On Model Based Clustering in a Spatial Data Mining Context

Gabriella Schoier and Giuseppe Borruso

DEAMS – Department of Economic, Business, Mathematic and Statistical Sciences "Bruno de Finetti", University of Trieste, Via A. Valerio, 4/1 – 34127 Trieste, Italy
{gabriella.schoier,giuseppe.borruso}@econ.units.it

Abstract[1]**.** In this paper we present the finite mixture models approach to clustering of high dimensional data. The mixture resolving approach to cluster analysis has been addressed in a number of different ways; the underlying assumption is that the patterns to be clustered are drawn from one of several distributions, and the goal is to identify the parameters of each and (perhaps) their number. Finite mixture models allows a flexible approach to the statistical modeling of phenomena characterized by unobserved heterogeneity in different fields of applications. In this analysis we consider the model based clustering on mixture models and compare it with the classical k-means approach. The application regards some aspects of the 218 Municipalities of the region Friuli Venezia Giulia in North-Eastern Italy with data based on the Italian population 2011 Census.

Keywords: Model based clustering, Finite mixture modeling, EM algorithm, Spatial data mining, GIS, Friuli Venezia Giulia Region, Italy.

1 Introduction

Clustering is the unsupervised classification of patterns - observations, data items, or feature vectors - into groups or clusters. The clustering problem has been considered in many contexts and by researchers in different disciplines. It is useful in several exploratory pattern-analysis, grouping, decision-making and machine-learning situations, including data mining, spatial data mining, document retrieval, image segmentation, and pattern classification.

Spatial data mining can be used for browsing spatial databases, understanding spatial data, discovering spatial relationships, optimizing spatial queries.

Clustering techniques have been recognized as primary Data Mining methods for knowledge discovery in spatial databases, i.e. databases managing 2D or 3D points, polygons etc. or points in some d-dimensional feature space (see e.g. [8]).

[1] The paper derives from joint reflections of the two authors. Gabriella Schoier realized paragraphs 1, 2, 3.2, 3.3 and 4, while Giuseppe Borruso wrote paragraphs 3.1 and 3.4.

The geographical visualization and analysis, where not otherwise specified, have been realized using Intergraph Geomedia Professional and Geomedia GRID 6.1 under the RRL (Registered Research Laboratory) agreement between Intergraph and the University of Trieste (Italy).

B. Murgante et al. (Eds.): ICCSA 2013, Part IV, LNCS 7974, pp. 375–388, 2013.

Cluster analysis can be defined as the organization of a collection of patterns - usually represented as a vector of measurements, or a point in a multidimensional space - into clusters based on similarity.

Different clustering algorithms have been proposed (see e.g. [7]); however several clustering methods have been criticized due to the lack of theoretical robustness both from a mathematical and a probabilistic point of view. For this reason model based clustering - which can be defined as clustering procedures based on finite mixture models - are being increasingly preferred over heuristic methods ([14], [9]). This type of models can be used in different fields concerning density estimation and clustering to high-dimensional data too (see e. g. [5], [15]).

In this analysis we consider the model based clustering on mixture models in a spatial data mining context ([3]) and compare it with the classical k-means approach. The application regards some aspects of the 218 Municipalities of the region Friuli Venezia Giulia while the data regards the Census of the Italian population of 2011.

2 The Model Based Clustering

The base assumption of these clustering methods is that the data are assumed to have generated by an underlying mixture of a finite number of distributions. The objective is to identify the parameters of each of them and their number. Usually the assumption is to take the component distributions to be multivariate normal ([1]). The basic concept of model based clustering is that of mixture model (see [10], [12]).

Given $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ a random sample of size n, dove $\boldsymbol{Y}_j$ is a p-dimensional a random vector with density probability function $f(\boldsymbol{y}_j)$ on $\mathbb{R}^p$. Let $\boldsymbol{Y}$ be a random vector consisting of p features

$$\boldsymbol{Y} = (\boldsymbol{Y}_1{}^T, \dots, \boldsymbol{Y}_n{}^T)^T$$

while let

$$\boldsymbol{y} = (\boldsymbol{y}_1{}^T, \dots, \boldsymbol{y}_n{}^T)^T$$

be an observed random sample of size n on $\boldsymbol{Y}$.

With the finite mixture model-based approach to density and clustering ([13]), the density $f(\boldsymbol{y}_j)$ of $\boldsymbol{Y}_j$ (one of the g density components of the mixture) can be written as:

$$f(\boldsymbol{y}_j) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}_j) \tag{1}$$

where $f_i(\boldsymbol{y}_j)$ are the component densities of the mixture and π_i are some unknown proportions such as:

$$0 \leq \pi_i \leq 1 \quad (i = 1, \dots, g)$$

$$\sum_{i=1}^{g} \pi_i = 1.$$

The number of components g can be taken sufficiently large to provide accurate estimate of the underlying density function ([4]). For clustering purpose each of the g components correspond to a cluster.

The posterior probability that an observation, on which $\boldsymbol{y}_j$ has been observed, belongs to the i-th component of the mixture is

$$\tau_i(\boldsymbol{y}_j) = \pi_i f_i(\boldsymbol{y}_j)/f(\boldsymbol{y}_j) \quad \text{for } i = 1, \dots, g;\ j = 1, \dots, n. \tag{2}$$

A probabilistic clustering of the data in g clusters can be obtained in terms of the fitted posterior probabilities of component membership for the data as given in (2).

It is possible to obtain a partition of the observations in g non-overlapping clusters $C_1, \dots, C_g$ assigning each observation to the component to which it has the highest estimated posterior probability of belonging. In this way the i-th cluster C_i contains all the observations assigned to group G_i.

Formally C_i contains those observations j with $\hat{z}_{ij} = (\hat{\boldsymbol{z}}_j)_i = 1$, where

$$\hat{z}_{ij} = \begin{cases} 1 \ if\ \hat{\tau}_i(\boldsymbol{y}_j) \geq \hat{\tau}_h(\boldsymbol{y}_j)\ \ (i, h = 1, \dots, g; h \neq i\ j = 1, \dots, n) \\ 0 \qquad\qquad otherwise \end{cases}$$

with $\hat{\tau}_i(\boldsymbol{y}_j)$ an estimate of $\tau_i(\boldsymbol{y}_j)$.

According to this notation $\hat{z}_{ij}$ can be viewed as an estimate of z_{ij} which, under the hypothesis that the observations come from a mixture of g groups $G_1, \dots, G_g$, is defined to be one or zero accordingly to the fact that the j-th observation does or does not come from G_i $(i = 1, \dots, g; j = 1, \dots, n)$.

The model can be fitted to the data using the maximum likelihood estimation method implemented via the *EM* (*Expectation Maximization*) algorithm ([2], [11]).

3 The Application Results and Discussion

3.1 The Study Region

The Friuli Venezia Giulia Region is located in North-Eastern Italy at the State border with Austria (North) and Slovenia (East). It hosts around 1.2 million people distributed over 219 municipalities (218 used in the present analysis). Most of the people live actually in few municipalities, these including the four Province capitals - Trieste, Udine, Pordenone and Gorizia - followed by Monfalcone, hosting the majority of population.

From a geomorphological point of view, the Region is divided in three areas, from North to South. Nearly one third of the surface is covered by a mountain area, followed by a series of hills and heading South with a flatland ending with the coastline. A major system of river follows mainly a North-South direction a part from the two origins in the extreme Eastern and Western corners of the Region. The five major urban areas mentioned above are located in the flatland in the Southern part of the Region and host one third of the total population. Transport infrastructures follow

mainly a West - East direction and also a North – South one, this latter following the major rivers flow in the mountain area. The human settlements as well as transport infrastructure can be spotted mainly located in the hill and flatland areas and therefore in the Central and Southern parts of the Region.

3.2 The Data and Some Preliminary Analysis

The database used was realized by the Research Centre of the Udine Chamber of Commerce from Italian National Statistics Institute (ISTAT) "Geodemo" data and contains qualitative and quantitative variables on demography, economy and income at municipality level for the Friuli Venezia Giulia Region. The data are detailed at municipality level and report figures on birth and death rates, average age, ageing and population growth forecasts. From such data a matrix for the 218 municipalities of Friuli Venezia Giulia was realized.

The database contains information on municipalities as surface in sq. km, population at 2011, divided in age classes, foreign people. Data were processed to obtain some population indices presented as follows:

- Density (in *R* programme, used for the application, appearing as den). Where *P (x)* is the population in the Municipality at year x, while S is the surface in sq. Km.

$$D\,(2011) = P(2011) / S$$

- Birth rate (nat). The birth rate can be expressed through the following formula:

$$n(x) = [N(x) / (P(x) + P(x\text{-}1))/2] \ast 1000$$

Where *N (x)* is the number of living people borne in the year *x*. Such value is actually estimated using an average value of the latest available years, as the updated value is not always available.

- Death rate (mor). Similarly as in the birth rate, it is realized using the average number of death in the last few years (T(x) is the number of dead people in the year *x*)

$$n(x) = [N(x) / (P(x) + P(x\text{-}1))/2] \ast 1000$$

- Percentage of young population. $P_{0\text{-}14}$ is the population 0-14 years old.

$$g(2011) = [P_{0\text{-}14}\,(2011) / P(2011)] \ast 100$$

- Ageing index (vec). Over 65 years old people over the population 0-14 years old.

$$v(2011) = [P_{65+}\,(2011) / P_{0\text{-}14}\,(2011)] \ast 100$$

- Variation of population between 2001 and 2011 (var).

$$D(2011, 2001) = [P\,(2011)/P\,(2001)]\text{- } 1$$

- Percentage of foreign people (str).

$$s(2011) = [S(2011)/P(2001)] \ast 100$$

Such variables were used for a preliminary analysis that highlighted some characters of the Region Friuli Venezia Giulia: 75% of the Municipalities holds a population density

lower than 220 inhabitants per square kilometre (Tab. 1), an average value of 186 that is lower than the Italian figures (201.7 inhabitants per square kilometre). Only the two main municipalities hosting the provinces' capitals as Trieste and Udine present the highest densities, respectively 2398 and 1734 inhabitants per square kilometre.

With reference to the variations intervened between the two census (2001 and 2011) only few municipalities present significant changes with a decrease in population higher than 25% (Drenchia, Ligosullo, Savogna and Dogna) while in other cases an increase of population can be registered (Vajont, Martignacco and Pravisdomini, with respectively an increase of 25%, 26% and 35%).

The growth in these municipalities is partly justified by an increase of the foreign population. On average the percentage of foreign population in Friuli Venezia Giulia Region is around 6 % - in line with the Italian value – but reaching 20% in some municipalities, particularly in the Western part of the Region in the Province of Pordenone: Pravisdomini (22.1%), Prata di Pordenone (20%), Vajont (19.6%), Pasiano di Pordenone (18.2%).

Table 1. - Summary of demographic variables

	Den	**var**	**nat**	**Mor**	**str**	**gio**	**vec**
	Barcis	Drenchia	Tramonti di Sopra	Vivaro	Ligosullo	Drenchia	Vajont
Minimum	2.54	-31.98	1.86	5.94	0.00	4.48	76.24
1° quartile	43.01	-2.92	6.98	9.97	3.40	11.48	160.54
Median	111.26	1.69	8.00	11.40	5.07	12.55	184.17
Average	186.57	1.32	7.95	12.37	6.01	12.34	209.67
3° quartile	220.13	6.72	9.27	13.77	7.48	13.65	216.37
Maximum	2397.98	34.74	14.91	39.80	22.15	17.60	1166.67
	Trieste	Pravisdomini	Vajont	Drenchia	Pravisdomii	Pravisdomini	Drenchia

When we observe the demographic indicators on age, we can notice that that ageing index is higher than the national, Italian Value (145). In some municipalities it reaches very high values, particularly in municipalities located into mountain areas (Drenchia, 1166.67; Tramonti di Sopra, 794.44; Barcis, 635.71; followed by Andreis, Dogna and Rigolato with values always higher than 500). In a parallel way, the presence of young people is lower than the national average. The Italian birth and death rates are respectively 9.1 and 9.7 and the regional values are lower in terms of birth rate while the death rate is higher than the national value. Crossing data we can notice a positive correlation between the birth rate, the population variation in the two census and the percentage of young population. A negative correlation can be found in the death and ageing rates. The higher values can be observed in the variables gio and vec (-0.854), gio and var (0.796) and mor and var (-0.737),(see Tab. 2).

Table 2. - Correlation matrix of the demographic variables

	Den	**Var**	**Nat**	**Mor**	**Str**	**gio**	**Vec**
den	1.000	0.305	0.269	-0.220	0.364	0.255	-0.193
Var	0.305	1.000	0.647	-0.737	0.451	0.796	-0.684
Nat	0.269	0.647	1.000	-0.420	0.505	0.644	-0.462
mor	-0.220	-0.737	-0.420	1.000	-0.250	-0.753	0.786
Str	0.364	0.451	0.505	-0.250	1.000	0.397	-0.250
Gio	0.255	0.796	0.644	-0.753	0.397	1.000	-0.854
vec	-0.193	-0.684	-0.462	0.786	-0.250	-0.854	1.000

3.3 Model Based versus K-means Clustering

All the demographic variables presented in the previous paragraph have been considered in the present application.

After the standardization of the variables the Mclust package of the *R* language has been used for the analysis (see [6]). In order to compare all the different models the BIC criterium has been applied. The results are presented in Tab. 3 and Fig. 1

As one can see from the previous table and from the next figure the best values for the BIC criterium regards model “VVV” with four components (-2670,388), model “VEV” with three components (-2711,766) and “VVV” with five components (-2727,848). The model choice is on the one with four components without restrictions as regards shape, dimension and orientation.

Table 3. - BIC values for the different models

	EII	VII	EEI	VEI	EVI
1	-4372.616	-4372.616	-4404.923	-4404.923	-4404.923
2	-3995.881	-3720.653	-3962.963	-3637.700	-3456.717
3	-3801.458	-3474.236	-3798.156	-3355.705	-3185.104
4	-3650.510	-3357.338	-3660.431	-3272.679	-3010.504
5	-3491.491	-3305.922	-3431.069	-3077.728	-2963.104
6	-3418.906	-3283.652	-3342.953	-3040.783	-2924.712
7	-3408.262	-3276.164	-3344.259	-3221.692	-2845.884
8	-3415.155	-3215.810	-3349.613	-3042.233	-2890.519
9	-3433.028	-3224.974	-3379.398	-3027.352	-2841.798
	VVI	EEE	EEV	VEV	VVV
1	-4404.923	-3482.667	-3482.667	-3482.667	-3482.667
2	-3571.534	-3344.251	-3014.889	-2834.307	-2860.717
3	-3247.110	-3346.602	-2804.079	-2711.766	-2751.574
4	-2890.298	-3167.857	-2763.850	-2752.092	-2670.388
5	-2845.976	-3030.844	-2756.196	-2826.174	-2727.848
6	-2837.431	-3069.793	-2961.225	-2872.833	-2880.964
7	-2868.480	-2979.938	-3049.321	-2952.191	-3020.440
8	-2857.736	-3027.461	-3116.015	-2999.034	NA
9	-2895.136	-3047.632	-3123.694	-3117.066	NA

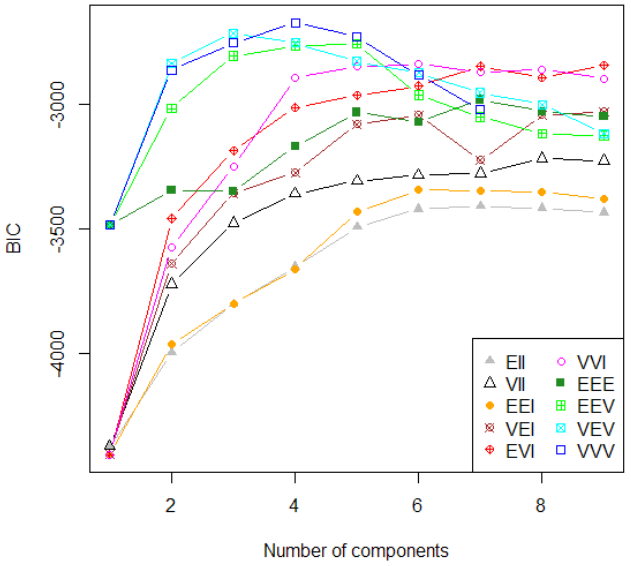

Fig. 1. BIC values for the different models and for different variance and covariance structure of the components

The parameters for the chosen mixture model estimated through the EM algorithm: the mean and the mixture proportion are reported in Tab. 4

Table 4. - Parameters values

```
$pro
[1] 0.45412844 0.31651376 0.16513761 0.06422018

$mean
            [,1]       [,2]       [,3]       [,4]
den -0.26668687  0.1884928 -0.6313611  2.5829286
var -0.05448485  0.6607826 -1.5153333  1.0242857
nat -0.29459596  0.6294058 -0.8750833  1.2315714
mor -0.22744444 -0.3805652  1.5655833 -0.5397143
str -0.37667677  0.4462609 -0.4846944  1.7119286
gio  0.01517172  0.6674928 -1.6165833  0.7590714
vec -0.15997980 -0.4481739  1.4668056 -0.4282143
```

In Tab. 5 some summary results of the clustering obtained using and in Fig. 1 a geographical representation are presented.

Table 5. – Dimensions and mean values of the clusters obtained using the model-based clustering

	N	**Den**	**Var**	**Nat**	**mor**	**str**	**gio**	**vec**
Cl 1	99	113,62	0,801	7,384	11,406	4,596	12,328	192,272
Cl 2	69	238,11	7,680	9,165	10,757	7,690	13,712	160,954
Cl 3	36	13,91	-13,250	6,267	18,992	4,190	9,068	368,976
Cl 4	14	892,45	11,175	10,324	10,084	12,448	14,001	163,139
Pop	***218***	***186,57***	***1,32***	***7,96***	***12,37***	***6,01***	***12,34***	***209,67***

In details in the first cluster (Cl. 1) there are 99 municipalities (75 in the province of Udine) having a mean density of 114 inhabitants/ sq. km. These municipalities have not registered significant variations in the population in the Census 2011; the birth rate and death rate, the percentage of young population and the ageing index are near the mean values registered for the whole region Friuli Venezia Giulia. The percentage of foreign people is in general less than the regional mean value. To this cluster belong municipalities like Aiello del Friuli, Codroipo, Osoppo e Villesse.

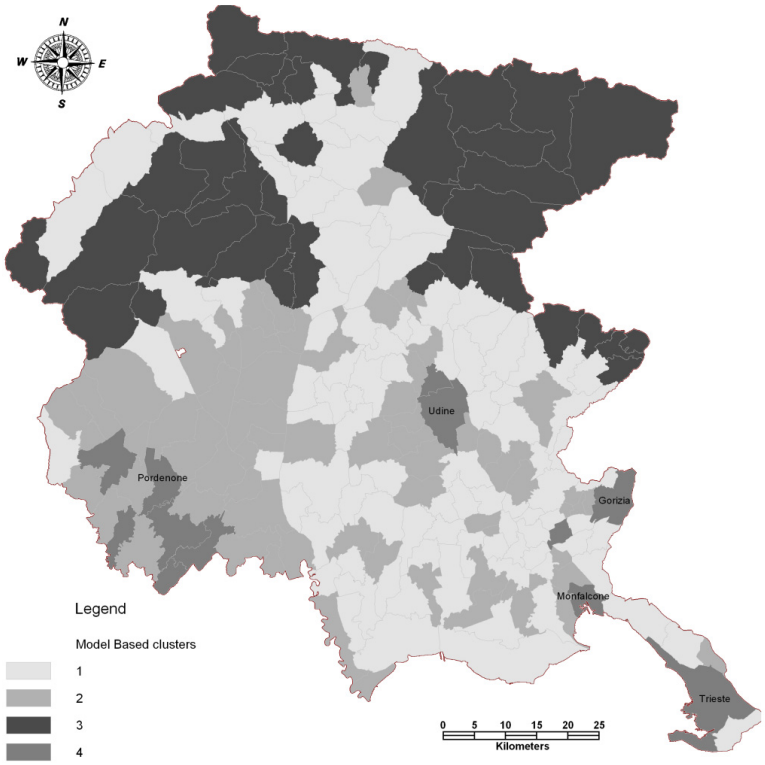

Fig. 2. - The clusters obtained using the model-based clustering

The second cluster, (Cl. 2), is formed by 69 municipalities having a mean of 238 inhabitants per square kilometre. This cluster è is characterized by a mean increase in the population with respect to 2001 of 7,7%: coherently with this value there are a percentage of foreign people, a birth index and a percentage of young population in mean slightly higher than the regional mean.

Obviously the death and ageing indices present lower values. Belong to this cluster the municipalities of Cervignano del Friuli, Palmanova, Maniago, Spilimbergo, Staranzano.

The third cluster (Cl.3), in formed by 36 observations, in this cluster there are the municipalities with the lowest population density and with a reduction intervened between the two census (2001 and 2011) of less than 10%.

The death (18.99) and ageing (368.98) indices are very high in mean while the other variables present mean values lower than those observed in the other municipalities. Belong to this cluster the almost all the municipalities located into mountain areas.

The fourth cluster (Cl.4) is formed by 14 municipalities with mean vs high density population. This cluster has opposite characteristics with respect to the previous cluster. In this cluster there are higher values as regards the variation intervened between the two census (2001 and 2011) (+11.2%), Percentage of young population (14%) percentage of foreign people (12.4%) and the birth index (10.3). Belong to this cluster the fourth province's capitals and the municipalities of Tavagnaco, Azzano Decimo, Chions, Fontanafredda, Prata di Pordenone, Pravisdomini, Vajont, Gradisca d'Isonzo, Monfalcone, Muggia.

Now we consider the results obtained with the *K-means* algorithm and compare these with the previous one.

Looking at Fig. 3 one can see that a partition with seven clusters would be the best one for describing the characteristics of our population but in order to compare the obtained partition with that of the *model-based*, the results of a *K-Means* clustering with four groups are reported.

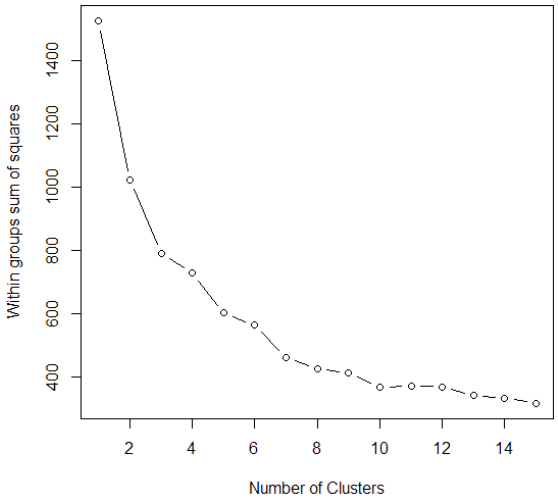

Fig. 3. – Within clusters residual Sum of Squares (WSS)

As one can see from the next table and Fig. 4 there is a clear difference between the dimension of the clusters; for instance the fourth cluster obtained with the model based clustering contains fourteen observation while with the *K-means* five observations are included these are municipalities with high density population; the municipalities of the Pordenone with the highest percentage of foreign people belong now to the second cluster. The fourth cluster is very similar to the third one found with model-based clustering; the first cluster has passed from 99 to 127 observations keeping more of less the same characteristics.

Table 6. - Dimensions and mean values of the *clusters* obtained using the *model-based clustering*

	N	Den	Var	Nat	mor	str	gio	vec
Cl 1	127	144.17	1.035	7.589	11.813	4.751	12.428	190.419
Cl 2	53	264.30	10.979	9.858	9.404	9.231	14.221	145.582
Cl 3	5	1569.63	5.875	10.019	11.520	14.985	13.346	184.715
Cl 4	33	15.38	-13.758	5.976	19.392	4.340	8.794	390.461
Pop	***218***	***186.57***	***1.32***	***7.96***	***12.37***	***6.01***	***12.34***	***209.67***

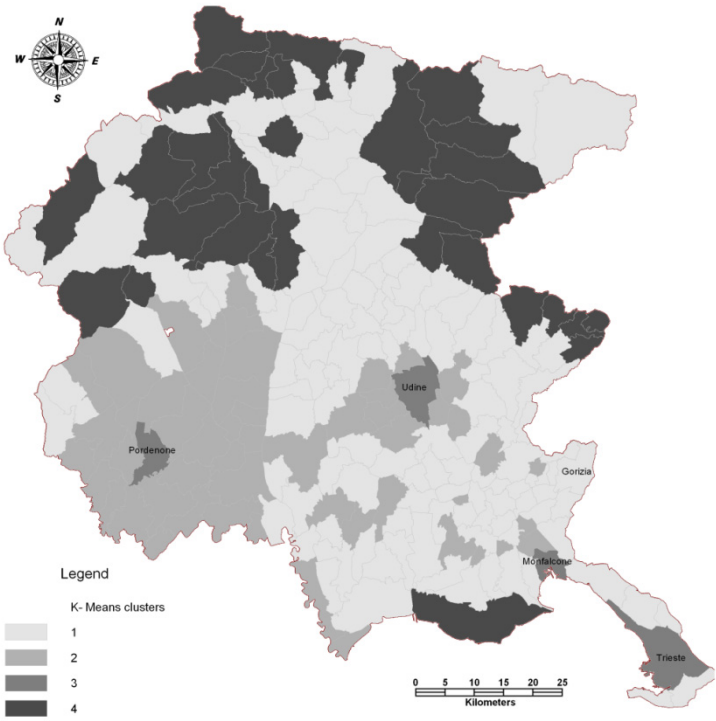

Fig. 4. - The clusters obtained using the k-means

3.4 A Spatial Comparison

Some interesting results can be observed observing in particular the pattern drawn after running the Model Based clustering algorithm and comparing the results with those obtained through the more 'traditional' K-Means one.

The analysis carried on revealed the presence of four clusters using both the algorithms. Very similar patterns arise, although the model based algorithm seems providing more refined and appreciable results. The four clusters deriving from the model based clustering present the following characteristics. Cluster 1 collects 99 municipalities, all presenting average values for all variables. In such set we can highlight municipalities located in all the areas of the region, mainly referred to medium-little dimensions and characterized by a good accessibility when located in mountain areas. However they are quite far from main centres and urban areas.

Cluster 2 is composed by 69 municipalities and is characterized by values higher than the average mainly in terms of density, variation of population in the two census, birth rates, presence of foreign nationals, young people. They are characterized mainly as peri-urban municipalities, located in the 'belts' around major centres and mainly located in the flatland areas. Cluster 3 counts for 36 municipalities and holds the higher values in terms of ageing population and death rate. The municipalities entirely belong to the mountain area. Cluster 4 presents the 14 municipalities with the higher values of population density, variation of population, birth rate, and presence of foreign nationals and younger people. Here we can notice an interesting mix of

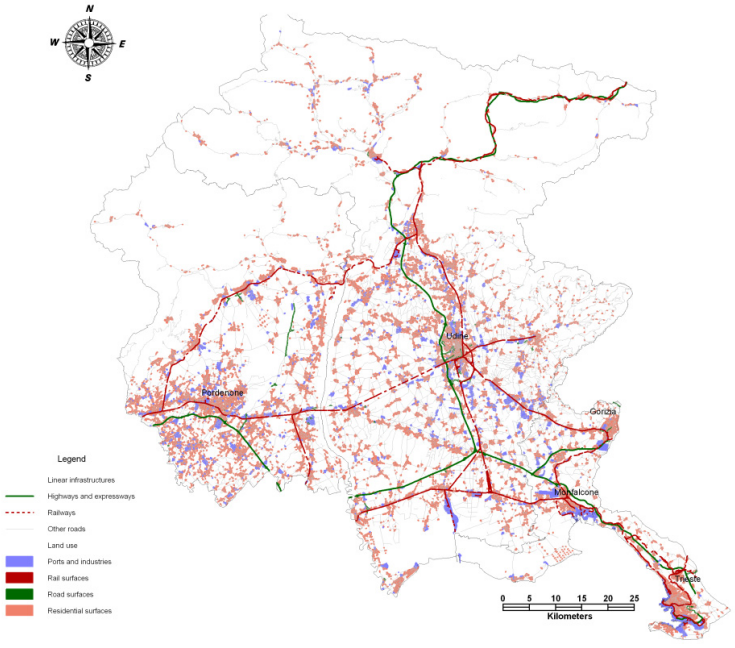

Fig. 5. - An 'infrastructure' view of the Region Friuli Venezia Giulia

municipalities including the four provinces' capitals and major municipalities, together with those municipalities considerable as belonging to the urban areas of such major cities. In Pordenone Province in the Western part of the Region we can notice a strong presence of such municipalities, also in an area that is characterized by an urban sprawl, although we are facing smaller dimensions than other urban areas where sprawl is considerable. The derived image is therefore consistent with what represented and visible in an 'infrastructure' view of the Region (Fig. 5), where particularly we can notice the Cluster 4 municipalities as overlapping with the areas presenting the higher level of urban and infrastructure density.

The results obtained and visualized in Fig. 2 after running the model based algorithm are consistent with the k-means approach, visible in Fig. 4. Here some differences can be noticed. Particularly a less disaggregated picture is visible, with cluster 4 hosting just four municipalities chosen among the most numerous and densely populated: in any case a consistent picture but demonstrating the better performances of the model based approach.

4 Conclusions

In this analysis we consider the model based clustering on mixture models in a spatial data mining context ([3]) and compare it with the classical k-means approach. The application regards some aspects of the 218 Municipalities of the region Friuli Venezia Giulia while the data regards the Census of the Italian population of 2011.

The results obtained after running the model based algorithm are consistent with the k-means approach. The classification obtained demonstrates the better performances of the model based approach.

Moreover the choice of model based clustering is supported from a theoretical point of view. In fact several clustering methods have been criticized due to the lack of theoretical robustness while the model based clustering which can be defined as clustering procedures based on finite mixture models have a strong mathematical and a probabilistic background. This type of models can be used in different fields concerning density estimation and clustering to high-dimensional data too.

References

1. Banfield, J.D., Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. Biometrics 49, 803–821 (1993)
2. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B (Methodological) 39(1) (1977)
3. Dasgupta, A., Raftery, A.E.: Detecting features in spatial point processes with cluster via model-based clustering. Journal of the American Statistical Association 93, 294–302 (1988)
4. Fraley, C., Raftery, A.E.: How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal 41(8) (1998)

5. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis and density estimation. Journal of the American Statistical Association 97(458) (2002)
6. Fraley, C., Raftery, A.E.: MCLUST Version 4 for R: normal mixture modeling and model-based clustering, classification and density estimation. Technical Report no. 597, Department of Statistics, University of Washington (2012)
7. Fung, G.: A comprehensive overview of basic clustering algorithms (2001), http://pages.cs.wisc.edu/~gfung/ (cited October 2012)
8. Han, J., Kamber, M., Tung, A.K.H.: Spatial clustering methods in data mining. A survey (2001), http://www.cs.uiuc.edu/homes/hanj/ (cited December 2012)
9. Ingrassia, S., Minotti, S., Vittadini, G.: Local statistical modeling via a cluster-weighted approach with elliptical distributions. Journal of Classification 29(3) (2012)
10. Lindsay, B.G.: Mixture models: theory, geometry and applications. In: NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics and American Statistical Association, vol. 5 (1995)
11. McLachlan, G.J., Krishnan, T.: The EM algorithm and extensions. Wiley, New York (1997)
12. McLachlan, G.J.: Model-based clustering (2007), http://www.maths.uq.edu.au/~gjm/ (cited October 2012)
13. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
14. McLachlan, G.J., Bean, R., Ng, S.K.: Clustering. Bioinformatics: Structure, Function, and Applications 2, 423–439 (2008)
15. McLachlan, G.J., Ng, S.K., Wang, K.: Clustering of high-dimensional and correlated data. In: Lauro, C., Palumbo, F., Greenacre, M. (eds.) Studies in Classification, Data Analysis, and Knowledge Organization: Data Analysis and Classification, pp. 3–11. Springer, Berlin (2010)