Carto-Som – Cartogram creation using self-organizing maps

Roberto HENRIQUES[1], Fernando BAÇÃO[1] and Victor LOBO[1,2]

[1]Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa
Campus de Campolide
1070-312 Lisboa
Portugal

[2]Portuguese Naval Academy
Alfeite
2810-001 ALMADA

The basic idea of a cartogram is to distort a map. This distortion comes from the substitution of area for some other variable (in most examples population. The SOM constitutes a very flexible tool that has been used in many different tasks. In this article we have presented a general method for constructing density-equalizing projections or cartograms, using the basic SOM algorithm, providing a tool for geographic data presentation and analysis.

KEYWORDS

Neural-networks, self-organizing maps, cartograms.

INTRODUCTION

The basic idea of a cartogram is to distort a map. This distortion comes from the substitution of area for some other variable (in most examples population). The objective is to scale each region according to the value it represents for the new variable, while keeping the map recognizable. The use of cartograms precedes the use of computerized maps and computer visualization. The first cartograms were created to show the geographic distribution of population, in the context of human geography [1]. Typically, cartograms are applied to portrait demographic [2], electoral [3] and epidemiological data [4]. Cartograms can be seen as variants of a map. The difference between a map and a cartogram is the variable that defines the size of the regions. In a map this variable is the geographic area of the regions, while in the cartogram any other georeferenced variable may be used.

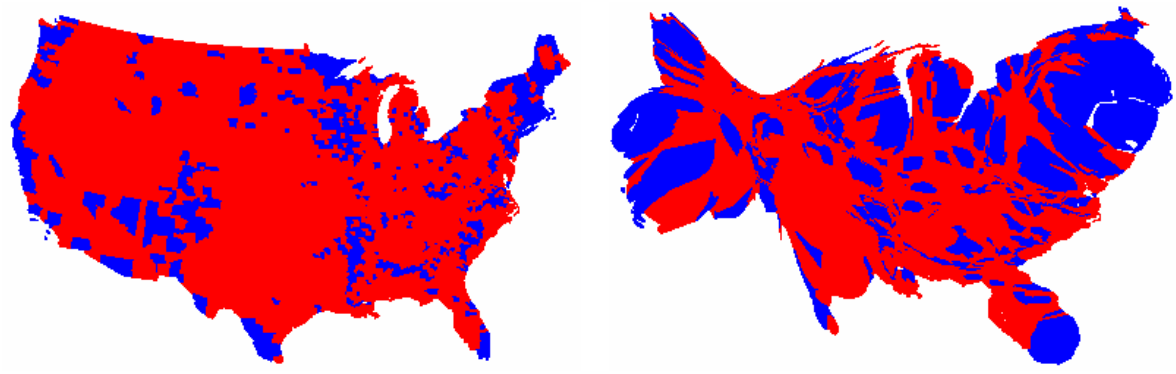
The Self-organizing map (SOM) [5] was introduced in 1981 and is a neural network particularly suited for data clustering and data visualization. The SOM's basic idea is to map high-dimensional data into one or two dimensions, maintaining the most relevant features of the data patterns. The SOMs may be used to extract and illustrate the essential structures in a dataset through a map, usually known as U-matrix, resulting from an unsupervised learning process [6]. The SOM constitutes a very flexible tool that has been used in many different tasks. Here we use the basic SOM algorithm to produce cartograms.

Cartograms

Maps have always been an important part of human life. The oldest known maps in existence today were created around 2300 BC in ancient Babylonia on clay tablets. Today, maps continue to play an important role in human communication, and recent technological developments, such as Geographic Information Systems (GIS), brought the possibility of making maps to a much wider group of users.

A cartogram is a purposely-distorted thematic map that emphasizes the distribution of a variable by changing the area of objects on the map [7]. Area cartograms are deliberate exaggerations of a map according to some external geography–related parameter that convey information about regions through their spatial dimensions[8]. These dimensions have no correspondence to the real world but are a representation of a variable other than area. A cartogram can be seen as a traditional map generalisation. While the map represents objects with the real area, cartograms use any variable instead [9].

In Figure 1a, a map of the counties of the United States in which each state is colored (conventionally) red or blue to depict the distribution of votes between the Republican presidential candidate (George W. Bush) and the Democratic candidate (John F. Kerry) respectively. Analysing Figure 1a, one cannot help to think that Bush has won with a large difference, since red states dominate, covering far more area than the blue ones. This map conveys a wrong impression, since the size of each county does not represent the number of voters. A cartogram using the voters in each county, is much more accurate (Figure 1b), it shows that in terms of population the differences are quite small. In fact, Kerry won in most of the more populated areas, while Bush won a large number of counties although these were essentially small, in terms of population. As pointed out in the Figure 1b, the cartogram will give us a different impression on the results.



a) Choropleth map **b)** Cartogram

Figure 1 – USA 2004 presidential election results. Represented in blue are counties where Kerry wins and in red Bush victories. In [10]

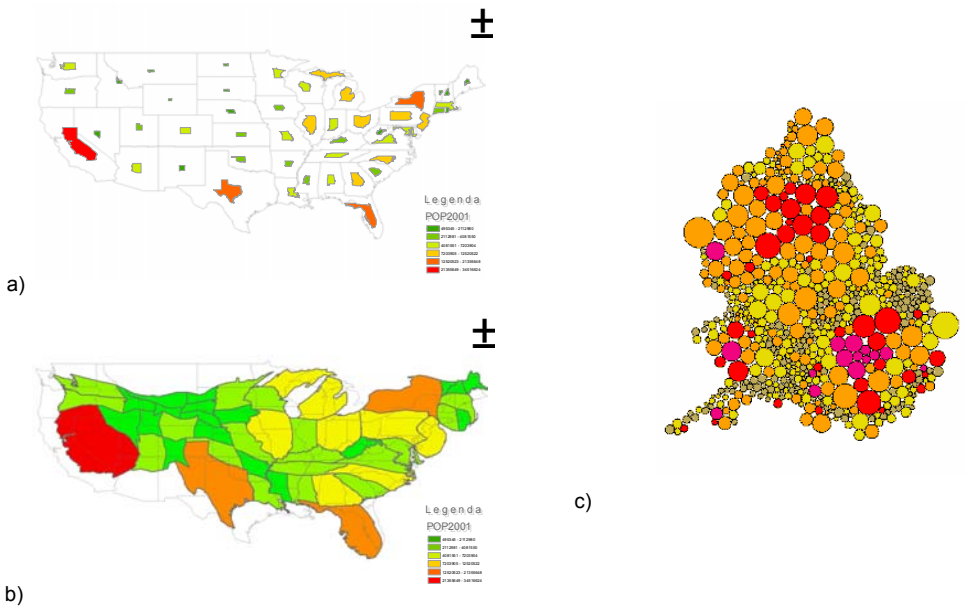
Cartograms can be grouped according to some characteristics like topology, perimeter and the shape. According to [11] cartograms can be divided into three types: non-continuous, contiguous and Dorling

Cartograms.

Non-continuous cartograms are the simplest cartograms to build. These cartograms do not necessarily preserve topology [12]. This means that object connectivity with adjacent objects is not maintained. Each object is allowed to grow or shrink, positioning its borders in a way that the result area represents one variable.

Continuous cartogramas differ from non-continuous cartograms because they preserve topology. In this case topology is maintained (the objects remain connected with each other) usually causing great distortion in shape. These are the most difficult cartograms to obtain since the objects must have the appropriate size to represent the attribute value and should maintain the shape of objects as best as possible, so that the cartogram can be easily interpreted.

Although Dorling cartograms area maintains neither shape, topology nor object centroids they have been classified as a cartogram. Shape expansion/shrink processes are also used to produce new shapes representing one variable. Dorling cartograms differ on the type of the shapes. Instead of preserving the shape of objects new regular shapes are used, normally circles.



a)

b)

c)

Figure 2 – Different types of cartograms. a) Non-continuous cartogram using USA population. b) Continuous cartogram using USA population. c) Dorling cartogram using England population in [11].

Several computer algorithms have already been developed to construct continuous area cartograms [8]. Contiguous Area Cartograms may be obtained using the Constraint-based Method [3], Rubber-map Method [13], Pseudo-cartogram method [2], Medial-Axes-based Cartograms [14], RecMap: Rectangular Map Approximations [9], Diffusion Cartogram [10] and Line Integral Method [4].

Self-organizing maps

Self Organizing Maps (SOM) were first proposed by Tuevo Kohonen in the beginning of the 1980s [5], and stemmed from his work on associative memory and vector quantization. The basic idea of a SOM is to map the data patterns onto a n-dimensional grid of units. That grid forms what is known as the output space, as opposed to the input space that is the original space where the data patterns are. This mapping tries to preserve topological relations, i.e., patterns that are close in the input space will be mapped to units that are close in the output space, and vice-versa. The output space will usually be 2-dimensional, and most of the implementations of SOM use a rectangular grid of units. So as to provide even distances between the units in the output space, hexagonal grids are sometimes used [15]. Single-dimensional SOMs are common (e.g. for solving the travelling salesman problem), and some authors have used 3-dimensional SOMs. Using higher dimensional SOMs, although posing no theoretical obstacle is rare, since it is not possible to easily visualize the output space.

Each unit, being an input layer unit, has as many weights or coefficients as the input patterns, and can thus be regarded as a vector in the same space as the patterns. When we train or use a SOM with a given input pattern, we calculate the distance between that pattern and every unit in the network. We then select the unit that is closest as the winning unit, and say that the pattern is mapped onto that unit. If the SOM has been trained successfully, then patterns that are close in the input space will be mapped to units that are close (or the same) in the output space, and vice-versa. Thus, SOM is “topology preserving” in the sense that (as far as possible) neighbourhoods are preserved through the mapping process.

The basic SOM learning algorithm is explained in [5]. Before training, the units may be initialized randomly. During the first part of training, they are “spread out”, and pulled towards the general area (in the input space) where they will stay. This is usually called the unfolding phase of training. After this phase, the general shape of the network in the input space is defined, and we can then proceed to the fine tuning phase, where we will match the neurons as far as possible to the input patterns, thus decreasing the quantization error.

METHODOLOGY

In this paper we present a new algorithm to create cartograms based on the SOM. Usually, when creating cartograms, areas with a high value on the selected variable “grow”, occupying the geographic space made available by areas with smaller values. This grow/shrink process is focused on creating an equal-density map where high values will be represented by larger areas, and small values by smaller areas. In our proposal the cartogram will be created based on the unit’s movement during the learning process of the SOM.

An example of this process is shown in Figure 3, showing a simplified example using only two regions. The two regions are geographically identical but have distinct values for the variable ***p*.** Assuming variable ***p*** to be population, the region color represents the value of population (dark color corresponds to a higher value of population). In the following step (Figure 3b) we randomly generate points (represented with triangles)

inside each region. The number of points generated is a linear function of the value of population. In Figure 3c a two-dimensional SOM (4 x 4) is initialized. A constant unit density is used to initialize the SOM, contrary to the usual practice, in which the units are randomly initialized in the input space. After the training phase (Figure 3d) the units are labelled (Figure 3e). Figure 3f represents the mapping of each unit to its initial position. A space transformation is then performed based on the unit's position and label resulting on a population cartogram (Figure 3h).

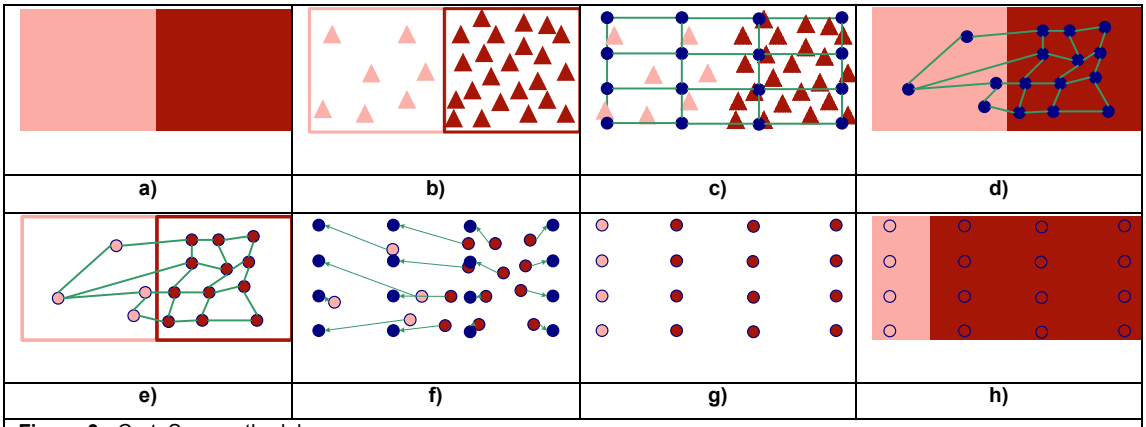


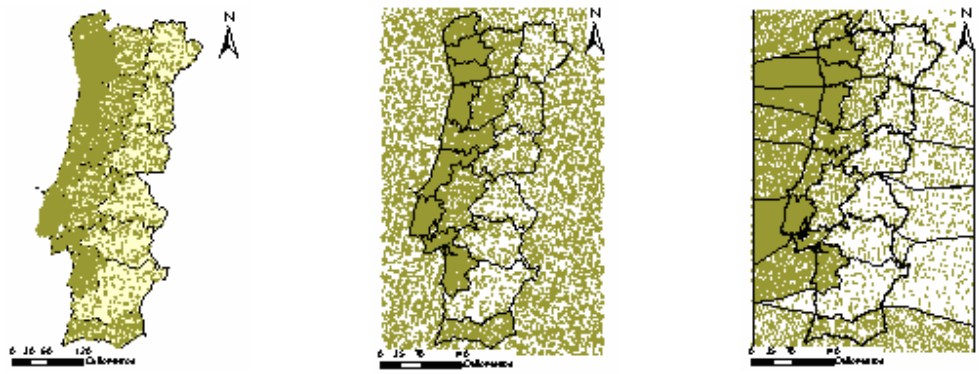
Figure 3 - CartoSom methodology

CARTO-SOM VARIANTS

In the example presented before the rectangles were used to represent the initial regions shape. The SOM is made of an *n*-dimensional grid of units or neurons. In this particular case study two dimensional neural networks is used and therefore the network will have a rectangle shape. This means that, in any complex shape, units will eventually be initially positioned outside the regions. In order to test SOM behaviour in these cases, four different variants were used to create cartograms. The term "ocean" data is assumed to be the point data generated randomly for the area outside geographic boundaries.

1. Standard SOM algorithm without "ocean" data
2. Standard SOM algorithm with median density "ocean" data
3. Standard SOM algorithm with differentiated density "ocean" data
4. Variant of SOM (ocean units are not used in the training process)

In Figure 4 we present point data randomly generated based on the population variable. In the first case (Figure 4a) points were generated only for the regions inside Portugal. On the second and third examples points were generated to the full area extent of Portugal. These examples differ in terms of the point density outside Portugal. On both cases points were generated inside Portugal based on the population variable. On Figure 4b an average density for Portugal was used while on the other case a division of outside space was made (based on Thiessen polygons) and each outside Portugal region was populated with points according to their costal region density.



a) Point data proportional to population inside Portugal's regions

b) Inside Portugal: point data proportional to population regions
Outside Portugal: point data proportional to the Portugal's regions median density

c) Inside Portugal : point data proportional to population regions
Outside Portugal: point data proportional to nearest costal region density

Figure 4 – Point data created on the 2001 Portuguese population.

In the first three variants of Carto-SOM, "quasi-standard" SOM algorithm was used. The SOM algorithm used only differs from the original one in the initialization phase. Usually, units are randomly initialized but in this methodology units were equally spread along the data space.

On the fourth variant the first point dataset (point data only inside Portugal) was used. Some changes were made on the original SOM algorithm. These changes are related to the initialization of the units (units were not randomly initialized but equally spread along the input space) [16] and before the training phase units positioned outside the original regions were excluded from the network. The algorithm is explained in bellow:

Let

X be the set of n training patterns $\boldsymbol{x_1, x_2}$

Y be the set of m *calibrating* patterns y_1, y_2 with labels S (for sea) and L (for land)

W be a p x q grid of units $\boldsymbol{w_{ij}}$ where i and j are their coordinates on that grid

a be the learning rate, assuming values in]0,1[, initialized to a given initial learning rate

r be the radius of the neighbourhood function $\boldsymbol{h(w_{ij}, w_{mn}, r)}$, initialized to a given initial radius

I be the list of $\boldsymbol{w_{ij} \in}$ W which when calibrated with Y have label L

1 Initialize W based on Geo-Som criteria (unit will be uniformly distributed in a geographic space)

2 Label the units of W with the calibrating patterns Y

3 Select for I all unit of W which have label L (select for I all units which are in land)

4 For k=1 to n

5 For all $\boldsymbol{w_{ij} \in I}$, calculate $d_{ij} = || \boldsymbol{x_k} - \boldsymbol{w_{ij}} ||$

6 Select the unit that minimizes d_{ij} as the winner $\boldsymbol{w_{winner}}$

7 Update each unit $\boldsymbol{w_{ij} \in I}$: $w_{ij} = w_{ij} + a\ h(w_{winner}, w_{ij}, r)\ ||\ x_k - w_{ij}\ ||$

8 Decrease the value of a and r

9 Until a reaches 0

In Figure 5 the output space after the training of the SOM is represented. As we can see, the units (red dots) that are geographically outside Portugal's boundaries were ignored in the training process.

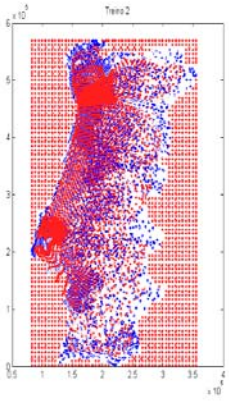


Figure 5 – Fourth variant output space. Units (red dots) and population dots (blue dots)

CARTO-SOM IMPLEMENTATION

The Carto-SOM methodology was implemented using Portugal and USA 2001 population data. Population variable was aggregated in Portuguese administrative regions (Distritos) for Portugal and in states for the USA. For each dataset the Carto-SOM four variants were applied using different SOM parameters. A table showing the used SOM training parameters (Table 1) is presented:

Variants	Test nº	x units	y units	iter1	iter2	n1	n2	L1	L2
1	1	50	25	50	75	20	8	0.5	0.1
	2	75	50	50	75	20	8	0.5	0.1
	3	50	25	50	75	25	10	0.5	0.1
	4	75	50	50	75	25	10	0.5	0.1
	5	50	25	50	75	25	10	0.7	0.2
	6	75	50	50	75	25	10	0.5	0.1
2	1	50	25	50	75	20	8	0.5	0.1
	2	75	50	50	75	20	8	0.5	0.1
	3	50	25	50	75	25	10	0.5	0.1
	4	75	50	50	75	25	10	0.5	0.1
	5	50	25	50	75	25	10	0.7	0.2
	6	75	50	50	75	25	10	0.5	0.1
3	1	50	25	50	75	20	8	0.5	0.1
	2	75	50	50	75	20	8	0.5	0.1
	3	50	25	50	75	25	10	0.5	0.1
	4	75	50	50	75	25	10	0.5	0.1
	5	50	25	50	75	25	10	0.7	0.2
	6	75	50	50	75	25	10	0.5	0.1
4	1	50	25	50	75	8	5	0.5	0.1
	2	75	50	50	75	8	5	0.5	0.1
	3	50	25	50	75	10	8	0.5	0.1
	4	75	50	50	75	10	8	0.5	0.1
	5	50	25	50	75	8	5	0.7	0.2
	6	75	50	50	75	8	5	0.5	0.1

Table 1 – SOM parameters used in the several tests.X units and y units are the number of units in x and y used in the SOM; iter1 and iter2 are the number of epochs for the first and second train (the unfolding and fine tune phases); n1 and n2 are the neighbourhood rate used in the first and second train and L1 and L2 are the first and second train learning rate.

The Carto-SOM methodology was implemented using the Matlab SOM Toolbox. Some routines were written or changed in order to perform variant 4.

RESULTS

In Figure 6 we present the USA population cartograms. First map concerns the original USA states projected in an equal-area projection and on the second one we represent the 2001 USA population using a choropleth map. Variant 1 to 4 result from cartograms using test 6 (Table 1). This test provides the best results when comparing the cartogram regions shape with the desired regions shape. On variant 1 USA states grow/shrink according to their population value. Ocean space is allowed to be occupied, so the final cartogram has a rectangular shape. Although the states changed their shape, topology was preserved. On variant 2 we use an equal density for the regions outside the USA boundaries. The chosen density was an average of all USA states density. Variant 3 is similar but instead of a similar density for the area outside USA we use densities according to nearest states. Variant 4 uses only the units that are inside the USA preserving on the final cartogram the USA borders. With this method we allow the states to grow only inside the country leaving the border unchanged.

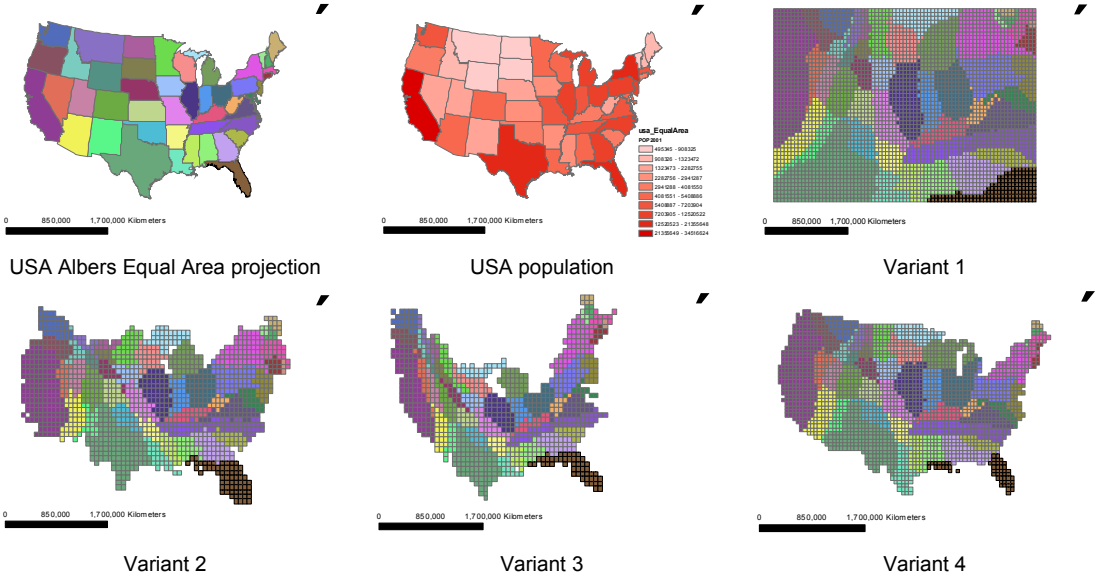


Figure 6 – USA population cartograms.

Some tests were performed in order to evaluate the cartograms quality. In Figure 7 we present a graph where density is represented for the original and cartograms regions. The blue line represents the average density of USA population. As we can see on the graph the tendency on the cartograms is to get a density more close to the average.

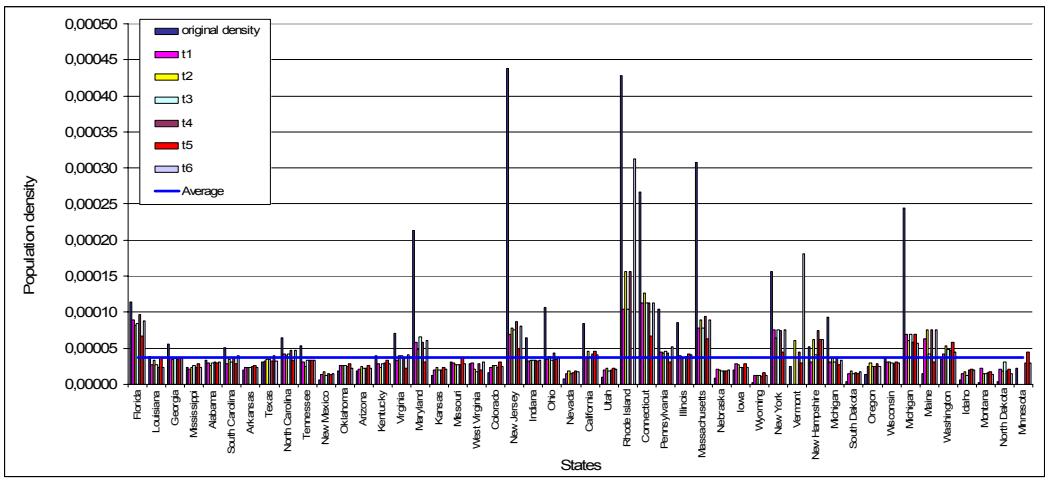


Figure 7 – USA population density. In blue is presented the original population density. T1 to t6 are several tests performed using the variant 4.

On Figure 8 we present a graph crossing state population with its area for the variant 4 tests. Ideally a

cartogram has an equal density in each region. If this was the case we would get a tendency line with a square of the Pearson correlation coefficient (R) equal to 1. As we can see in the graph the best R obtained is 0.882.

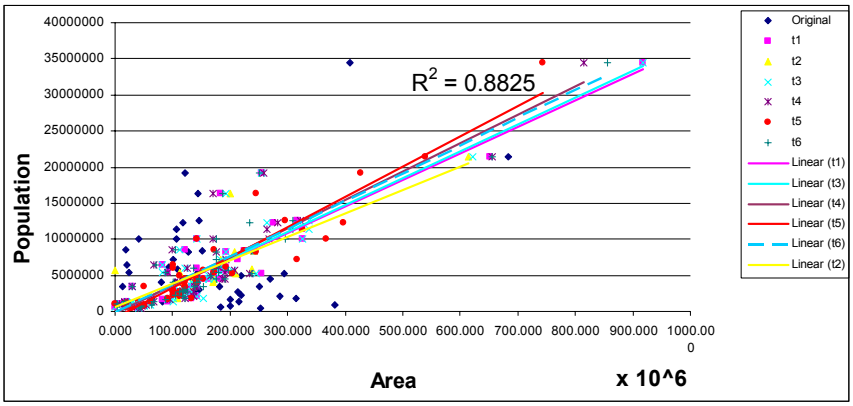
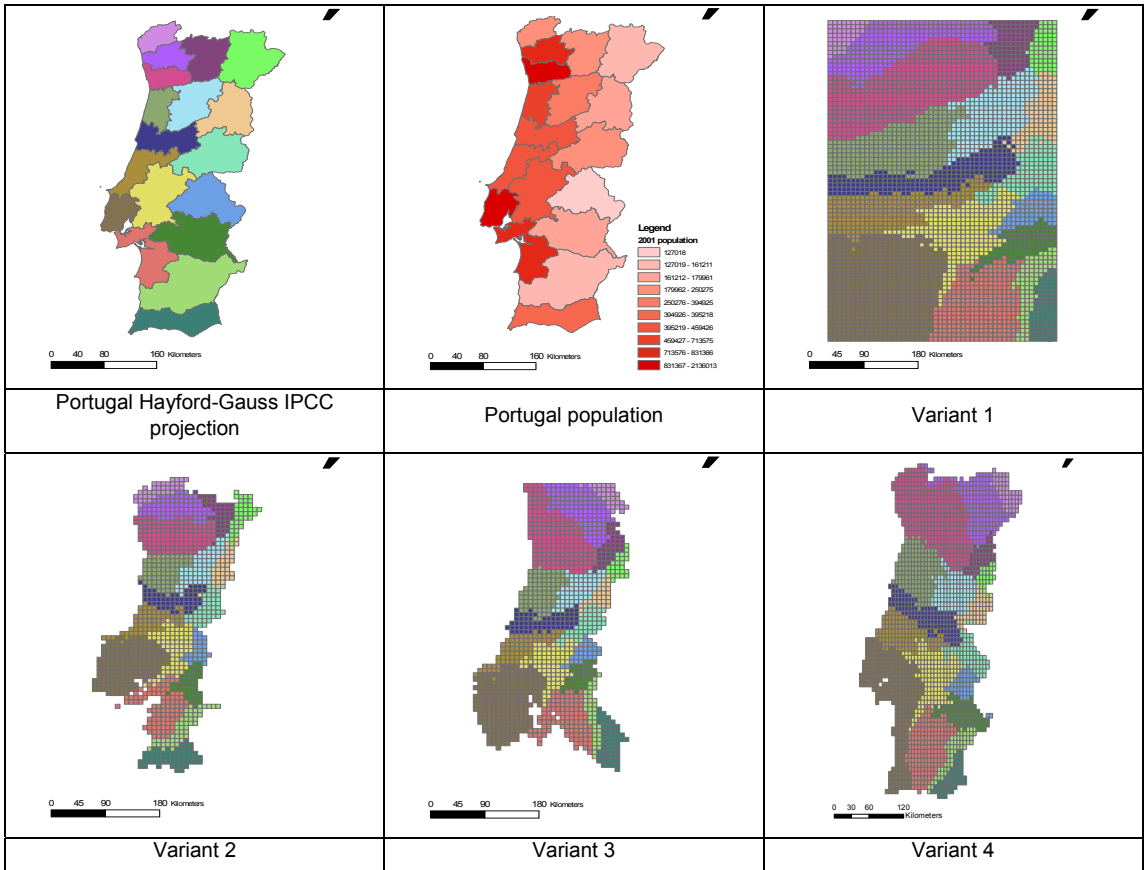


Figure 8 – USA population versus state area.

The Carto-SOM methodology was also tested with Portugal 2001 population data producing the cartograms shown in Figure 9. First map concerns the original shape of Portugal using the Hayford-Gauss IPCC Lisbon projection. A choropleth map of population density, is also present, symbolizing the value of population throw the color lightness. Variant 1 to 4 was also applied, resulting in the cartograms in Figure 9.



Portugal Hayford-Gauss IPCC projection	Portugal population	Variant 1
Variant 2	Variant 3	Variant 4

Figure 9 – Portugal population cartograms.

In Figure 10 we present each region density for the original map and cartograms. The blue line represents the average density of Portugal's population. As in the USA study case, the regions densities tend to be closer to the average, meaning that high-density regions tend to grow and low-density regions compress its shape.

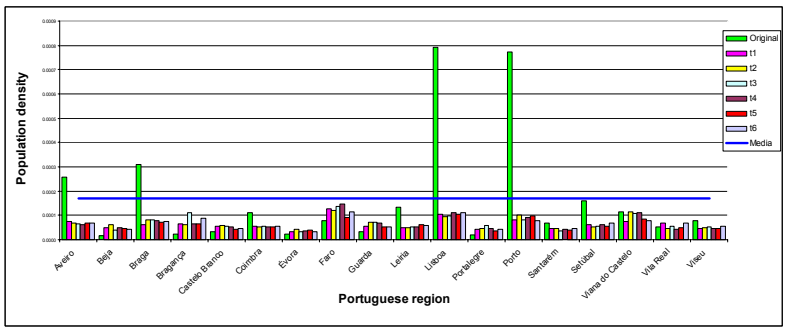


Figure 10 – Portugal population density. In blue is presented the original population density. T1 to t6 are several tests performed using the variant 4.

In Figure 11 each region population and its area is shown for the variant 4 tests. For this variant the best square of the Pearson correlation coefficient (R) equals 0.891

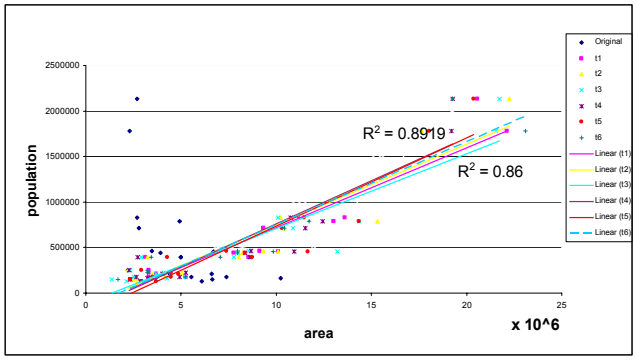


Figure 11 – Portugal population versus region area.

CONCLUSIONS

In this article we have presented a general method for constructing density-equalizing projections or cartograms providing a tool for geographic data presentation and analysis. We have presented tests made with the cartograms which indicate that the cartograms created are good representations of the study variables. Although the proposed algorithm is an effective cartogram generation algorithm, promising directions for further research still remain. It would be interesting to include in the algorithm methods for computing the final cartogram shape giving it a more realistic boundary instead of using cells. Another improvement in this method would be to increase of the number of units used in the SOM training. This change requires a deeper change in the SOM standard algorithm.

REFERENCES

1. **RAISZ, E.**, *The rectangular statistical cartogram.* The Geographical Review, 1934. **24**: p. 292-296.

2. **TOBLER, W.**, *Pseudo-Cartograms.* The American Cartographer, 1986. **13,1**: p. 43-50.

3. **HOUSE, D. and C. KOCMOUD**. *Continuous cartogram construction*. in *Proceedings of IEEE Visualization*. 1998.

4. **GUSEIN-ZADE, S. and V. TIKUNOV**, *A New Technique for Constructing Continuous Cartograms.*

Cartography and Geographic Information Systems, 1993. **20 (3)**: p. 167-173.

5. **KOHONEN, T.**, *Self-organizing formation of topologically correct feature maps.* Biol. Cyb, 1982. **43 (1)**: p. 59-69.

6. **KASKI, S. and T. KOHONEN**, *Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world*, in *Neural Networks in Financial Engineering*, N. Apostolos-Paul, et al., Editors. 1996, World Scientific: Singapore. p. 498-507.

7. **Changming, D. and L. Lin**. *Constructing contiguous area cartogram using arcview avenue*. in *The Proceedings of Geoinformatics'99 Conference*. 1999. Ann Arbor: Li, B., et al., (eds.).

8. **DOUGENIK, J., N. CHRISMAN, and D. NIEMEYER**, *An algorithm to construct continuous area cartograms.* Professional Geographer, 1985: p. 37:75-81.

9. **Heilmann, R., D.A. Keim, C. Panse, and M. Sips**. *RecMap: Rectangular Map Approximations*. in *IEEE Symposium on Information Visualization*. 2004. Austin, Texas, USA.

10. **Gastner, M. and M.E.J. Newman**. *Diffusion-based method for producing density-equalizing maps*. in *Proceedings of the National Academy of Sciences of the United States of America*. 2004.

11. **NCGIA USGS**, *Cartogram Central*. 2002.

12. **Olson, J.**, *Noncontiguous Area Cartograms.* The Professional Geographer, 1976. **28, 4**: p. 371-380.

13. **TOBLER, W.**, *A Continuous Transformation Useful for Districting.* Annals, New York Academy of Sciences, 1973. **219**: p. 215-220.

14. **Keim, D.A., S.C. North, and C. Panse**, *Medial-axes based Cartograms.* AT&T Labs Research, 2003.

15. **Kohonen, T., J. Hynninen, J. Kangas, and J. Laaksonen**, *The Self-Organizing Map Program Package*. 1995, University of Technology: Helsinki.

16. **Bação, F., V. Lobo, and M. Painho**, *The Self-Organizing Map, the Geo-SOM, and relevant variants for geosciences*, in *Computers and Geosciences*. 2005, Elsevier. p. 155-163.

AUTHORS INFORMATION

Roberto HENRIQUES

Research assistant at the New Technologies Laboratory from the Institute of Statistics and Information Management (ISEGI-UNL) teaches Geographic Information Science (GIS). UNIGIS C&SIG master attendee, presently working on the master's thesis.

Fernando BAÇÃO

Professor at the Institute of Statistics and Information Management (ISEGI-UNL) and UNIGIS Portugal teaches Geographic Information Science (GIS) and Knowledge Discovery related courses. He holds a PhD. from the New University of Lisbon in Information Management and has published mostly on topics related with the application of Knowledge Discovery Tools within the GIS field. He's currently the director of the undergraduate program at ISEGI-UNL. For more information please visit http://www.isegi.unl.pt/ensino/docentes/fbacao/index.html

Vitor LOBO

Victor José de Almeida e Sousa Lobo is an Associate Professor at the Portuguese Naval Academy, and an invited Professor at the Institute of Statistics and Information Management of the New University of Lisbon (ISEGI-UNL). He has a PhD and MSc in Computer Science Engineering by the Faculty of Science and Technology of the New University of Lisbon, and an Engineering Degree in Electrical Engineering by the Technical University of Lisbon. His main interests are in Neural Networks, Self-Organization, Pattern Recognition, Clustering, and more recently in Knowledge Discovery Tools within the GIS field. For more information, please visit http://www.isegi.unl.pt/ensino/docentes/vlobo