartographic Spatialization Metaphors

Athanasia Darra, Marinos Kavouras, Eleni Tomai
School of Rural and Surveying Engineering
National Technical University of Athens
15780 Zografou Campus, Athens, Greece
{nancyd, mkav}@survey.ntua.gr, etomai@mail.ntua.gr
http://ontogeo.ntua.gr/

SUMMARY

An important subject in geographic analysis is the representation of socioeconomic units (SEUs) on the basis of their characteristics. As a result, similarities and heterogeneities need to be revealed and portrayed accordingly. In this paper, we present a method for the analysis of a particular kind of socioeconomic units, which are the islands of the Greek archipelago. In order to facilitate the analysis but also to convey its results, SEUs are depicted using a two-dimensional "map" metaphor whose dimensions are not geographic but the dimensions of the socioeconomic analysis. The spatialization method used is multi-dimesional scaling (MDS). The resulted map can be actually considered as a cartogram for it does not preserve geographic location or topology but rather the semantic similarity of SEUs.

KEYWORDS: socioecononic units, similarity, spatialization, cartograms, MDS

BACKGROUND

A great category of phenomena which exhibit a special behaviour and are of importance to geography, are those expressing the socioeconomic dimensions of space. The definition of geographic entities on the basis of their socioeconomic characteristics creates the so called socioeconomic units (SEUs). When - and it is often the case – the spatial characteristics play a defining role in SEUs, these are also known as spatial SEUs or simply SSEUs (Frank et al., 2001). Ontological research at NTUA (http://ontogeo.ntua.gr) has focussed on clarifying different kinds of SEUs and their semantic relations/properties. The objective is to determine the defining/essential characteristics of SEUs, that is, those which provide their identity. This is the only way to be able to compare effectively different SEUs and truly integrate them. An important process in the semantic integration framework presented by Kavouras (2003) is the ontology comparison and a formal determination of their similarity. An important means for revealing semantic similarity is visualization. The purpose of this paper is a cartogrammatic portrayal of SEUs representing their semantic similarity using spatialization techniques. The ultimate goal is the study of heterogeneities, data mining and compression, as well as, the ability to interpret the socioeconomic differences of the units under study. Assessing and visualizing semantic similarity is a real challenge for geographic information science and modern cartography.

According to Kuhn & Blumenthal (1996), "Spatialization maps physical space to abstract domains .. in user interfaces .. through spatial metaphors". According to Skupin & Buttenfield (1997), spatialization is "... a projection of elements of a high-dimensional information space into a low-dimensional, potentially experiential, representational space". The two definitions are not different but complimentary. They just view spatialization from different viewpoints. In our case, the first definition expresses the creation of the cartogram, while the second reflects more the spatialization procedure followed.

As an application SEUs for testing this methodology, we selected the inhabited islands of the Greek archipelago. This large number of islands presents great differences in size, population, tourism activity, and economic significance, while they are administratively and geographically arranged into groups. They are also extremely interesting SEUs for the following reasons:

- They constitute (bona fide) (Smith, 2001) units of multidimensional socioeconomic activity.
- They possess definite boundaries.
- They do not share boundaries with other SEUs.
- They exhibit interior coherence.
- They form wider networks or clusters.
- They have an easily recognizable figure (essential to cartograms).
- Some may potentially play the role of an attraction pole (especially if it is far from other existing centres such as the large cities in mainland Greece).

In order to determine semantic similarity of the SEUs, it was thus necessary to select data pertaining to their socioeconomic activities and development-growth. The primary indicators of the degree and potential of development are usually related the existing infrastructures. An infrastructure is defined as the combination of material installations together with a system of institutional and additional supportive services with the aim to cover social needs and service economic activities, contributing thus to growth.

The spatialization method employed was MDS (Multi-Dimensional Scaling). Its rationale is to portray multidimensional data in a space of limited dimensions. Specifically, multidimensional data are depicted as points in 2D or 3D space. Such visualization converses information about data similarity. The basic metaphor for visualizing similarity is distance. This is in accord with the well-known low of geography: «Everything is related to everything else, but closer things are more closely related» (Tobler, 1970). As an extension, distance between non-geographic data can be considered, in some sense, as proportional to their similarity.

In the rest of the paper, we first present the data categories used in the present analysis, as well as, the spatialization method employed. In the forth and fifth sections, we present the application and the results of the analysis. The conclusions about the methodology developed in this study and the evaluation of the final product are presented in the last section, where some future research plans are also outlined.

DATA

The study area is the insular Greece. Crete and Evia, two very large islands of Greece - much larger than the rest - were excluded from the study due to their difference in growth in all sectors, a fact that would distort the analysis and solution. For the remaining 74 inhabited islands, we collected the necessary spatial and descriptive data. These data refer to the following 4 dimensions:

- INFRASTRUCTURES
- TOURISM
- DEMOGRAPHIC INFORMATION (POPULATION DENSITY)
- DISTANCE FROM ATTRACTION POLES (MAJOR CITIES)

The indication for INFRASTRUCTURES is the combination of Health, Education, Transportation and Administration. Without getting into detail, for each infrastructure type, a weighted index was computed based on quantitative and qualitative information on the existence/presence of certain related components. This is not an easy task for there are various ways of evaluating transportation infrastructure for example, or distance from attraction poles. With respect to the last parameter, the distance to the nearest pole of attraction was considered among the three major cities/ports, i.e., Athens, Thessaloniki or Patras.

Population density varied from 3 to 357 inhabitants per Km^2. Tourism growth values were estimated by combining the number of visitors, the facilities available and the island size. The values of all these data were converted to become dimensionless and normalized (dividing by the largest value) to a 100-scale in order to become comparable. An indicative portion of the results shows in table 1.

#	Island	Population Density	Standardised distance	Infra-structures	Tourism growth
1	ANGISTRI	50	5	4	29
2	AEGINA	90	3	26	43
3	ANTIKYTHIRA	10	41	0	14
4	POROS	50	7	11	29
5	SPETSES	90	14	7	57
6	HYDRA	30	11	11	57
7	KYTHIRA	10	36	24	57
8	SALAMINA	100	0	19	14
9	RHODES	50	77	89	100
10	KOS	70	59	54	86

Table 1: Final values of variables for each island.

METHODOLOGY

As mentioned above, the selected method for visualizing our data is Multi-Dimensional Scaling (MDS) (Kruskal, 1978). This method is based on multivariate statistics with the intention of reducing dimensionality of the data, in our case, of the socioeconomic data that characterise the Greek islands. This MDS algorithm depends on a similarity measure. More specifically, the goal is to represent multivariate data into a projection space of two or three dimensions. The relative position of two elements in the projection space expresses their similarity. The farther the two elements are, in the projection space, the more dissimilar they are.

If we examine *n* data elements that have *m* dimensions, we form a matrix of our data P nxm. Then we create a Similarity matrix (or a dissimilarity matrix, depending on the data characteristics) S nxn, in which each element S (i, j) corresponds to the results of comparing (under a certain context) the data elements i and j. Therefore, in the matrix at position i,j we find a value that corresponds to the degree elements i and j are similar. It is obvious that:

- The elements S (i, i) for every i=1, 2,…, n are equal to 116 since every element is identical to itself.
- The similarity matrix S is symmetric.

The process of determining the similarity matrix presupposes a way of measuring the similarity of data. There are several possibilities; we mention here only the distance-type measure and the matching-type measure. The distance-type measure is based on the Euclidean Distance equation. There are also other ways of counting distance like city block or Minkowski distance. The use of distance-type measure is usually applied to quantitative data, only when they refer to the same units and they are normalized. The normalization of the dataset equalizes their variability; in this way, we secure the equipollent contribution of all the dimensions (characteristics) of the dataset when determining similarity. In addition, we avoid putting in the shade those characteristics that exhibit low variability and favour the characteristics of high variability so that they contribute exclusively in the determination of similarity. The matching-type measure is applied when we want to compare data whose characteristics are ordinal.

[16] Reversely, for the dissimilarity matrix this value is equal to 0.

Further analysis of the matching-type measure is out of the scope of this paper since we have not used it for our dataset.

After defining the way of accounting for the data similarity, a multivariate dataset can be portrayed as a set of points in a two dimensional and three-dimensional space. This consists in projecting the points in a way so that the distances among them approximate the relevant values of the Similarity matrix. This is feasible when the Stress function becomes minimal. The stress function is defined as the degree of approximation of the points' distances to the values of the similarity matrix. Kruskal (1978), gives the following definition:

$$\text{Stress} = \sqrt{(\Sigma\Sigma(f(x_{ij}) - d_{ij})^2 / \Sigma\Sigma d_{ij}^2)},$$

where, $f(x_{ij})$, is a function that depends on the type of our dataset (quantitative or ordinal).

The mathematical explanation when Stress does not reach the zero values is one - the representation space has not enough dimensions. However, it is not necessary to achieve a zero value of the stress function for the spatialization, to be feasible in a representation space of specific dimensions. We can afford a small distortion of the result, and usually, we follow the rule: if the stress is less than 0.1 the approximation is very good and distortion negligible, while if it is more than 0.15 the solution is unacceptable.

APPLICATION

In the current approach, a dissimilarity matrix of dimensions 74x74 (corresponding to the 74 Greek islands) was formed. In this case, the similarity measure used the following distance-type measure between two island i, j:

$$S = \sqrt{(P1_i - P1_j)^2 + (P2_i - P2_j)^2 + (P3_i - P3_j)^2 + (P4_i - P4_j)^2}$$

where, P1 to P2 are the values of the four parameters (population density, distance, infrastructure, and tourism growth).

Therefore, from table 1, we can create the dissimilarity matrix; table 2 shows an excerpt. The software used for the application is PERMAP version 9.7[17]. The spatialization of the dataset using MDS gave a solution where the Stress function was equal to 0.0011447. As mentioned above, this value provides a high approximation with a slight distortion of the final position of the given socioeconomic units regarding their initial values, so the solution is acceptable.

The output is a two dimensional map where socioeconomic units that are close share common characteristics (the four variables that we chose to examine), while the opposite is true for distant ones. This principle can give us an account of the similarity of Greek islands under the examined context. Figure 1 shows the output of the method for our dataset.

[17] The homepage of PERMAP http://www.ucs.louisiana.edu/~rbh8900/ (new versions of the software are often). Details of the software are not given herein, but for further information, the reader can refer to (Heady & Lucas, 1997).

	ANGISTRI	AEGINA	ANTIKYTHIRA	POROS	SPETSES	HYDRA	KYTHIRA	SALAMINA	RHODES	KOS
ANGISTRI	0.00									
AEGINA	47.96	0.00								
ANTIKYTHIRA	55.66	96.34	0.00							
POROS	7.82	45.17	55.09	0.00						
SPETSES	50.13	25.71	94.88	49.74	0.00					
HYDRA	36.10	63.84	57.17	35.02	60.21	0.00				
KYTHIRA	61.71	87.65	49.36	58.40	84.66	34.97	0.00			
SALAMINA	54.30	31.36	100.54	53.06	47.54	83.10	106.22	0.00		
RHODES	132.52	119.40	134.80	126.49	118.52	112.72	96.48	144.05	0.00	
KOS	95.40	78.19	109.20	90.28	73.42	81.15	76.29	103.66	46.52	0.00

Table 2: Excerpt of the dissimilarity matrix for the data of table 1.

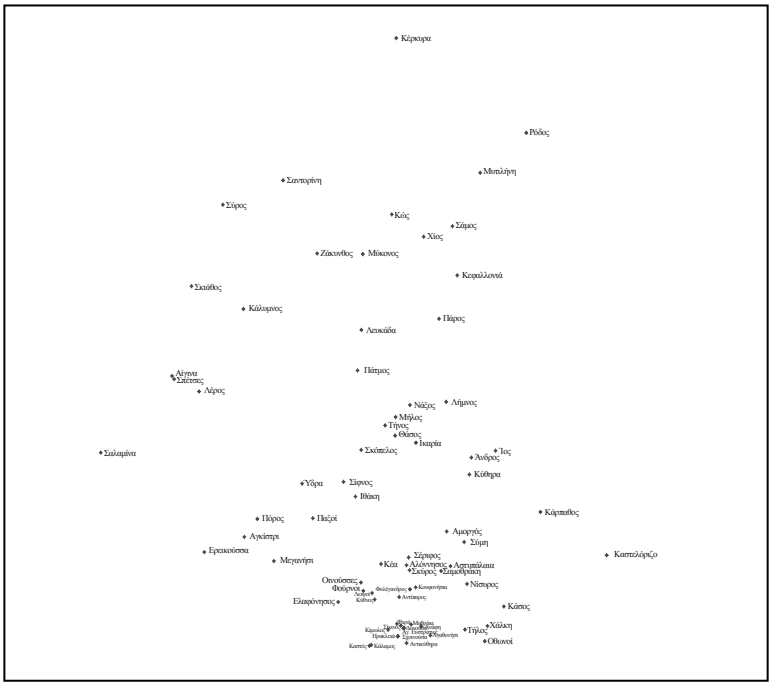
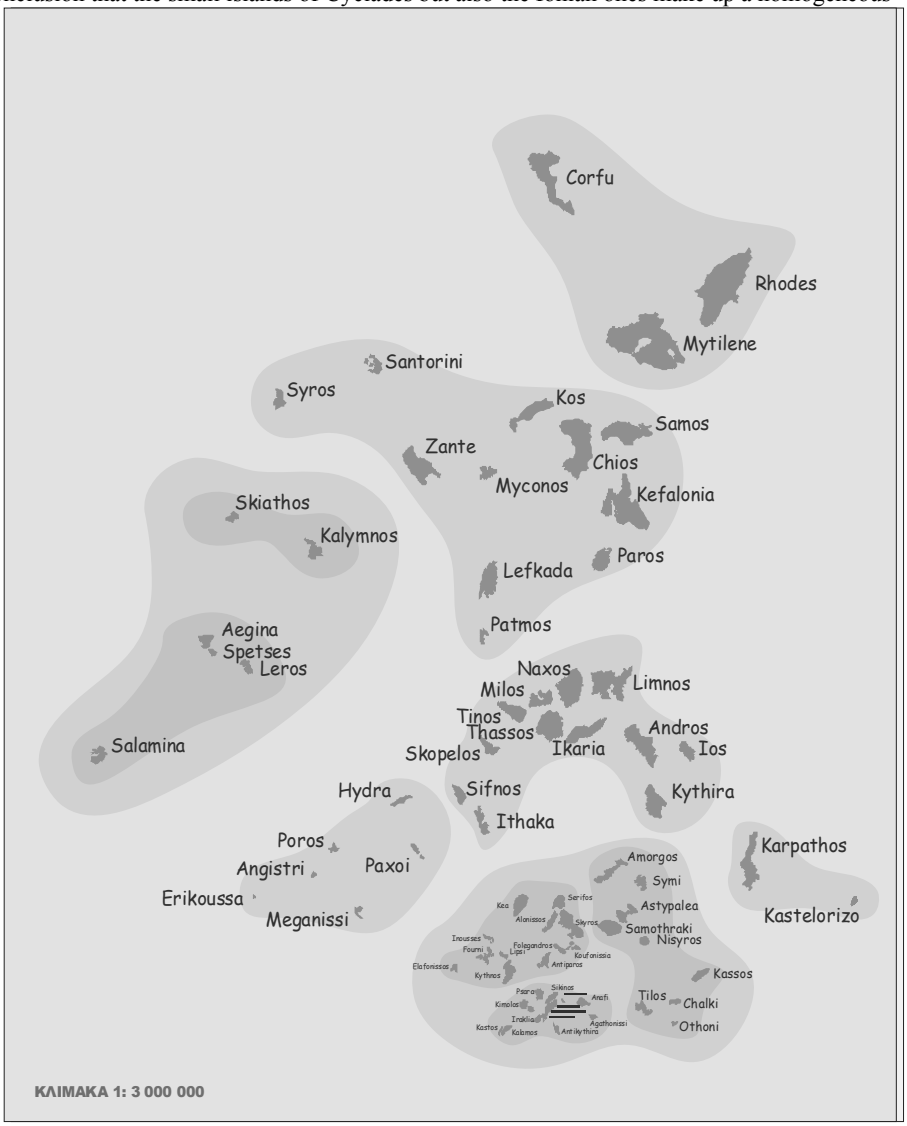


Figure 1: The island positions as resulted from MDS.

RESULTS

The position of SEUs on the produced map is not real in the geographic sense, but depends on the variables/values used for the cartogram. The cartogram illustrates insular groupings, different from the known ones, which possess common properties with respect to their socioeconomic variables used. Based on the use of the cartographic metaphor, map 1 shows the resulting insular groups. Their analysis leads to the conclusion that the small islands of Cyclades but also the Ionian ones make up a homogeneous

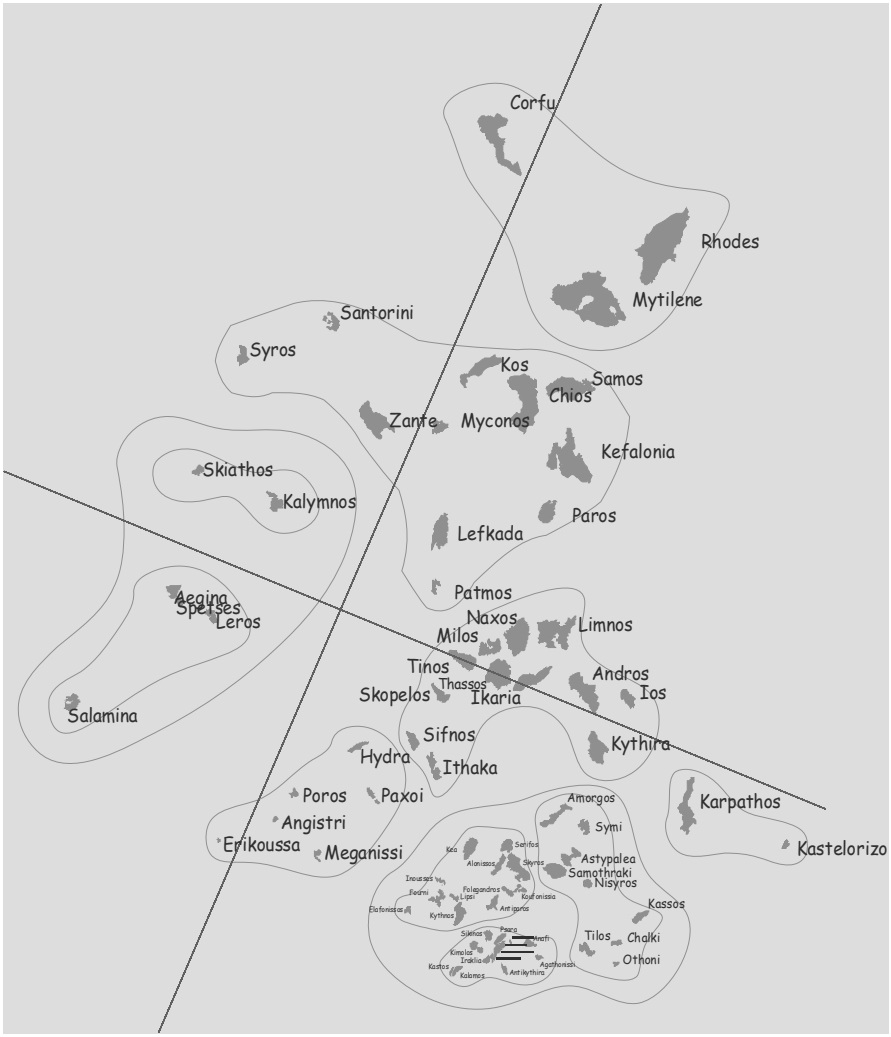


Map 1: *Insular groupings of the cartogram.*

agglomeration comprising of islands with minimal to no infrastructures, very low tourism activity, being relatively far from attraction poles without air-flight connections. The distances among these islands are very short, which practically translates that this group is much more homogeneous that the others and its members are not significantly differentiated in any of the variables considered.

Besides the interpretation of the results with respect to (a) similarity and (b) group formation, we can also identify axes so that SEUs projected on these are differentiated in an interpretable manner. The two axes of the example presented show on map 2, expressing the following:

- The almost horizontal axis (from up left to down right) expresses mainly population density variation with a slight influence of the standardized distance from an attraction pole.
- The almost vertical axis (from up right to down left) expresses the combined influence of infrastructures and tourism growth.



Map 2: *Interpretation of resulting map.*

Thus, we pinpoint in the upper left quarter of the map the islands exhibiting a satisfactory level of infrastructures and higher population density. Here are island such as Syros and Skiathos with high population density and satisfactory infrastructures. In the upper right area, we locate islands with satisfactory infrastructures but lower population density. Also, in the upper part, close to the vertical axis we see a group formation of the three larger islands (Corfou, Rhodes and Lesbos) with higher level

infrastructures but average population density. The islands, because of their distance from attraction poles, exhibit greater independence and have the potential to act as future poles. In the lower left quarter of the map, there are islands such as Salamina, Aegina and Leros, which have higher population density but their infrastructures are below average. Finally, at the lower right area, there are the majority of the small islands of the territory with poor infrastructures, small population and far from the attraction poles.

CONCLUSIONS AND FUTURE RESEARCH

The method of spatialization facilitates:

- The exposal of semantic relations.
- The formation of clusters of entities with common characteristics, as well as revealing their hidden structure.
- The visualization of very large amounts of data using less information with reduced dimensions.

The spatialization method used here was from the technical point of view simple to implement, yet the value of the findings is important. The formation of homogeneous clusters from the initial data helps study various socioeconomic phenomena. Furthermore, the process of identifying the semantic similarity of SEUs, such the one used in the present study, can create similarity zones in space and expose not apparent relations or associations.

Our future research focuses on the

- application of the proposed methodology in other SEU types and problems,
- formal ontological definition of SEUs based on their essential properties,
- use of other parameters in the similarity measure,
- investigation of other spatialization techniques such as SOM (Self-organizing Maps), and finally,
- comparison of the results produced by different techniques.

REFERENCES

Frank A. Raper J. & Cheylan J.P. (Eds.), 2001, Life and Motion of Socio-Economic Units, London: Taylor & Francis, GISDATA Series 8, pp. 353.

Heady R.B. & Lucas J.L., PERMAP: An interactive program for making perceptual maps. In Behavior Research Methods, Instruments & Computers, Vol. 29, 450-455, 1997.

Kavouras, M., Understanding and Modelling Spatial Change". In Frank A., Raper J. & Cheylan J.P. (eds.). Life and Motion of Socio-Economic Units, London: Taylor & Francis, GISDATA Series 8, Chapter 4:49-59, 2001.

Kavouras, M., A Unified Framework for Semantic Integration, International Workshop on Next Generation Geospatial Information, October 19-21, Cambridge (Boston), MA, USA, 2003. http://ontogeo.ntua.gr/publications/kavouras-boston-extended%20abstract.pdf

Kruskal, J. B. and Wish, M., 1978, Multidimensional Scaling, Sage University Paper series on Quantitative Applications in the Social Sciences, number 07-011. Sage Publications, Newbury Park, CA.

Kuhn, W. & Blumenthal B., 1996, Spatialization: Spatial Metaphors for User Interfaces. Geoinfo Series, 8, Department of Geoinformation, Technical University of Vienna, Vienna, Andrew U. Frank (series ed.).

PERMAP, PERceptual MAPping Software, http://www.ucs.louisiana.edu/~rbh8900/

Skupin, A. & Buttenfield, B. P., Spatial Metaphors for Display of Information Spaces. Proceedings, AUTO-CARTO 13,116-125, Seattle, Washington, Apr. 7-10, 1997.

Smith, B., Objects and their Environments: From Aristotle to ecological Ontology. In Frank A. Raper J. & Cheylan J.P. (eds.). Life and Motion of Socio-Economic Units, Chapter 4. London: Taylor & Francis, GISDATA Series 8, Chapter 6: 79-97, 2001.

Tobler, W. A Computer Model Simulating Urban Growth in the Detroit Region. Economic Geography, 46(2): 234-240, 1970.