Space Models as a tool for Sustainability Development

Adriano Moreira*, Maribel Yasmina Santos, Sofia Carneiro
Department of Information Systems
University of Minho, Guimarães, Portugal
e–mail: {adriano, maribel, sofia}@dsi.uminho.pt

ABSTRACT

Space Models are new space geometries that are created to emphasize the particularities of the geo-referenced data analysed. A Space Model integrates groups of regions that present similar behaviour attending to a specific characteristic. Each group represents a cluster aggregating regions that are similar regarding to the analysed characteristic, and regions in different clusters are as dissimilar as possible.

This paper proposes the creation of Space Models, through the STICH (*Space Models Identification Through Hierarchical Clustering*) algorithm, as an alternative approach for data visualization, where the geometry of the maps is created from the data itself. Space Models are new space geometries that are created to emphasize the particularities of the analysed data, and integrate groups of regions that present similar behaviour attending to a specific characteristic.

The achieved results are illustrated through a set of examples that are compared with conventional representations, showing that Space Models provide real added-value over conventional approaches, namely by facilitating the identification of peculiarities in the data.

INTRODUCTION

Maps are used in many application areas to support the visualization of geo-referenced data. However, the geometry explicit in each map is defined for or with a specific purpose, being sometimes later used in applications for which it was not conceived. This is often the case with maps representing administrative subdivisions of the geographic space. Administrative subdivisions inside a country – parishes, municipalities, and districts, for example – are defined following specific criteria and an accumulated number of changes along the years, and not following a natural division of the space. They are later used as a reference for data analysis, even when they are not the most suitable resource for that. As an example, consider the 'Emissions of greenhouse gases' indicator distribution across the 15 European countries[1] represented in Figure 1, and collected at NUTS II level.

In the map shown in Figure 1, the conventional method for data visualization was used. Since the data is available at NUTS II level, the obvious choice for visualization is to use a map with the NUTS II geometry, classify the data values within a given number of classes (6 in this case), and paint the NUTS II regions accordingly to the class of the data value associated to each region. Geographic Information System (GIS) tools, provide means to build these maps automatically, including the classification of the data values in different classes using a few different methods (Equal Interval, Quantile, Natural Breaks (Jenks), Standard Deviation).

* Corresponding author
[1] These are the 15 European countries that integrated the European Union until April 2004.

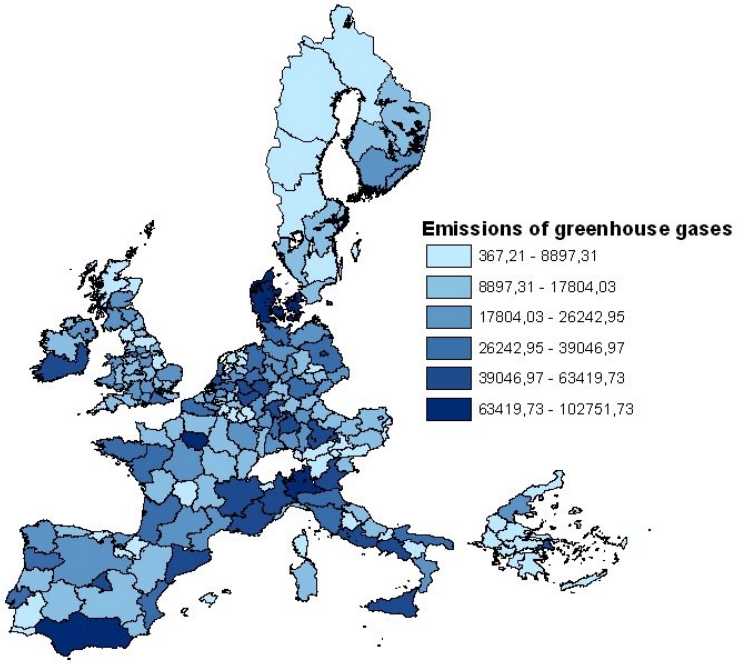

Figure 1. 'Emissions of greenhouse gases' indicator across Europe

The first problem with this approach is that the NUTS II regions are of very different sizes: large regions in countries like Spain and Sweden, and very small regions in countries like Germany and Belgium. The second problem is the choice for the number of classes used to group the data values (and consequently the number of different colours used to paint the regions). In this example, regions with data values between 367,20 and 8897,31 are classified within the same class (more than an order of magnitude between the lower and upper limits). Therefore, the observation of the map does not provide any information about what regions are close to the lower limit (performing better) or close to the upper limit. To observe these differences, a larger number of classes must be used, probably turning the map very confusing and loosing the whole picture. Finally, even with a considerable small number of classes (6 in this example), this map does not clearly shows the best and worst regions in terms of the indicator under analysis.

To overcome these limitations this work describes an approach for the identification of Space Models - new space geometries that are created to emphasize the particularities of geo-referenced data. Each Space Model integrates groups of regions that present similar behaviour attending to a specific characteristic. Each group represents a cluster aggregating regions that are similar regarding to the analysed characteristic, and regions in different clusters are as dissimilar as possible.

In order to identify Space Models the hierarchical clustering algorithm STICH – *Space Models Identification Through Hierarchical Clustering* – was developed, allowing the identification of sets of regions with similar behaviour.
Space Models are of great importance for the analysis of indicators (environmental or social, for instance) associated to regions and to understand the main differences between these regions. This is the objective of the EPSILON (*Environmental Policy via Sustainability*

Indicators on a European-wide NUTS III Level) project[2], in which the collected sustainability indicators must be analysed in order to identify regions with similar behaviour, and regions that exhibit different trends in data, in order to achieve a sustainable development across Europe. This project contributes to the better understanding of the European Environmental Quality and Quality of Life, by delivering a tool aimed to generate environmental sustainability indices at NUTS-III level. The work described in this paper was developed as part of the EPSILON project and is being integrated into that tool as an added-value functionality for data analysis.

This paper is organised as follows. In the next section, the fundamentals of Space Models are introduced and the algorithm used for their creation is described. Then, the application of Space Models in the analysis of sustainability indicators is presented through a set of examples. These examples compare the results obtained through the use of Space Models with the results that are achieved by conventional visualization approaches based on administrative maps. The following section is dedicated to the description of the Space Models Tool implementation and its integration into the EPSILON Tool. Finally, we present some concluding remarks.

SPACE MODELS

Human beings mentally use space models to simplify reality and to perform spatial reasoning more effectively. When we look to the birth rate of the several districts of a country and try to analyse the available data our first thought is to group districts with similar birth rate. This procedure allows us the creation of a space model.

Space Models are new geometries of the space that are created to emphasize particularities of the analysed data. A Space Model integrates groups of regions that present similar behaviour attending to a specific characteristic explicit in the data. Each group represents a cluster aggregating regions that are similar regarding to the analysed characteristic. Regions in different clusters must be as dissimilar as possible. Besides the role of Space Models in data analysis, their creation also allow their use as a mean for data visualization when the available maps are not suitable or do not fit specific purposes.

Concepts associated to Space Models

The creation of Space Models [1, 2] through clustering techniques [3, 4, 5] allows the identification of groups of regions that emerge from the analysed data and not groups of regions that are imposed by either analysis techniques or human constraints.

The process of creation of Space Models presented in this paper is completely autonomous and automatic (its algorithm is resumed in the next subsection) and assumes several assumptions that guarantee the quality of the obtained Space Models. The principles are:

- Space Models must be created from the data values available for analysis, and no constraints can be imposed for their identification;
- The created Space Models must be the same, independently of the order by which the available data is processed in the clustering process;
- Space Models can include clusters of different shapes and sizes;
- Space Models must be independent of specific domain knowledge, like the specification of the final number of clusters.

[2] A project founded by the European Commission through the IST program (contract IST-2001-32389).

The identification of Space Models with the STICH algorithm

The STICH algorithm [1, 2] is based on an iterative process in which the several regions are grouped according to their similarity with respect to a specific characteristic. In each step of this process the clusters are formed joining the *k* most similar regions, being *k* a value identified from the data and dependent of each cluster.

This iterative process starts with all regions in different clusters, i.e. each region constitutes a cluster, and ends with all regions grouped into the same cluster. As a hierarchical clustering algorithm it allows the identification of several Space Models, one at each iteration of the clustering process.

The formal specification of STICH can be found in [1, 2]. Figure 2 depicts a simple example in which the iterative process of STICH and the clusters identification is exemplified. In this example 5 steps were needed to aggregate all data in one final cluster. At each step different data aggregations are obtained based on the calculation of the Similarity Matrix of the regions. This matrix contains all the distances existing between each pair of regions. After this calculation STICH identifies the *k-nearest-neighbours* of each region, being *k* a value that can be different from one region to another (for more details see [1, 2]). The average of the *k-nearest-neighbours* of each region is calculated and the region is afterwards assigned to the cluster in which it appears with the minimum *k-nearest-neighbour* average. At this point new clusters are obtained and new centroids are calculated for them.

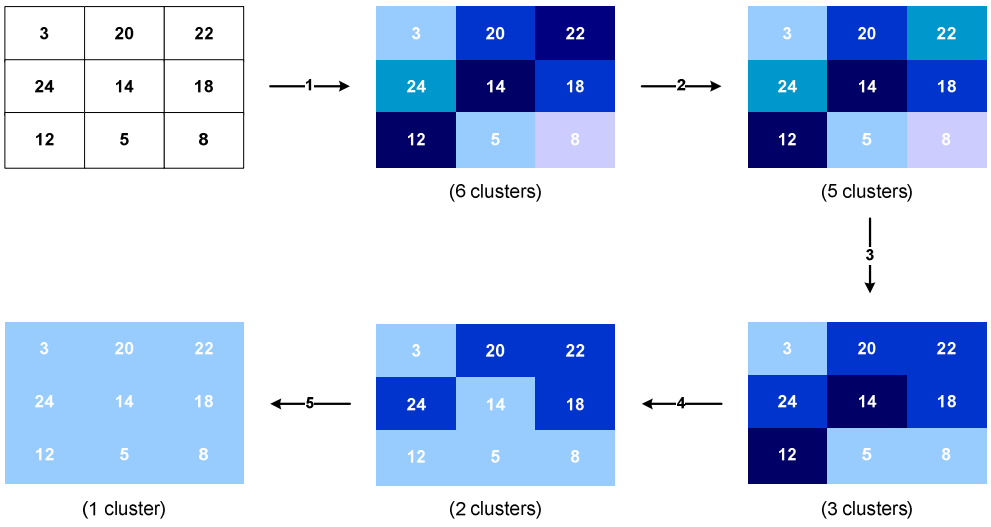

Figure 2. The clustering process of STICH

SPACE MODELS IN SUSTAINABILITY DEVELOPMENT

The work described in this paper, including the STICH algorithm, was developed within the framework of the EPSILON project with the aim of turning the analysis and visualization of sustainability indicators easier and more powerful. Although Space Models can be used in many application areas, in this section we present some examples within the area of sustainability indicators analysis. With these examples, some of the benefits of Space Models are demonstrated and compared with conventional approaches based on administrative maps. However, note that conventional techniques continue to be useful in many scenarios, and that Space Models approaches are, in many situations, a complement to these techniques.

Space Models created by using the STICH algorithm have some characteristics that are illustrated in the following examples, namely:

- They facilitate the identification of peculiarities in the analysed data by highlighting regions where the geo-reference data is considerably different from the average. This allows the easy identification of regions with particular problems or regions that perform much better than others for a given indicator.
- The same data can be visualized at different levels of aggregation, corresponding to different Space Models, where the number of clusters and their limits are extracted from the data itself. These different levels of aggregation allow the analysis of the same data from a very detailed (although maybe confusing) level to a very broad level (less detail) without hiding the most different regions.
- A Space Model created from a given indicator can be used to assist on the analysis of another indicator.

To illustrate these characteristics, some data was extracted from the EPSILON database and used to create example Space Models. The selected datasets are described next.

The data available at the EPSILON project

One of the outputs of the EPSILON project is a database that stores a large amount of information related to the sustainable development assessment across the 15 European countries. The data in this database reflects the sustainability model developed by the project and is organized accordingly to a 4x4x4 structure: four pillars (Economic, Environmental, Institutional and Social), four themes per pillar and four sub-themes per theme. Each sub-theme is supported by a number of indicators and sub-indicators from where the corresponding value is calculated. Additionally, the database includes a number of quality assessment parameters that provide information about the quality of the data. This structure is replicated at four levels of detail (NUTS 0 to NUTS III) with different amounts of data for each level. For more information on the sustainability model developed by the project see [6].

The data used in the examples described below was extracted from the EPSILON database. The first selected dataset (dataset 1) is the 'Soil Toxicity Index' indicator distribution across the 15 European countries collected at NUTS I level. This dataset has a total number of 74 records. The second dataset (dataset 2) is the 'Groundwater Quality' indicator distribution across the 15 European countries collected at NUTS I level. This indicator represents the quality of the groundwater by means of the level of hazardous substances. This dataset has a total number of 74 records. The third example uses two datasets (datasets 3 and 4): one with the indicator 'Average number of units in all economic activities' (activities such as, mining, quarrying, manufacturing, electricity, gas, transport, restaurants, etc), and the other with the indicator 'Emissions of greenhouse gases'. These two dataset are available at NUTS II level and have 204 records each.

The Space Models obtained through the STICH algorithm

This section describes some examples of Space Models obtained with the STICH algorithm for the datasets identified above.

Figure 3 presents a Space Model obtained by the STICH algorithm after grouping the dataset 1 data values in 4 clusters. Note that the number of clusters is not an initial parameter of the STICH algorithm – the Space Model with 4 clusters was chosen among all the Space Models created by the STICH algorithm (Space Model obtained at the 19^{th} iteration).

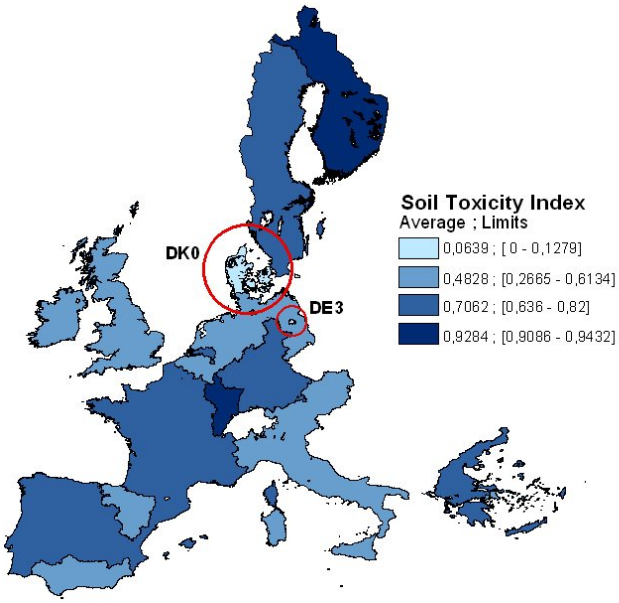

Figure 3. Space Model obtained by STICH

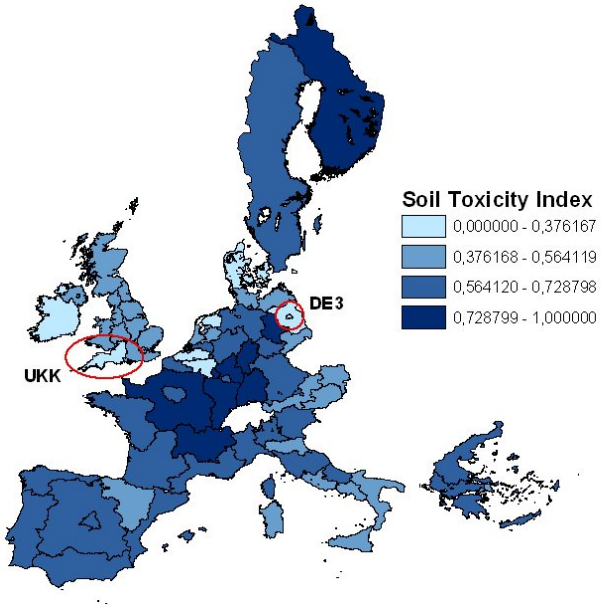

Figure 4. The same data using Natural Breaks classification of the map in 4 classes

This example shows that the Space Model obtained by the STICH algorithm highlight a region (the one formed by two NUTS I regions, DK0 and DE3, marked with two circles) where the average value of the 'Soil Toxicity Index' is much lower than the same value in all other regions. This allows the easy and immediate identification of the areas that perform much better than all others in terms of this indicator. In Figure 4, a conventional map, with the data values classified in 4 classes, is presented. With this classification, some important information is lost. For example, note that the region DE3 belonging to the lower class in Figure 3, is now (in Figure 4) aggregated to other regions for which the performance in terms

of the analysed indicator is much worse. In this particular case, regions like DE3 are displayed with the same performance as UKK. However, the first one has a value of 0 while the last has a value of 0,376167. In summary, the map shown in Figure 3 allows a more easy identification of the most relevant (best and worst) cases, therefore overcoming the second problem identified in the map shown in Figure 1.

The next two figures (Figures 5 and 6) shows the same data (dataset 2) at different levels of aggregation, that is, different Space Models with different number of clusters (11 and 3 clusters respectively) and different limits for each cluster. These two Space Models correspond to the output of the clustering process at the end of two different iterations. This example illustrates the versatility of the Space Models approach.

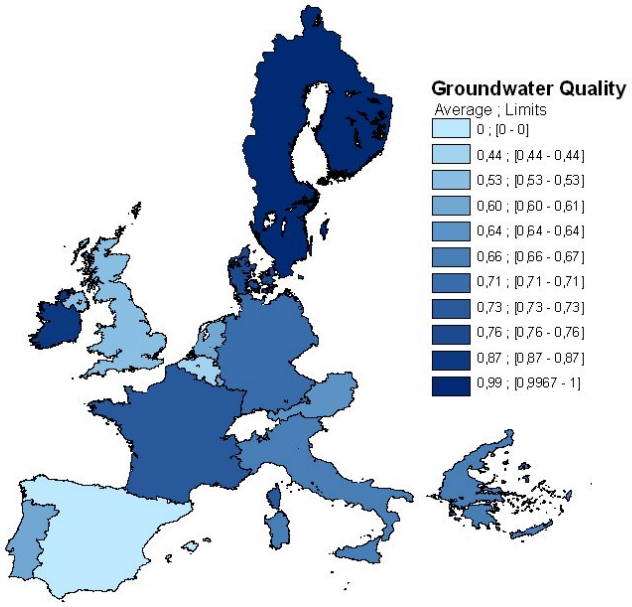

Figure 5. Space Model obtained by STICH at the 2nd iteration

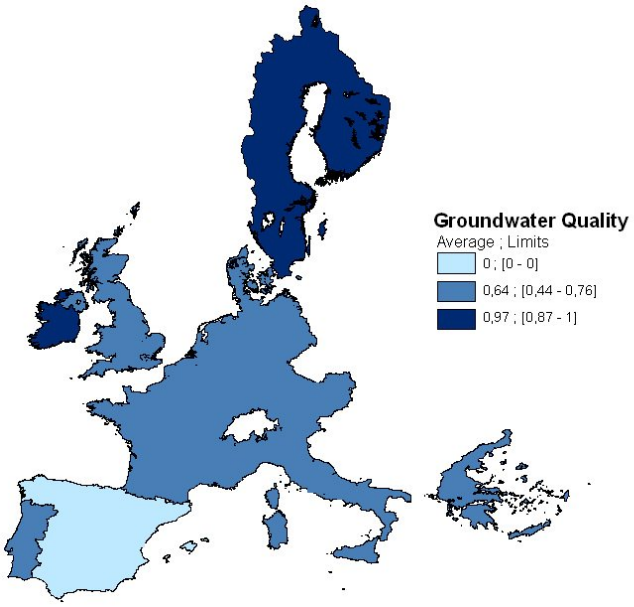

Figure 6. Space Model obtained by STICH at the 9th iteration

By choosing the appropriate iteration of the clustering process, the user is given the opportunity to analyse the same data at different levels: from a very detailed analysis to a broad view of the data. Moreover, with this approach, the regions that are more different than the others are not hidden by the choice of a small number of clusters (broader view in Figure 6). Note that, in this example, Spain stills highlighted as the best region, even when only three clusters are considered.

The next example shows how a Space Model created from a given indicator – the 'Emissions of greenhouse gases' indicator at NUTS II level (shown in Figure 7) – can be used to assist on the analysis of another indicator – the 'Average number of units in all economic activities' indicator also at NUTS II level (shown in Figure 8).

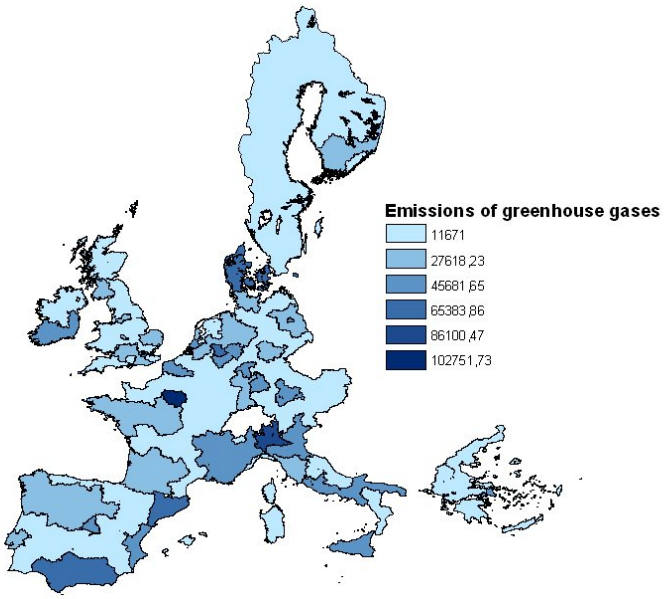

Figure 7. Space Model obtained by STICH for the indicator 'Emissions of greenhouse gases'.

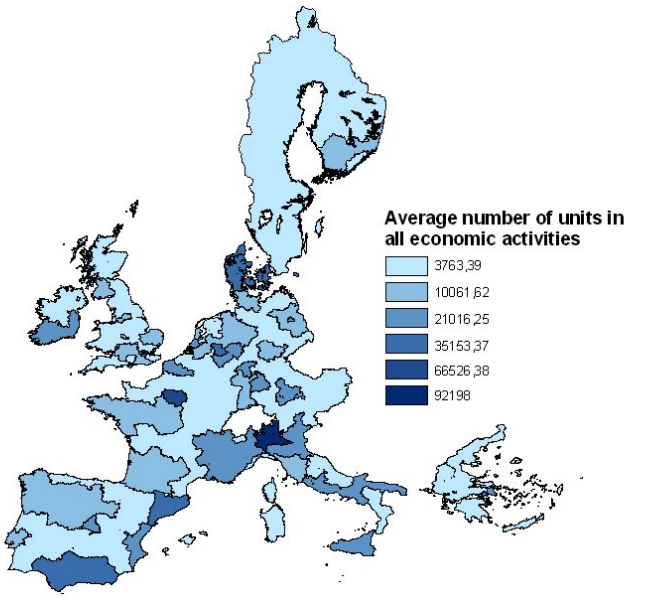

Figure 8. Analysis of the indicator 'Average number of units in all economic activities' in the Space Model obtained for the indicator 'Emissions of greenhouse gases'.

In this example, the 'Emissions of greenhouse gases' indicator was used to create a Space Model. Then, the data related to the 'Average number of units in all economic activities' indicator was grouped exactly the same way (the same original regions) as the STICH algorithm did for the first indicator. Finally, this aggregated data was shown on top of the created Space Model.

The indicator analysed in Figure 8 is related to the average units of all activities such as, mining, quarrying, manufacturing, electricity, gas, transport, restaurants, etc. It is therefore expected that the regions with more activities be more favourable to the emissions of greenhouse gases. Actually, as seen in Figure 8, the darker regions (regions with high number of activities) are closely related to the darker regions in Figure 7[3] (regions with high emissions of greenhouse gases). In summary, this usage of the STICH approach facilitates the cross analysis of related indicators.

TECHNOLOGICAL CHARACTERIZATION

The example results described in the previous section were obtained from an implementation of the STICH algorithm called Space Models Tool – SM-Tool. The SM-Tool is a software module implemented in Visual Basic for Applications (VBA), the language embedded in the ESRI ArcView 8.2 (the Geographic Information System adopted by the project). The implementation of the STICH algorithm included in this module was developed as a set of DLL's (Dynamic Link Libraries) implemented in Visual Basic 6.0 (VB6). The result is a functionality that can be easily added to the ArcView working environment to perform data analysis. More detailed information about the implementation of the SM-Tool is available in [7].

The same set of DLL's, implementing the STICH algorithm, are now being integrated into the EPSILON Web Tool being developed by the project. This Web tool, that will be available in the Internet using a simple browser, will provide the possibility for users to perform benchmarking between different countries regarding their sustainable development. The benchmarking is supported by all the data available in the EPSILON database.

CONCLUSION

In this paper, a new technique for geo-referenced analysis and visualization, based on a clustering algorithm, was presented. This technique, available through a Space Models Tool, was developed within the context of the EPSILON project and is being integrated into the EPSILON Web Tool being developed by the project. The benefits of this technique were demonstrated through a set of examples oriented towards the analysis and visualization of sustainability indicators. These examples and other results have shown that the described technique provides real added-value over conventional approaches.

NOMENCLATURE

EPSILON – *Environmental Policy via Sustainability Indicators on a European-wide NUTS III Level*
NUTS – *Nomenclature of Terrestrial Units for Statistics*
STICH – *Space Models Identification Through Hierarchical Clustering*

[3] The Space Model presented in Figure 7 can also be compared with the map in Figure 1 that represents the normal classification of the European Countries in 6 classes for the same dataset.

REFERENCES

1. Santos, Maribel Yasmina, Adriano Moreira and Sofia Carneiro, STICH – Clustering in the identification of Space Models, in John Wang (editor), Encyclopedia of Data Warehousing and Mining, Idea Group Publishing, In Press.
2. Santos, Maribel Yasmina, Adriano Moreira and Sofia Carneiro, STICH – A Hierarchical Clustering Algorithm, *Proceedings of the Workshop "Data Gadgets 2004 – Bringing Up Emerging Solutions for Data Warehousing Systems", IX Conference on Software Engineering and Database*, Málaga, Spain, November 9, 2004.
3. Estivill-Castro, V., & Yang, J., Fast and Robust General Purpose Clustering Algorithms. *Data Mining and Knowledge Discovery*, 8 (2), 2004, 127-150.
4. Grabmeier, J., Techniques of Cluster Algorithms in Data Mining. *Data Mining and Knowledge Discovery*, 6 (4), 2002, 303-360.
5. Jain, A. K., Murty, M. N., & Flynn, P. J., Data Clustering: A Review. *ACM Computing Surveys*, 31 (3), 1999, 264-323.
6. Isabelle Blanc, Damien Friot, Manuele Margni and Olivier Jolliet, How to assess the Environmental State of EU regions with the global concept of sustainability, *18th International Conference Informatics for Environmental Protection – EnviroInfo 2004*, Geneva, October 21-23, 2004.
7. Moreira, Adriano, Maribel Yasmina Santos and Sofia Carneiro, Delivery of the Clustering Model, s/w and documentation, Epsilon project Report D42, December 9, 2004.