Application of Clustering Analysis in Gravity Databases

Jorge Luiz de Lima Matias, Eder Cassola Molina (IAG-USP)

This paper was prepared for presentation during the 12th International Congress of the Brazilian Geophysical Society held in Rio de Janeiro, Brazil, August 15-18, 2011.

Abstract

Clustering is a Data Mining technique which aims to analyze and describe data sets automatically, with or without *a priori* knowledge about them. Automatic analysis are used to simplify the human job in a data set study, which often is difficult due to the nature of the data or its complexity.

Clustering analysis methods are used in several areas of knowledge to search and analyze patterns of a data set, and proved to be a powerful tool in multivariate analyses. These methods describe the data arranging them in several clusters, using a similarity metric. In this work some of these kind of clustering are discussed, and two of them are discussed in details: partitional and hierarchical agglomerative methods.

To study the applicability of Data Mining in geophysical data sets clustering analysis methods are used to describe a gravity data set of an offshore area near of the northeast Brazilian coast, and the results demonstrates that clustering can become a powerful tool in geophysical research.

Introduction

The importance of data analysis is always rising significantly, as is its size and complexity. One can be rich in data, but many times poor in knowledge. Many areas of knowledge and business have the same problem: too much data for analyze. Data Mining tools are used to solve this problems, usually with interesting results.

The Data Mining goal is to help in the knowledge discovery process into databases, finding new and useful patterns automatically, what would be difficult to do in a trivial way by a human analyst; this method usually permits the discovering of new and very interesting results that was not clearly available into a huge amount of data (Fayyad et al, 1996).

There are several techniques in Data Mining, and they can be classified in two principal types (Rezende, 2003, Mitchell, 1997): predictive and descriptive. A predictive method (e. g. Classification and Regression) uses *a priori* knowledge to classify the new data in already known classes. A descriptive tool describes the data in new classes, e. g. Clustering, Association Rules and Summarization.

Each method in Clustering analysis aims to describe the data in a different way, with a different metric, so the decision of the right method to use in a specific data set is not trivial. This results in a pattern where the data are described as clusters, that are grouped by some similarity criteria. A clustering algorithm can be classified by the way it describes the data (Jain et. al., 1999): partitional, grade based, density based, hierarchical, and so far.

Partitional clustering aims to describe the data in a optimal pattern with a determined number of clusters, dividing the data in mutually exclusive groups (the same data cannot be in more than one cluster) with an iterative method that optimizes the classification. In this work a classical partitional algorithm based in the distance between the clusters' means is used, named K-means, and also a variation of this method that uses a probability based in a multivariate Gaussian, the EM clustering method. The main advantage of these methods is that they optimize the result patterns, but they also have a big disadvantage: the need of *a priori* knowledge about the number of clusters into the data set (Aldridge, 2005).

Other method that is investigated is a hierarchical clustering method, the GHBC (Gaussian Hierarchical Bayesian Clustering). It is an agglomerative method, what means that the process starts with each element forming separate clusters that in successively algorithm iterations are merged in larger clusters based at a similarity criteria, what forms a hierarchy (Metz, 2006). So, the hierarchical method provides not only one optimal pattern, like the partitional methods, but a hierarchy of possible patterns that can describe the data at several levels, which is very useful in data sets with anomalies with different densities (Ankerst et. al., 1999). Hierarchical methods have some problems, as they do not optimize the patterns, and when some data is put into a cluster, it will be in it during all the next patterns of the hierarchy.

There is not a perfect method in data analysis, so different methods give different types of results, and some methods can work better for different types of data. Some methods of clustering analysis are discussed here to show a way that one can use the best features of each one to obtain a better result in some cases.

These techniques were used to analyze a dataset with free air gravity anomaly, geoid height and topography in an offshore area near the northeastern Brazilian coast. The results showed the potential of using this type of technique in geophysical studies.

GHBC Algorithm

Gaussian Hierarchical Bayesian Clustering is an agglomerative algorithm based in the optimization of the posterior similarity of a pattern. The origin of this algorithm is the HBC (Iwayama & Tokunaga, 1995), successfully used in text classification. This algorithm is based on maximizing of the Bayesian posterior probability. Following what is described in Iwayama & Tokunaga (1995), Christ et. al. (2007) and Everitt et al. (2001), maximizing that probability is the same that maximizing the argument U (eq. 1).

$$U = SC(c_iUc_j) / [SC(c_i)*SC(c_j)] \tag{1}$$

where *SC(c)* is the probability of the existence of the cluster *c*, or, in other words, is the probability of all the elements classified into the cluster *c* be produced by this same cluster. *SC* is defined by the multiplication of all these probabilities, named elemental probability (eq. 2).

$$SC(ci)=\pi\ p(d_j|c_i) \tag{2}$$

The difference between them is that the GHBC uses a Gaussian model, where each cluster is described by a multivariate normal distribution, so the probability of a data be part of a cluster (elemental probability) is defined as stated in Equation 3, where d_j is a vector that represents the data, μ_{c_i} is the mean vector of the cluster c_i, Σ_{c_i} is the covariance matrix of the same cluster, and *N* represents the multivariate normal distribution (Murtagh & Raftery, 1984; Banfield & Raftery, 1993; Dasgupta & Raftery, 1998; Villanueva, E. R., 2007):

$$p(d_j|c_i) = N(d_j;\mu_{c_i};\Sigma_{c_i}) \tag{3}$$

The clusters cannot be unitary because the way the elemental probability is calculated, so the process must have an input of an initial pattern of small non-unitary clusters. To produce this initial cluster partition the K-means algorithm was used in this work. Figure 1 illustrates the GHBC algorithm

```
Input:
$D=\{d_1,d_2,...,d_n\}$     Data set
$C_0=\{c_1,c_2,...,c_k\}$     Initial clusters

Output:
Hierarchy of patterns: $\{C_0,C_1,C_2,...,C_{k-1}\}$

Do:
$U = SC(c_iUc_j) / [SC(c_i)*SC(c_j)]$ ,  $c_i,c_j \in C_{k-1}$, $c_i \neq c_j$

$(c_x,c_y) \leftarrow \mathrm{argmax}_{ci,cj}(U)$
$c_z \leftarrow c_x \cup c_y$
$C_k \leftarrow C_{k-1}+c_z-c_x-c_y$

Until:
n-1 iterations, when all elements of the data set Will be at the same cluster.
```

***Fig. 1**: Representation of the GHBC clustering (modified for Villanueva, E. R., 2007).*

K-means algorithm

The K-means algorithm is a partitional algorithm based in the distance between the means of clusters. It describes the cluster by a mean vector, and each data is described as a vector, so a data vector is merged into the cluster that shows a mean vector closest to its value. The input of the algorithm is a range of means vector estimated by the user, and in each iteration the data vectors are merged and, after that, the means vector are recalculated, and these iterations continue optimizing the pattern until the means vector do not change anymore. A representation of this algorithm is showed in figure 2.

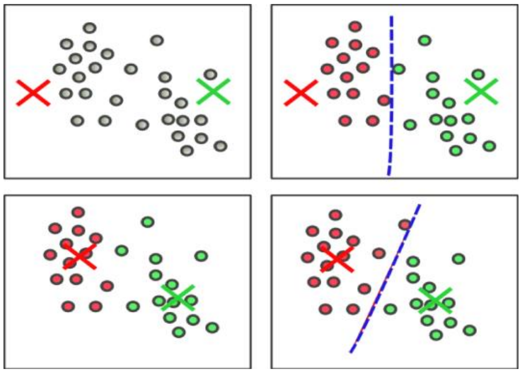

***Fig. 2:** Representation of the iterations of K-means. The Colored "X"s represent the means of the clusters.*

There are a lot of distance metrics that can be used in this method, such as the Euclidian, the square-euclidian, the city-block (manhattan), and so far (Shen, 2005).

The K-means algorithm was used here with a huge number of initial clusters to form a pattern of small but non-unitary clusters that was used as input into the GHBC. This pattern is also used as a form of optimizing the patterns of the top of the hierarchy classified by GHBC.

EM Clustering

The Expectation – Maximization Clustering is very similar to the K-means algorithm in the logical mechanism. The main difference is the metrics used. EM is not distance-based as is K-means. A cluster in EM is described by a normal distribution (Gaussian), and on each iteration the data vectors are merged into the cluster that presents the higher elemental probability (eq. 3) for this data, so the mean vector and the covariance matrix of the cluster are recalculated (Witten, 2000).

In this work the EM Clustering and K-means were used to optimize the hierarchical result of GHBC algorithm. These algorithms need an initial partition to optimize, but we usually do not know this initial partition, neither the best number of clusters to describe the data correctly. Even so, the associated use of these algorithms with GHBC provides good results.

Data set

The data sets used here consist of Free Air gravity anomaly (figure 3.a), geoid height (figure 3.b), and topography (figure 3.c) grids of an offshore area near the northeastern Brazilian coast.

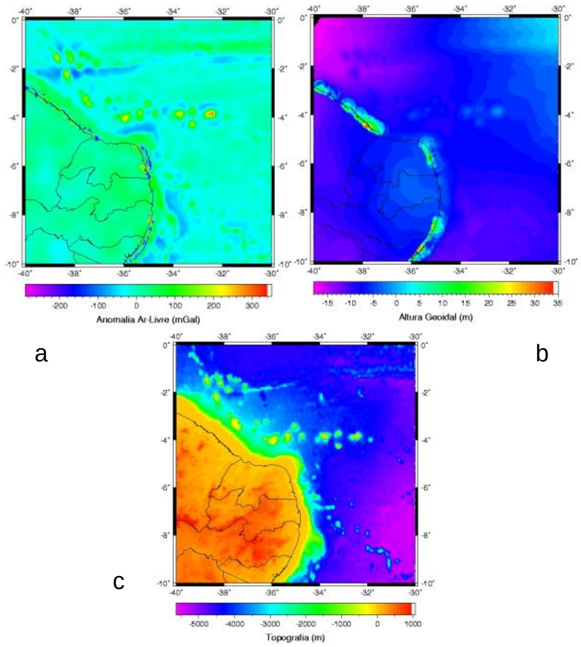

***Fig. 3**: Data set formed by: **a)** Free Air Anomaly **b)** Geoid height **c)** topography.*

Results

The application of the GHBC algorithm in these data set results in a hierarchy of patterns, and some of these patterns is illustrated in figure 4. The input of the algorithm was the data from figure 4.a, formed by use of the K-means algorithm using Euclidian distance, and the resulting input data consisted of 650 vectors of equally spaced data points.

Figure 4 shows only some examples of the results, but is possible see how the algorithm works. The initial cluster (fig. 4.a) is formed by 650 clusters; after 150 iterations of the algorithm 150 clusters were merged into others, and some big groups have been formed, as the clusters pointed by white, black and red arrows (fig. 4.b).

At the sequence, it is clear in figures 4.c and 4.d that some clusters were merged into others to form a bigger cluster marked by white arrows. After that, on figure 4.e, there is a cluster formed near the continental shelf, pointed by a white arrow, and in figure 4.f it is pointed a cluster that follows the continental shelf. Figures 4.g and 4.h demonstrate that this cluster continues to separate at the end of the hierarchy.

All the clusters pointed above (and others) are very closely related to anomalies in the original data set (fig. 3). With the hierarchy tool it is possible to analyze the anomalies in different levels of detail, and the cluster is a good guide to the analyst to notice important patterns into the database.

The next step was to use the K-means algorithm, with different distance metrics. Figure 5 demonstrates the improvement in the similarity of the patterns (the parameter B is the similarity exponent, SC=A*10^B, where SC of a pattern is defined by the multiplication of the similarity of all the clusters of this pattern, that is defined by the eq. 2). As expected, the K-means algorithm improved the patterns of GHBC, and figures 7 and 8 illustrates the difference between the patterns before and after the application of this partitional method.

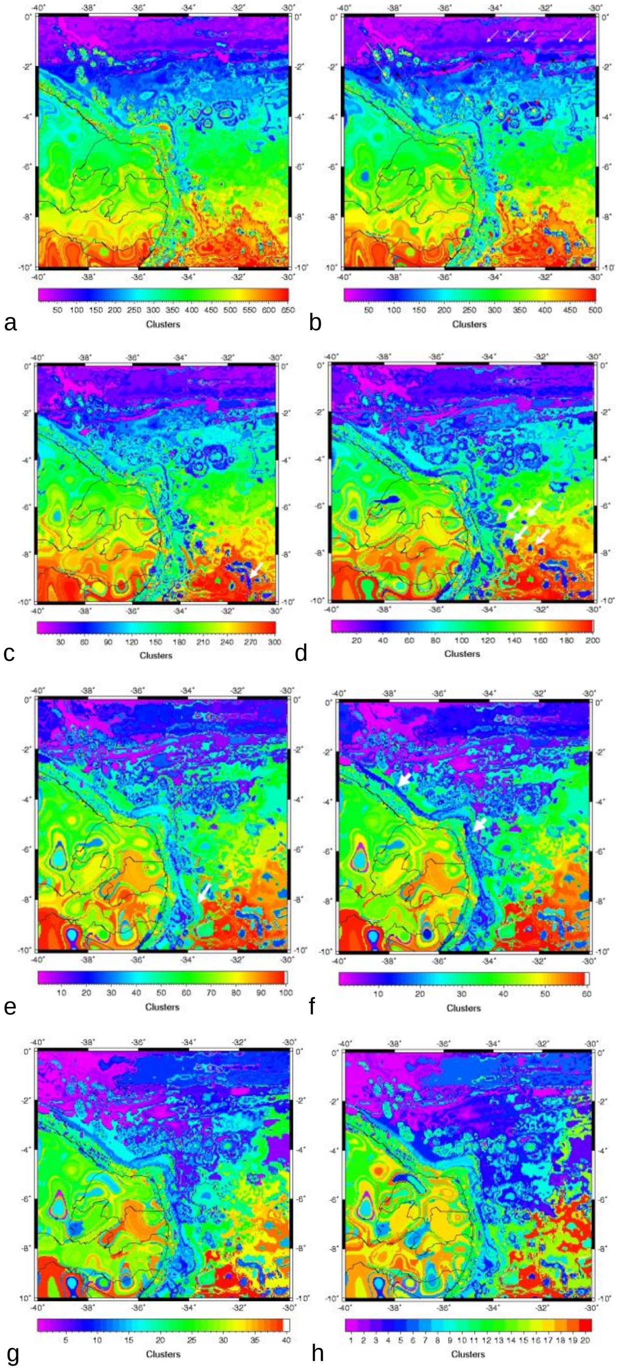

Fig. 4: Illustrations of the hierarchy by GHBC: **a)** Input of the algorithm, by K-means, with 650 clusters; **b)** after 150 iteration, the pattern with 500 clusters; **c)** with more 300 iteration, the pattern with 300 clusters; **d)** pattern with 200 clusters; **e)** pattern with 100 clusters; **f))** pattern with 60 clusters; **g))** pattern with 40 clusters; **h))** pattern with 20 clusters.

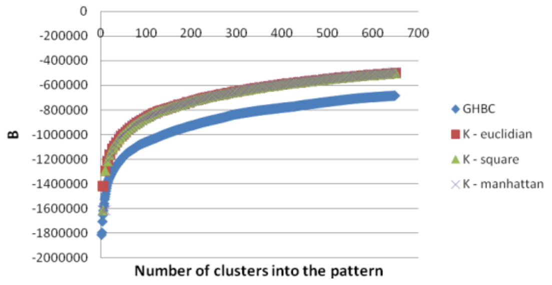

Fig. 5: *Difference of pattern simillarities before (only with GHBC) and after application of the K-means algorithm (with euclidian, square euclidian and Manhattan metrics).*

The EM algorithm was also applied in the GHBC hierarchy, with similar (but more inconstant) results (figure 6). These patterns are showed in figures 7 and 8.

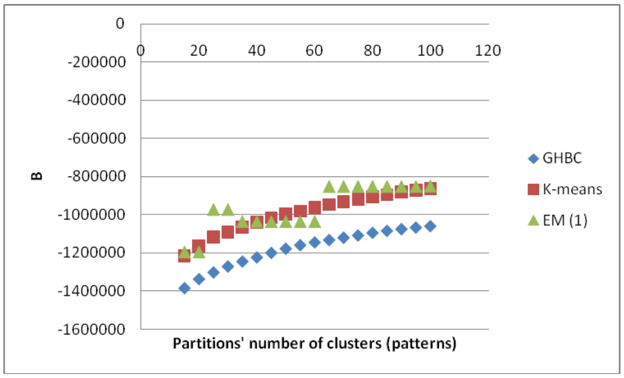

Fig. 6: *Applying EM clustering in the GHBC output also get a gain in similarity.*

In figure 7 and 8 the difference of the algorithms is evident. GHBC, like any other hierarchical algorithm, maintains the form of the initial clusters, and only put the most similar ones together. K-means and EM clustering were used to reclassify the output of GHBC, and the final result obtained was very different. K-means appears to have a tendency to result in clusters more concentrated in space, where most of the clusters that were not well-defined in the initial GHBC result show a better definition with K-means. The final result with EM clustering was different, where some clusters were only slightly better defined, like in K-means, but some clusters changed significantly, but in a consistent way. Neither of the results seems to be much better than the others, but if the computational cost is accounted into, using GHBC and EM clustering shows a better cost-benefit, because EM clustering is computationally faster than K-means and results in something different but consistent with the data set, with results that seem to be better than the original pattern calculated by GHBC.

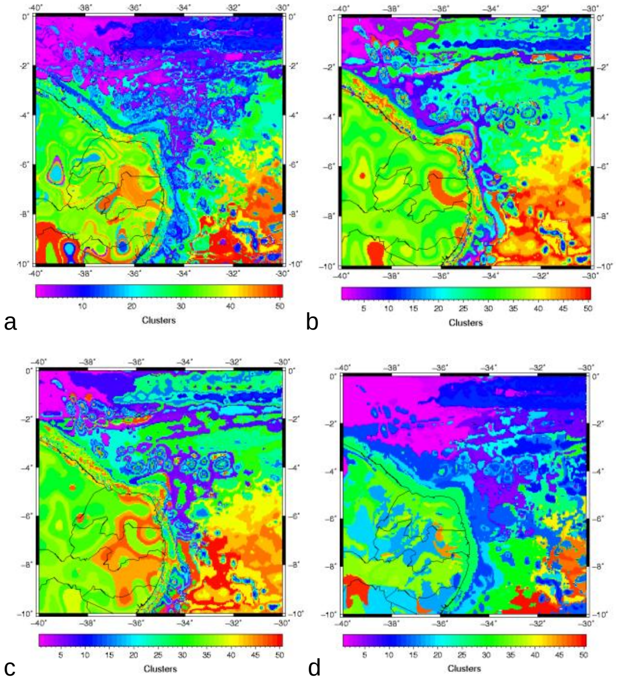

Fig. 7: a) *original pattern calculated by GHBC with 50 clusters,* ***b)*** *this pattern after apply K-means using Euclidian distance,* ***c)*** *this same pattern using K-means with Manhattan distance,* ***d))*** *this pattern using EM-clustering.*

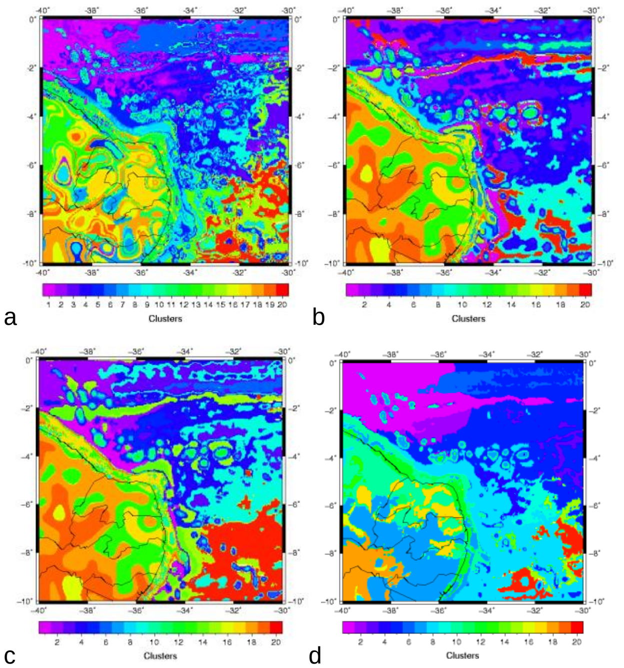

Fig. 8: a) *original pattern by using GHBC with 20 clusters;* ***b)*** *this pattern after applying K-means using Euclidian distance,* ***c)*** *this pattern using K-means with Manhattan distance,* ***d)*** *this pattern using EM-clustering.*

Conclusions

The obtained results demonstrate that the clustering techniques can be used with success on gravity databases. The use of a hierarchical technique (GHBC) provides the most important and interesting result obtained in this work, and the estimated hierarchy brings a huge piece of information for analysis, at different detail levels. The posterior application of partitional methods results in more interesting material for analysis, but, as expected, they maintained the big structures found in GHBC. These techniques demonstrated to be a great guide to the analyst, helping one to easily notice the relevant structures into the data set.

Other advantage of these methods is that they provide not only maps of the clusters but clusters with a quantitative description, with a mean vector and a covariance matrix, forming a normal distribution model for each cluster (see fig. 9 an example of different Gaussian models to clusters).

The results obtained here suggest that the Data Mining techniques can turn to be a very important tool in geophysical studies.

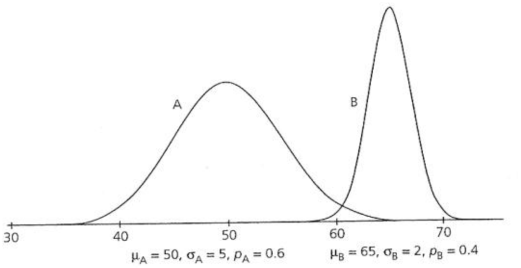

Fig. 9: *Gaussian models for two different clusters (Witten, 2000).*

Acknowledgments

This work was carried on with FAPESP grant 2009/14944-5. The figures were made using GMT software (Wessel & Smith, 1998).

References

Aldridge, M (2005). Clustering: An Overview. In *Lectures Notes in Data Mining*, 2005, p.106. Edited by M. W. Berry & M. Browne. University of Tennessee, USA.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). Optics: Ordering points to identify the Clustering structure. In Delis, A., Faloutsos, C., & Ghandeharizadeh, S., editors. Sigmod 1999, proceedings ACM Sigmod International Conference on Managemente of Data, june 1-3, 1999, Philadephia, Pennsylvania,USA, pages 49-60. ACM Press.

Banfield, J. D., Raftery, A. E., 1993. Model based Gaussian and non-gaussian Clustering. Biometrics, vol. 49: 803-821.

Christ, R. E., Villanueva, E. R., Maciel, C. D., 2007. Gaussian Hierarchical Bayesian Clustering algorithm. Proceedings of The seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007), Rio de Janeiro(RJ), Brazil, 2007. In press.

Dasgupta, A.; Raftery, A. E., 1998. Detecting features in spatial point processes with clutter via model-based Clustering. American Statistical Association, vol. 93: 294-302.

Everitt, B. S., Landau, S. & Leese, M, 2001. Cluster Analysis. Oxford University Press Inc., New York, USA, 4 edition.

Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, 1996. The kdd process for extracting useful knowledge from volume of data. In *Communications of the ACM*, 39(11), 27- 34.

Iwayama, M., & Tokunaga, T., 1995. Hierarchical bayesian Clustering for automatic text classification. International Joint Conference on Artificial Intelligence, vol. 2:1322-1327.

Jain, A. K., Murty, M. N., & Flynn, P. J., 1999. Data Clustering: a review. ACM Computting Surveys, vol. 31(3):264-323.

Metz, J., 2006. Interpretação de clusters gerados por algoritmos de Clustering hierárquico. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – USP São Carlos.

Mitchell, T. M., 1997. Machine Learning. WCB McGraw-Hill. Pyle, D., 1999. Data preparation for data mining. San Francisco, CA, USA: Morgan Kaufmann Publisher Inc.

Murtagh, F., Raftery, A. E. (1984). Fitting Straight Lines to Point Patterns. Pattern Recognition, vol.17(5): 479-483.

Rezende, S. O., J. B. Pugliese, E. A. Melanda, & M. F. Paula, 2003. Mineração de dados. Em S. O. Rezende (ed.), Sistemas Inteligentes: Fundamentos e Aplicações, 1, 307-335. Barueri, SP: Editora Manole.

Shen, Z. Distance-Based Algorithms. Lectures Notes in Data Mining, 2005, p.73. Editado por M. W. Berry & M. Browne. University of Tennessee, USA.

Villanueva, E. R. (2007). Métodos Bayesianos aplicados em taxonomia molecular. Dissertação de Mestrado, Escola de Engenharia de São Carlos – USP São Carlos.

Wessel, P. & Smith, W.H.F., 1998. New, improved version of Generic Mapping Tools released, EOS Trans. Amer. Geophycs. U., vol. 79(47): 579.

Witten, I. H., Frank, E., 2000. Data Mining: practical machine learning tools and techniques with Java. Academic Press: Morgan Kaufmann Publishers.