Clustering census data: comparing the performance of self-organising maps and k-means algorithms

Fernando Bação[1], Victor Lobo[1,2], Marco Painho[1]

[1] ISEGI/UNL, Campus de Campolide, 1070-312 LISBOA, Portugal
bacao@isegi.unl.pt
[2] Portuguese Naval Academy, Alfeite, 2810-001 ALMADA, Portugal
vlobo@isegi.unl.pt

Abstract. Clustering techniques are frequently used to analyze census data and obtain meaningful large scale groups. One of the most used techniques is the widely known k-means clustering algorithm. In recent years, Kohonen's self-organizing Maps have been successfully used to perform similar tasks. Nevertheless, evaluation studies comparing these two approaches are rare and usually inconclusive. In this paper an experimental approach to this problem is adopted. Through the use of synthetic data a particular environment is set up, and these two approaches are compared. Additional, tests are performed using real-world data based on the small area Portuguese census data. The tests focus on two main issues. The first concerns the quality of the K-means algorithm as an intra-cluster variance minimization tool, when compared with Self-Organizing Maps. The second issue deals with the structural impact that sub-optimal solutions, found using k-means, can have in the resulting clustering. The results of the empirical experiments suggest that Self-Organizing Maps (SOM) are more robust to outliers than the k-means method.

1 Introduction

The connection of Geographical Information Systems (GIS) and census data made available a huge volume of digital geo-referenced data [1]. This created opportunities to improve the available knowledge on a number of socio-economic phenomena that are at the heart of Geographical Information Science (GISc). Nevertheless, it also shaped new challenges and raised unexpected difficulties on the analysis of multivariate spatially referenced data. Today, the availability of methods able to perform sensible data reduction, on vast amounts of high dimensional data, is a central issue in science generically and GIScience is no exception. The need to transform into information the massive digital databases that result from decennial census operations has stimulated work in a number of research areas.

The term cluster analysis encompasses a wide group of algorithms (for a comprehensive review see [2]). The main goal of such algorithms is to organize data into meaningful structures. This is achieved through the arrangement of data observations into groups based on similarity. These methods have been extensively applied in different research areas including data mining [3, 4], pattern recognition [5, 6], statistical data analysis [7]. Geographers and urban researchers are among those who have heavily relied on cluster algorithms within their research work [8, 9]. Research on geodemographics [10-12] [13], identification of deprived areas [14], and social services provision [15] are examples of the relevance that clustering algorithms have within today's GISc research. For this reason possible improvements on existing clustering methodologies and the introduction of new tools constitute a relevant issue in GISc.

Recently, Self-Organizing Maps have been proposed as a step forward in the improvement of small-area typologies based on census data [12], traditionally developed using k-means algorithms. In fact, there have been tests comparing SOM's with other clustering methods such as k-means [16-18]. Conclusions seem to be ambivalent as different authors point to different conclusions, and no definitive results have emerged from extensive testing. Some authors [19] [16, 17] suggest that SOM performs equal or worst than statistical approaches, other authors conclude the opposite [18] [12].

The main objective of this paper is to evaluate the performance of the SOM and k-means in the clustering problem, under specific conditions. Especially relevant is the possibility of providing empirical evidence to support the allegations that SOM can be a more effective tool in census-based data clustering. It is well known that data quality is often a problem and has negative effects on the quality of the results. Robustness to outliers and poor quality data is certainly an important characteristic of any algorithm used in census data clustering. Clustering methods should be capable of providing satisfactory results and tackle the challenges prompted by census data.

The issue of proving the superiority of one clustering method over another is difficult and the criteria to establish comparisons elusive. The methodology used here to compare the performance of the two algorithms consists in using two synthetic datasets, and three real-world datasets are used to evaluate and confirm findings made in the synthetic datasets.

In the next section an overview of the characteristics and problems affecting census-based data is made, followed by a presentation of the formal clustering problem. In section three the methods tested are presented, with emphasis on the description of the SOM. Section four presents the different sets of data used in testing the algorithms. Section five deals with the results from the tests and finally, section six addresses the conclusion that can be drawn from this work.

2 Census Data Characteristics and the Standard Process of Classification

Census data is a fundamental source of information in numerous research areas within GISc. Because of this, algorithms used by geographers for clustering should be capable of dealing with specific problems associated with the use of census data [13]. For the purpose of this work we would like to highlight the problems which result from dealing with high dimensional datasets that may have measurement errors. Additionally, and more closely related with the special nature of spatial data [20], in census datasets one should expect variations in size and homogeneity in the geographical units and also non-stationary in the relations between variables, which are bound to change across regions. All these problems concur to the complexity which is involved in clustering census data. Emphasis should be put on the importance of using robust clustering algorithms, algorithms which, as much as possible, should be insensible to the presence of outliers.

Closely related with robustness is the capability of modelling locally, preserving the impact of errors and inaccuracies in data within local structures of the clustering, rather than allowing these problems to have a global impact on the results. The idea is to find algorithms which degrade progressively in the presence of outliers instead of abruptly disrupting the clustering structure. Improvements in clustering algorithms will yield benefits in all research areas which use census data as part of their analysis process [12]. Although the performance of the clustering methods in itself is not enough to solve all the problems related with the quality of census based clusterings, it is definitely a relevant issue.

The common procedure of clustering census data includes the following 7 steps [21]:

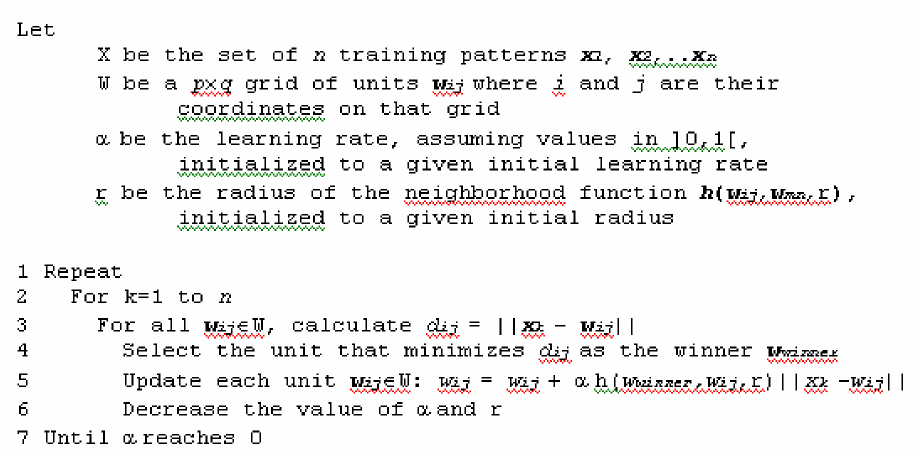
1. Definition of the clustering objective;
2. Careful choice of the variables to use;
3. Normalizing and orthogonalizing the data;
4. Clustering the data;
5. Labelling, interpretation and evaluation;
6. Mapping the results;
7. Regionalizing.

Here we concentrate in step 4, the clustering algorithm that should be used to achieve the desired data reduction. In recent years alternatives to the K-means algorithm have been proposed. A number of authors have pointed out the potential of using SOM's in clustering tasks, e.g. [22]. Specifically in GISc, SOM has been proposed as an improvement over k-means method on the grounds that it provides a more flexible approach to census data clustering [13], a property which can be a particularly useful. Nevertheless, hard evidence of the superiority of SOM over k-means in clustering census data is still missing. There are well known properties which characterize the k-means clustering algorithm. First, and due to the use of Euclidean distance, k-means is especially effective dealing with Gaussian distributions [23]. Secondly, k-means performance is especially sensible to the presence of outliers [2, 24]. Thirdly, initialization conditions have an important impact on the performance of the method.

3 Self-Organizing Map and K-Means Algorithm

Although the term “Self-Organizing Map” could be applied to a number of different approaches, we shall use it as a synonym of Kohonen’s Self Organizing Map [25] [22], or SOM for short, also known as Kohonen Neural Networks. These maps are primarily visualization and analysis tools for high dimensional data [22], but they have been used for clustering, dimensionality reduction, classification, sampling, vector quantization, and data-mining [22, 26].

The basic idea of a SOM is to map the data patterns onto a n-dimensional grid of neurons or units. That grid forms what is known as the output space, as opposed to the input 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. So as to allow an easy visualization, the output space is usually 1 or 2 dimensional. The basic SOM training algorithm can be described as follows:



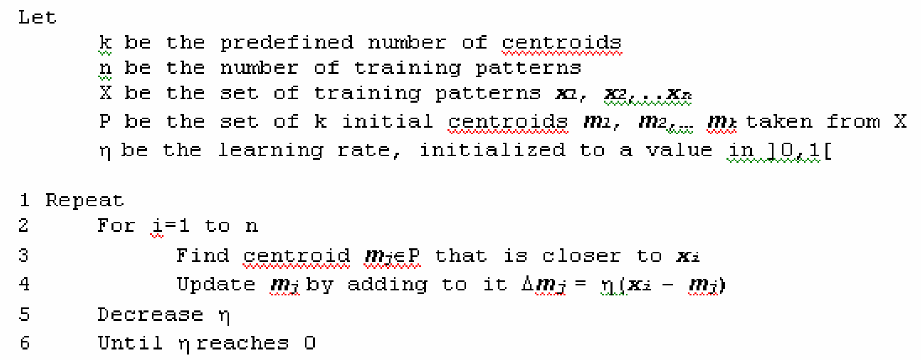
```
Let
      X be the set of $n$ training patterns $x_1$, $x_2$,..$x_n$
      W be a $p$x$q$ grid of units $w_{ij}$ where $i$ and $j$ are their
            coordinates on that grid
      $\alpha$ be the learning rate, assuming values in ]0,1[,
            initialized to a given initial learning rate
      r be the radius of the neighborhood function $h(w_{ij}, w_{mn}, r)$,
            initialized to a given initial radius

Repeat
  For k=1 to $n$
    For all $w_{ij} \in W$, calculate $d_{ij} = ||x_k - w_{ij}||$
      Select the unit that minimizes $d_{ij}$ as the winner $w_{winner}$
      Update each unit $w_{ij} \in W$: $w_{ij} = w_{ij} + \alpha h(w_{winner}, w_{ij}, r) ||x_k - w_{ij}||$
      Decrease the value of $\alpha$ and r
Until $\alpha$ reaches 0
```

The neighborhood function h is usually a function that decreases with the distance (in the output space) to the winning unit, and is responsible for the interactions between different units. During training, the radius of this function will usually decrease, so that each unit will become more isolated from the effects of its neighbors. It is important to note that many implementations of SOM decrease this radius to 1, meaning that even in the final stages of training each unit will have an effect on its nearest neighbors, while other implementations allow this parameter to decrease to zero. The learning rate a must converge to 0 so as to guarantee convergence and stability for the SOM [22].

The k-means is widely known and used so only a brief outline of the algorithm is presented (for a thorough review see [5-7]. K-means is an iterative procedure, to place cluster centers, which quickly converges to a local minimum of its objective function [24, 27]. This objective function is sum of the squared Euclidean distance (L2) between each data point and its nearest cluster center [24, 28] this is also known as “square-error distortion” [29]. It has been shown that k-means is basically a gradient algorithm [28, 30] which justifies the convergence properties of the algorithm.

The original online algorithm [31] is as follows:



```
Let
      k be the predefined number of centroids
      n be the number of training patterns
      X be the set of training patterns $x_1$, $x_2$,..$x_n$
      P be the set of k initial centroids $m_1$, $m_2$,… $m_k$ taken from X
      $\eta$ be the learning rate, initialized to a value in ]0,1[

1 Repeat
2     For i=1 to n
3           Find centroid $m_j \in$P that is closer to $x_i$
4           Update $m_j$ by adding to it $\Delta m_j = \eta(x_i - m_j)$
5     Decrease $\eta$
6     Until $\eta$ reaches 0
```

There are a large number of variants of the k-means algorithm. In this study we use the generalized Lloyd’s algorithm [6, 32], which yields the same results as the algorithm above [30]. The popularity of this variant in statistical analysis is due to its simplicity and flexibility. As the generalized Lloyd’s algorithm doesn’t specify the placement of the initial seeds, in this particular application the initialization is done through randomly assigning observations as a cluster seeds.

It must be noted that SOM and k-means algorithms are rigorously identical when the radius of the neighborhood function in the SOM equals zero [33]. In this case the update only occurs in the winning unit just as happens in k-means (step 4).

4 Experimental Setting

4.1 Datasets used

The data used in the tests is composed of 4 basic datasets, two synthetic and two real-world. The real-world datasets used are the well known iris dataset [34] and sonar dataset [35]. The iris dataset has 150 observations with 4 attributes and 3 classes, while the sonar dataset has 208 observations with 60 attributes and 2 classes. Two synthetic datasets were created to compare the robustness of the two clustering methods. The first dataset, DS1, comprises 400 observations in two-dimensions with 4 clusters. Each of these clusters has 100 observations with a Gaussian distribution around a fixed center, as shown in figure 1. The variance of these Gaussians was gradually increased during our experiments, yielding quite scattered clusters as depicted in figure 2. The second data set, DS2, consists of 750 observations with 5 clusters with Gaussian distributions defined in a 16 dimensional space.

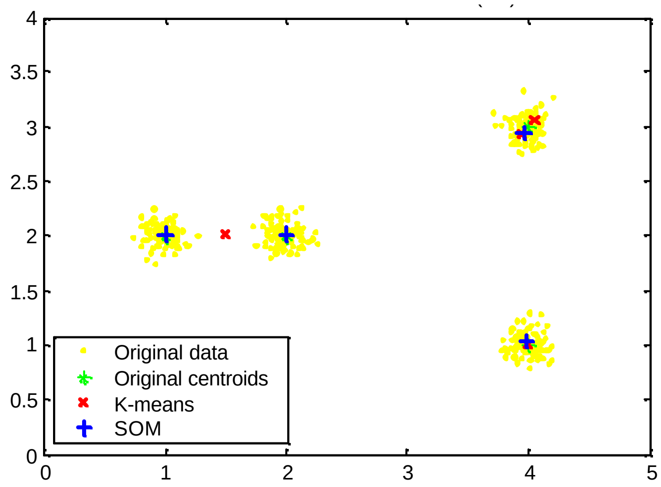


Fig. 1. The DS1 with the lowest value of standard deviation, showing 4 well defined clusters of 100 observations

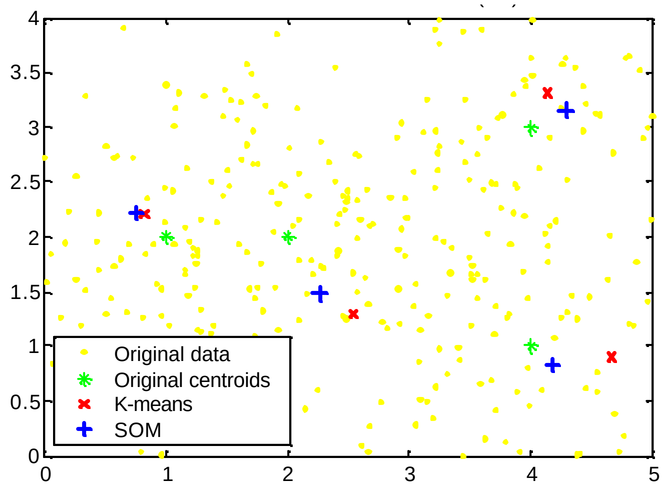


Fig. 2. The DS1 with the highest value of standard deviation, the 4 clusters are not identifiable as observations are very scattered

4.2 Robustness assessment measures

In order to access the performance of the two methods a set of three measurements was used. The first one is the quadratic error i.e., the sum of the squared distances of each point to the centroid of its cluster. This error is divided by the total dispersion of each cluster so as to obtain a relative measure. This measure is particularly relevant as it is the objective function of the k-means algorithm. Additionally, the standard deviation of the mean quantization error is calculated in order to evaluate the stability of the results found in the different trials. The second measure used to evaluate the clustering is the mean classification error. This measure is only valid in the case of classification problems and is the number of observations attributed to a cluster where they do not belong. Finally, a structural measurement is used in order to understand if the structural coherence of the groups is preserved by the clustering method. This measure is obtained by attributing to each cluster center a label based on the labels of the observations which belong to its Voronoi polygon. If more than one centroid receive a given label (and thus at least one of the labels is not attributed) then the partition is considered to be structurally damaged.

5 Results

Each one of the datasets was processed 100 times by each algorithm, and the results presented in table 1 constitute counts or means. Table 1 presents a summary of the most relevant results. A general analysis of table 1 shows a tendency for SOM to outperform k-means. The mean quadratic error over all the datasets used is always smaller in the case of the SOM, although in some cases the difference is not sufficiently large to allow conclusions. The standard deviation of the quadratic error is quite enlightening showing smaller variations in the performance of the SOM algorithms. The class error indicator reveals a behavior similar to the mean quadratic error. Finally, the structural error is quite explicit making the case that SOM robustness is superior to k-means.

Looking closer at the results in different datasets, there is only one data set in which k-means is not affected by structural errors. The reason for this is related with the configuration of the solution space. In the sonar dataset the starting positions of the k-means algorithm are less relevant than in the other 3 datasets.

Table 1.

Dataset	Method	Quadratic error	Std(Qerr)	ClassErr	Struct Err
IRIS	SOM	86.67	0.33	9.22	0

	k-means	91.35	25.76	15.23	18
SONAR	SOM	280.80	0.10	45.12	0
	k-means	280.98	3.18	45.34	0
DS1	SOM	9651.46	470.36	1.01	0
	k-means	11341.49	2320.27	12.77	58
DS2	SOM	27116.40	21.60	7.40	0
	k-means	27807.97	763.22	15.51	49

Figure 3 shows the results achieved by the two methods in dataset DS1. It is quite clear that SOM is more stable than k-means as the structural coherence of the clustering varies very little. With low levels of standard deviation (all observations very close the clusters centroids) k-means shows a poor performance failing structural coherence in more than 50% of the runs. On the contrary the SOM fails to get the right structure only 10% of the runs. As the standard deviation grows k-means improves the percentage of runs in which the structural coherence is right. Nevertheless, it never gets to the 100% level in which SOM scores in every run between 0.2 and 0.9 standard deviation.

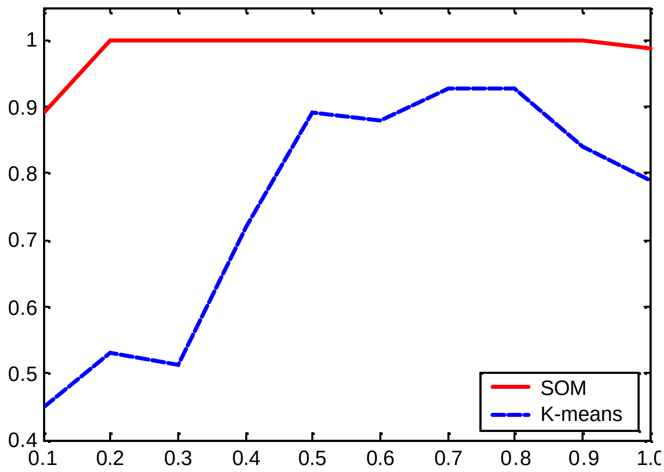


Fig. 3. One kernel at x_S (*dotted kernel*) or two kernels at x_i and x_j (*left and right*) lead to the same summed estimate at x_S. This shows a figure consisting of different types of lines. Elements of the figure described in the caption should be set in italics, in parentheses, as shown in this sample caption. The last sentence of a figure caption should generally end without a period

Through the tests it became clear that initialization conditions play a major role in the quality of the results produced by the k-means algorithm, as it has been noted by different authors (e.g [23]). A number of strategies have been proposed in order to improve k-means tolerance to initial conditions. These are beyond the scope of this paper. Clearly the gradient nature of the k-means algorithm, which largely accounts for its computational efficiency, is also responsible for its sensitivity to local optima.

The real-world dataset refers to enumeration districts (ED) of the Lisbon Metropolitan Area and includes 3968 ED's which are characterized based on 65 variables, from the Portuguese census of 2001. Exploratory analysis of this dataset using large size SOMs and U-Matrices suggests that we should consider 6 clusters within this dataset. To find the exact locations and members of these 6 clusters we applied a batch k-means algorithm to this data, and compared the results with those obtained with a 6x1 SOM. In both cases we repeated the experiment 100 times with random initializations. The quadratic error obtained with k-means was 3543 ± 23 with a minimum of 3528, whereas with SOM we obtained 3533 ± 6 with a minimum of 3529. In figure 4 we present a histogram of the quadratic errors obtained with both approaches.

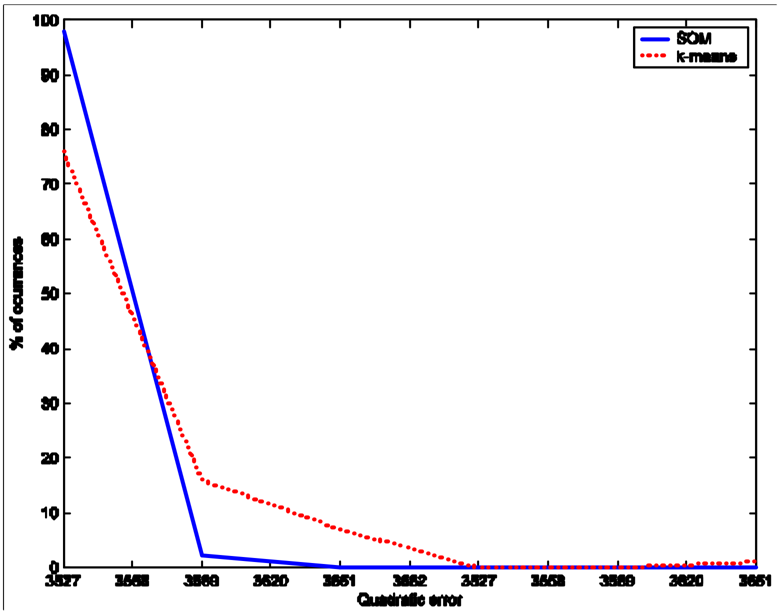


Fig. 4. Histogram of the quadratic errors using k-means and SOM to cluster Lisbon's census data into 6 groups

These results show that the best clustering obtained with each method is practically the same, but on average SOM outperforms k-means and has far less variation in it's results.

6 Conclusions

The first and most important conclusion that can be drawn from this study is that SOM is less prone to local optima than k-means. During our tests it is quite evident that the search space is better explored by SOM. This is due to the effect of the neighborhood parameter which forces units to move according to each other in the early stages of the process. This characteristic can be seen as an "annealing schedule" which provides an early exploration of the search space [36]. On the other hand, k-means gradient orientation forces a premature convergence which, depending on the initialization, may frequently yield local optimum solutions.

It is important to note that there are certain conditions that must be observed in order to render robust performances from SOM. First it is important to start the process using a high learning rate and neighborhood radius, and progressively reduce both parameters to zero. This constitutes a requirement for convergence [22] but also raises the probability of reaching optimal results.

SOM's dimensionality is also an issue, as our tests indicate that 1-dimensional SOM will outperform 2-dimensional matrices. This can be explained by the fact that the "tension" exerted in each unit by the neighboring units is much higher in the case of the matrix configuration. This tension limits the plasticity of the SOM to adapt to the particular distribution of the dataset. Clearly, when using a small number of units it is easier to adapt a line than a matrix.

These results support Openshaw's claim which points to the superiority of SOM when dealing with problems having multiple optima. Basically, SOM offers the opportunity for an early exploration of the search space, and as the process continues it gradually narrows the search. By the end of the search process (providing the neighborhood radius decreases to zero) the SOM is exactly the same as k-means, which allows for a minimization of the distances between the observations and the cluster centers.

1. Batty, M. and P. Longley, *Analytical GIS: The Future*, in *Spatial Analysis: Modelling in a GIS Environment*, P. Longley and M. Batty, Editors. 1996, Geoinformation International: Cambridge. p. 345-352.
2. Jain, A.K., M.N. Murty, and P.J. Flynn, *Data Clustering: A review*. ACM Computing Surveys, 1999. **31**(3): p. 264-323.
3. Fayyad, U., et al., *Advances in Knowledge Discovery and Data Mining*. 1996: AAAI/MIT Press.
4. Han, J. and M. Kamber, *Data Mining : Concepts and Techniques*. 2000: Morgan Kaufmann.
5. Fukunaga, K., *Introduction to statistical patterns recognition*. 2nd ed. 1990: Academic Press Inc.
6. Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern Classification*. 2001: Wiley-Interscience.
7. Kaufman, L. and P. Rousseeuw, *Finding Groups in Data*. 1989: John Wiley and Sons.

8. Han, J., M. Kamber, and A. Tung, *Spatial clustering methods in data mining*, in *Geographic Data Mining and Knowledge Discovery*, H. Miller and J. Han, Editors. 2001, Taylor & Fancis: Londor. p. 188-217.
9. Plane, D.A. and P.A. Rogerson, *The Geographical Analysis of Population: With Applications to Planning and Business*. 1994, New York: John Wiley & Sons.
10. Feng, Z. and R. Flowerdew, *Fuzzy geodemographics: a contribution from fuzzy clustering methods*, in *Innovations in GIS 5*, S. Carver, Editor. 1998, Taylor & Francis: London. p. 119-127.
11. Birkin, M. and G. Clarke, *GIS, geodemographics and spatial modeling in the UK financial service industry.* Journal of Housing Research, 1998. **9**: p. 87-111.
12. Openshaw, S., M. Blake, and C. Wymer, *Using neurocomputing methods to classify Britain's residential areas*, in *Innovations in GIS*, P. Fisher, Editor. 1995, Taylor and Francis. p. 97-111.
13. Openshaw, S. and C. Wymer, *Classifying and regionalizing census data*, in *Census Users Handbook*, S. Openshaw, Editor. 1994, Geo Information International: Cambridge, UK. p. 239-270.
14. Fahmy, E., D. Gordon, and S. Cemlyn. *Poverty and Neighbourhood Renewal in West Cornwall*. in *Social Policy Association Annual Conference*. 2002. Nottingham, UK.
15. Birkin, M., G. Clarke, and M. Clarke, *GIS for Business and Service Planning*, in *Geographical Information Systems*, M. Goodchild, et al., Editors. 1999, Geoinformation: Cambridge.
16. Balakrishnan, P.V., et al., *A study of the classification capabilities of neural networks using unsupervised learning: a comparison with k-means clustering*. Psychometrika, 1994. **59**(4): p. 509-525.
17. Waller, N.G., et al., *A comparison of the classification capabilities of the 1-dimensional Kohonen neural network with two partitioning and three hierarchical cluster analysis algorithms.* Psychometrika, 1998. **63**(1): p. 5-22.
18. Openshaw, S. and C. Openshaw, *Artificial intelligence in geography*. 1997, Chichester: John Wiley & Sons.
19. Flexer, A., *On the use of self-organizing maps for clustering and visualization*, in *Principles of Data Mining and Knowledge Discovery*, Z. J.M. and R. J., Editors. 1999, Springer. p. 80-88.
20. Anselin, L., *What is special about spatial data? Alternative perspectives on spatial data analysis*, in *Spatial Statistics, Past, Present and Future*, D.A. Griffith, Editor. 1990, Institute of Mathematical Geography: Ann Arbor, ML. p. 63-77.
21. Rees, P., et al., *ONS classifications and GB profiles: census typologies for researchers*, in *The Census Data System*, P. Rees, D. Martin, and P. Williamson, Editors. 2002, Wiley: Chichester. p. 149-170.
22. Kohonen, T., *Self-Organizing Maps*. Information Sciences. 2001: Springer.
23. Bishop, C.M., *Neural Networks for Pattern Recognition*. 1995: Oxford University Press.

24. Bradley, P. and U. Fayyad. *Refining initial points for K-means clustering*. in *International Conference on Machine Learning (ICML-98)*. 1998.

25. Kohonen, T. *Clustering, Taxonomy, and Topological Maps of Patterns*. in *Proceedings of the 6th International Conference on Pattern Recognition*. 1982.

26. Vesanto, J. and E. Alhoniemi, *Clustering of the Self-Organizing Map*. IEEE Transactions on Neural Networks, 2000.

27. Kanungo, T., et al., *An efficient k-means clustering algorithm: analysis and implementation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002. **24**(7): p. 881-892.

28. Selim, S.Z. and M.A. Ismail, *k-means type algorithms: a generalized convergence theorem and characterization of local optimality*. IEEE Trans. Pattern Analysis and Machine Intelligence, 1984. **6**: p. 81-87.

29. Jain, A.K. and R.C. Dubes, *Algorithms for clustering data*. 1988: Prentice Hall.

30. Bottou, L. and Y. Bengio, *Convergence Properties of the K-Means Algorithms*, in *Advances in Neural Information Processing System*. 1995, MIT Press: Cambridge, MA. p. 585-592.

31. MacQueen, J. *Some methods for classification and analysis of multivariate observation*. in *5th Berkeley Symposium on Mathematical Statistics and Probability*. 1967: University of California Press.

32. Loyd, S.P., *Least Squares quantization in PCM*. IEEE Transactions on Information Theory, 1982. **28**(2): p. 129-137.

33. Bodt, E.d., M. Verleysen, and M. Cottrell. *Kohonen Maps versus Vector Quantization for Data Analysis*. in *ESANN'1997*. 1997. Bruges.

34. Fisher, R.A., *The use of Multiple Measurements in Taxonomic Problems*. Annals of Eugenics, 1936. **VII**(II): p. 179-188.

35. Sejnowski, T.J. and P. Gorman, *Learned Classification of Sonar Targets Using a Massively Parallel Network*. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1988. **36**(7): p. 1135 -1140.

36. Bodt, E.d., M. Cottrell, and M. Verleysen. *Using the Kohonen Algorithm for Quick Initialization of Simple Competitive Learning Algorithms*. in *ESANN'1999*. 1999. Bruges.