Rural Development Statistics (RDS) for Policy Monitoring: a Rural-Urban Territorial Classification and Farmers' Income Data

Edoardo Pizzoli
FAO, Food and Agriculture Organization of United Nations, Rome, Italy
ISTAT, National Institute of Statistics, Rome, Italy[1]

Gabriele Palmegiani
Tor Vergata University, Rome, Italy [2]

Abstract

National Institutions and International organizations classify administrative areas, basically, for territorial policy monitoring and decision making. The identification and the supervision of different groups of regions, becomes the starting point to implement specifics economics reforms which may allow to governments a significant improvement of economic conditions at territorial level. Furthermore, date on income at regional level, better if collected and available at households and farmers' level, are essential for analysis. Using cross-section data (2004) for Italy, Morocco, Tunisia and Senegal on residential population, educational level and agricultural employement, regions are explicitly classified demonstrating that it is possible to gain a statistical accuracy considering more than two groups of regions. After that, it is investigated whether the optimal numbers of clusters of regions is predictive of a proxy of farmers' income.

Keywords

Rural-Urban regions, Proxy variables, Normality, Outliers, Cluster Analysis, K-means clustering, Analysis of variance (ANOVA)

[1]Senior Statistician, contact: pizzoli@istat.it
[2]FAO volunteer, contact: gabriele.palmegiani@gmail.com

Introduction: why a territorial statistical classification and income data?

Rural Development Statistics (RSD) is becoming more and more important for policy monitoring at different decision making's levels. At United Nation (UN) level the policy objectives indicated in the Millennium Development Goals (MDGs) clearly suggest that it is necessary to monitor variables not just at the country level but also at the regional and local levels. There are statistically identifiable territorial classifications that are significant to understand issues relevant to policymakers.

Furthermore, to collect farmers' income and their households' total income to have information on rural households' wellness can be a good practical solution with some advantages in terms of measurement. This is because incomes from agricultural and related activities are the most difficult to calculate, so the task will be easier if it is joint to normal data collection for this specific sector of activity.

In a multivariate data framework, to classify means to find out properties or features which may belong to only one or more than one observation and to group them in two or more homogenous clusters. The methods which allow researchers to discover groups in the data are called cluster methods. In this specific application, observations are administrative subdivisions of a country.

From the properties of the groups found, the policymakers can formulate specific treatments or advices, to improve the conditions of the objective population. In botany, for example, discovering that flowers in a sample differ significantly in their properties means that, likely, there are a large number of species or groups. Then, if a researcher wants promote the growth of all the flowers he should treat them in a differentiate way (differentiate the level of water used, sun exposure, etc...). In a similar way, cluster analysis can be applied to territorial classification for differentiated development policies actions.

In this paper regional properties will be studied. Common features will be found which allow identifying significantly diffent groups of regions. Policy hints can be formulated according to the clusters' characteristics, as a botanist who promotes the growth of different species of flowers.

Data

Data is collected using the two sources of information available that is from the Nationals Offices of Statistics and from FAO's specifics databases. The constructed data-set consists of eight measurements on each region of Italy, Tunisia, Morocco and Senegal. The total number of regions are 70. The covariates (all in percentage) are as follows:

- **s_pop_d** = share of residential population (regional population density).
- **s_illit** = share of regional illiterates over total population.
- **s_degr** = share of regional graduates over total population.
- **s_dipl** = share of regional high school certified over the total population.
- **s_empl** = share of regional actives employed over total population.
- **s_ag_e** = share of regional employment in agriculture over total population.
- **s_sau** = share of regional harvested surface over regional surface.
- **s_s_y** = the harvested surface's performance. (share of regional yield in agriculture over regional harvested surface).

Since data on regional income, such as gross domestic product (GDP) or aggregates of total regional production, is not available, we shall use the covariate *s_s_y* as a *proxy* of regional income. It implies that a greater harvested surface perfarmance will be interpreted as a signal of a greater amount of income at least for the regional farmers.

In order to express covariates using the same scaling, the share of population density (*s_pop_d*) is defined as the ratio of residential population over the regional surface divided by one hundred.

Given the covariates available, the main questions of interest is whether there is any evidence of different types or clusters of regions. Then, a further issue to be investigated will be whether the clusters of regions found are predictive of regional farmers' income.

Statistical analysis

As a preliminary work we have checked both the *normality* of the covarites and the existence of *outliers*.
To check normality, histogram plots of the frequency distribution have been applied *(Figure (1))*. Since a histogram is a crude form of a density estimate, a smooth line is passed through the rectangles, while for the width default value can work reasonably well. Some covariates, as: *s_pop_d*, *s_empl*, *s_ag_e*

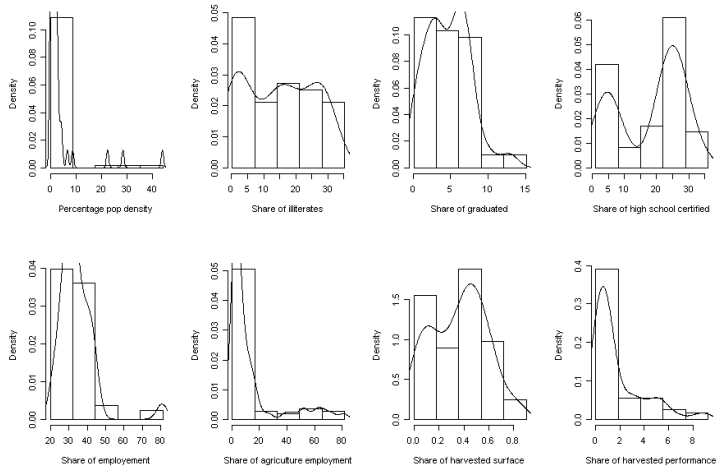

Figura 1: Normality check of the covariates: histogram plot of the frequency distributions.

and *s_s_y*, show a positive skewness in the distribution. Others covariates,

such as: *s_degr*, *s_dipl* and *s_sau* point out a distribution which is closer to a weighted sum of two normals. The estimated distribution of *s_illit* seems to be quite stable.

From this analysis, it can be concluded that the covariates considered do not seem to follow a well defined normal distribution. Some of the distributions look like as a weighed normal distribution, at least. Then, results from t-tests and F-tests should be taken with caution.

To check outliers a scatterplot matrix *(Figure (2))* and the DFBETA-statistic[3] *(Figure (3))* are provided. Outliers are observations which appears to be inconsitent with the remainder of the dataset. The scatterplot matrix *(Figure (2))* evidences that some regions are far from the bulk of the data. Since bivariate scatterplots may not be sufficient to detect outliers, to verify if observations are inconsistent, we have estimated the DFBETA-statistics *(Figure 3)* in a multivariate regression model which treats the harvested performance as dependent variable. Since there are point observations inward or outward the central band[4] we can state that inconsistent point should appear.

Another statistics widely used to check the influence of a data point is the

[3]The DFBETA-statistic for a given covariate x is defined as:

$$DFBETA - statistic = \frac{\hat{\beta} - \hat{\beta}_{(i)}}{s.e(\hat{\beta}_{(i)})}$$

Since $\hat{\beta}$ is the ordinary least squares (OLS) coefficient of covariate x, $\hat{\beta}_{(i)}$ is the OLS coefficient of covariate x when i-th observaton is dropped out and $s.e(\hat{\beta}_{(i)})$ is the standard error of the latter coefficient, this statistic measures the distance that this regression coefficient would shift when the i-th observaton is exluded from the regression, scaled by its standard error.

[4]It has been suggested (Belsley D. A. & oth.(1980)) that point observations with:

$$|DFBETA| > 2/\sqrt{n}$$

take special attention. For example *(Figure 3)*, for the share of active employment in agriculture, the Veneto region is a downward influential point. It may be a possible outliers, at least, for this regressor.

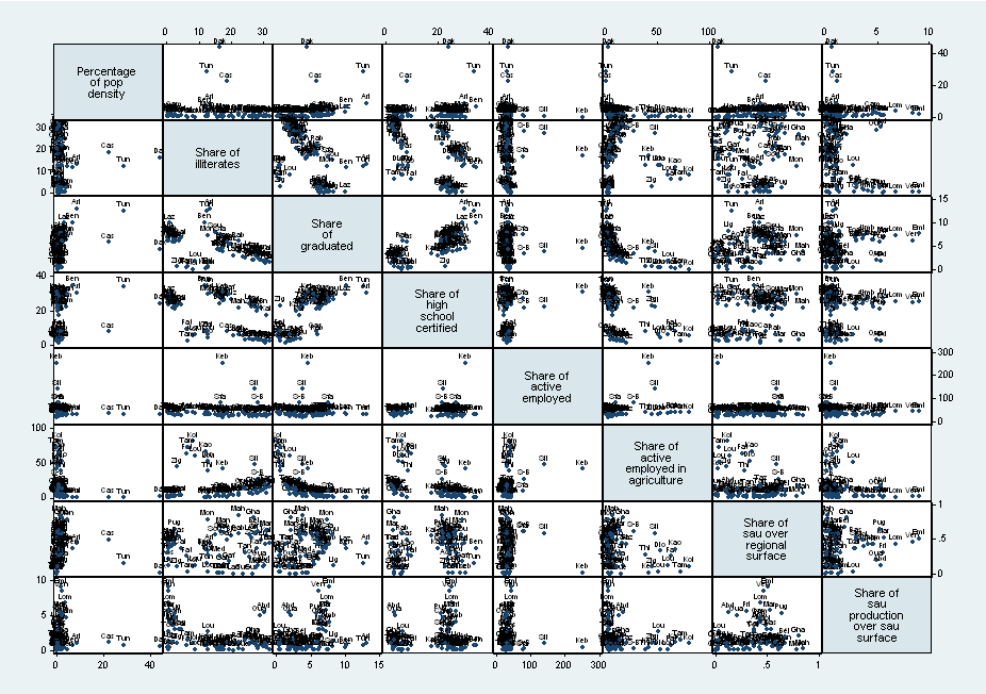

Figura 2: Outliers check of the covariates: the scatterplot matrix.

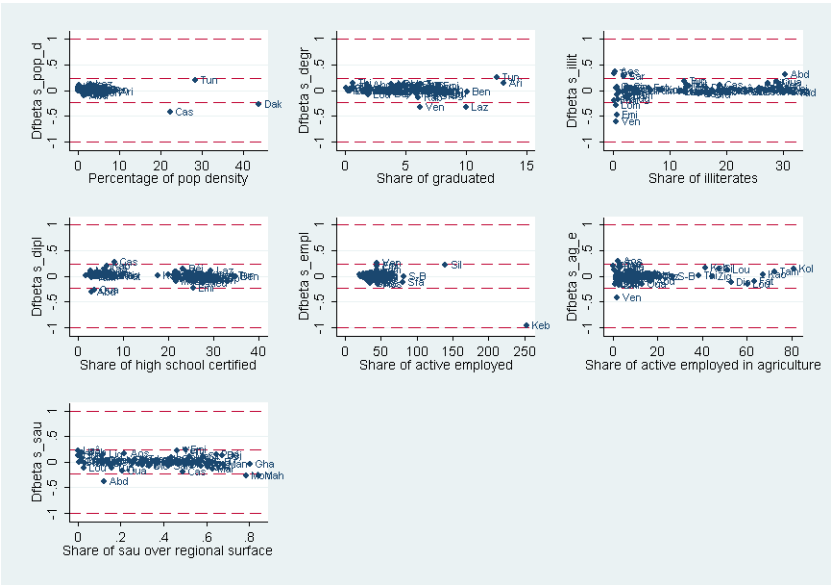

Figura 3: Outliers check of the covariates: the DFBETA-statistic.

Cook's distance[5]. This measure is composed by two part: a ratio which enphasizes the leverage of an observation and its standardized squared residual. The scatterplot shown in *(Figure 4)* plots these components. Every points

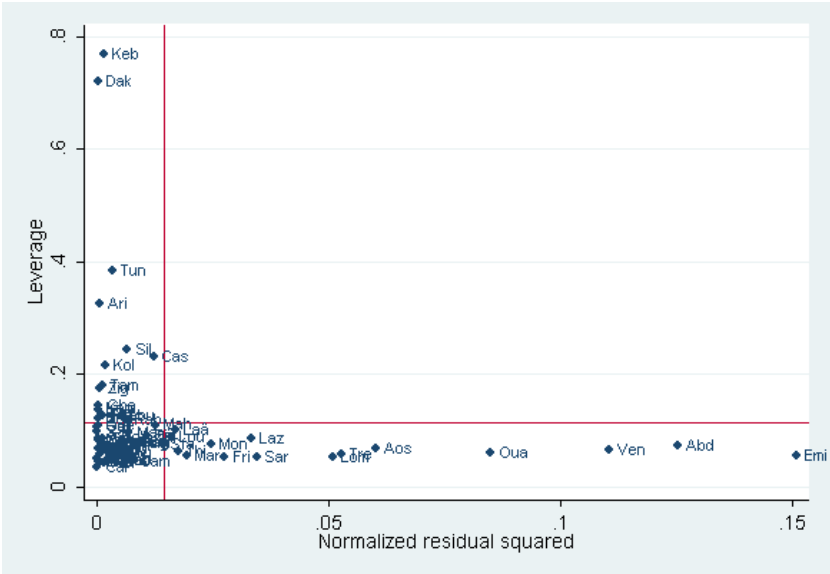

Figura 4: Outliers check of the covariates: the leverage versus regression residuals squared with harvested performance as dependent variable.

are into the bands, but there are data points with a leverage greather than 0.5. This is too much leverage. Taking in consideration, all of these graphs, we have decided to drop out the regions of Kebili and Siliana (both belong in Tunisie). Kebili is canceled out since has too much leverage, Siliana is drop out since has too much residential population respect the people employe-

[5]Cook (1997) defined a influence measure of a given data point based on the norm of the vector $\hat{\beta} - \hat{\beta}_{(i)}$. He proposed:

$$Cook_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})' \boldsymbol{X'X} (\hat{\beta} - \hat{\beta}_{(i)})}{ks^2} = \frac{h_{ii}}{1 - h_{ii}} \frac{\hat{\epsilon}_{si}^2}{k}$$

Where k is the number of covariates, s^2 is the sampling variance and $X'X$ is the metric in which the OLS regression is runned. Instead, h_{ii} are the diagonal elements of the *projection matrix* which produces the fitted values of a regression and $\hat{\epsilon}_{si}^2$ are the squared of standardized residuals. A large value of $Cook_i$ means that i-th observation has an relatively high influence.

ment. Dakar is kept in analysis since capitals represent, especially in african countries, a reality itself.

After this preliminary work, we have started the applied analysis. To discover the evidence of different cluster of regions we have used three different types of *agglomerative clusters methods* and we have chosen the optimal number of clusters using the Duda-Hart's (1973) stop rule statistics *(Table 1)*. According that, the value of pseudo t-test should be minimized and at the

	single linkage (Duda&Hart statistic)		**average linkage** (Duda&Hart statistic)		**complete linkage** (Duda&Hart statistic)	
numb. clusters	Je(2)/Je(1)	pseudo t-test	Je(2)/Je(1)	pseudo t-test	Je(2)/Je(1)	pseudo t-test
1	0.9081	6.68	0.5380	56.68	0.5380	56.68
2	0.4818	68.83	0.8038	13.91	0.8038	13.91
3	0.0000	.	**0.8759**	7.79	**0.8684**	**8.34**
4	**0.9726**	**1.55**	0.4953	54.02	54.02	79.03
5	0.9110	5.28	0.5157	**6.57**	0.3987	10.56

Tabella 1: The optimal number of clusters: the Duda-Hart statistic.

same time the ratio of the scores should be maximized (highlighted values). Looking the table, we can state that the optimal number of cluster should be three. Their graphical representation (*Figure 5*), the belonging region *(Table 2)* and clusters means *(Table 3)* are provided following the complete linkage algorithm.

The interpretation of conditional means groups *(Table 3)* plays a fondamental role in our classification.
Point observations which belongs in cluster one have the highest mean level of both population density and harvested performance. The cluster mean of the share of graduated remains relatively high, instead the cluster mean of employed in agriculture is the lower. This features, arise typically, for more developped areas, where agriculture is specialized, productivity is high and people involved are few. Farmers are skilled and agricultural tools are already developped. We shall label this cluster as the *urban cluster*.

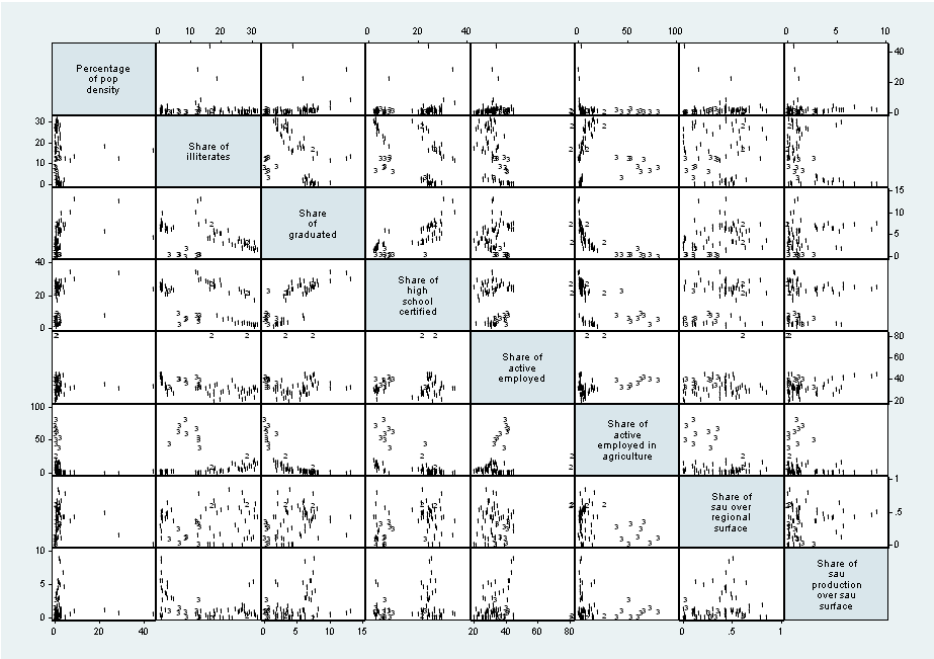

Figura 5: The optimal number of clusters. Scatterplot representation of complete linkage result.

region	group	region	group	region	group
Piemonte	1	Manouba	1	Gharb	1
Valle d'Aosta	1	Nabeul	1	Casablanca	1
Lombardia	1	Zaghouan	1	Guglielmin	1
Trentino	1	Bizerte	1	Laâyoune	1
Veneto	1	Béja	1	Marrakech	1
Friuli	1	Jendouba	1	Meknes	1
Liguria	1	Le Kef	1	Oriental	1
Emilia Rom.	1	-	-	Oued	1
Toscana	1	Sousse	1	Rabat	1
Umbria	1	Monastir	1	Souss	1
Marche	1	Mohdia	1	Tadla	1
Lazio	1	Sfax	2	Tanger	1
Abruzzo	1	Kairouan	1	Taza	1
Molise	1	Kassarine	1	Dakar	1
Campania	1	Sidi Bouzid	2	Diourbel	3
Puglia	1	Gabes	1	Fatick	3
Basilicata	1	Mednine	1	Kaolack	3
Calabria	1	Tataouine	1	Kolda	3
Sicilia	1	Gafsa	1	Louga	3
Sardegna	1	Tozeur	1	St-Louis	3
Tunisi	1	Ouar	1	Tambacounda	3
Ariana	1	Abda	1	Thiès	3
Ben Arous	1	Boulemane	1	Ziguinchor	3

Tabella 2: The optimal cluster result. Regions and groups.

Cluster(w)	E(s_pop_d|w)	E(s_illit|w)	E(s_degr|w)	E(s_dipl|w)	E(s_empl|w)
w=1	3.2	15.2	5.3	19.4	32.1
w=2	0.8	22.6	5.4	24.3	80.6
w=3	0.9	9.3	0.6	8.5	37.6

Cluster(w)	E(s_ag_e|w)	E(s_sau|w)	E(s_s_y|w)
w=1	5.6	0.4	1.9
w=2	18.0	0.6	0.1
w=3	59.1	0.2	0.9

Tabella 3: The condictional means groups.

Cluster two is interesting in both: educational and agricultural features. On the one hand, the highest mean level of illiterates is accompanied by the highest level of graduated. This may be explained by a intracluster spread in schooling. Then, for policy, these regions should uniforme their human capital level, basically, insuring and enlarging the schooling at a larger part of people.
On the other hand, the cluster mean of employment in agriculture remains high, but the mean of the harvested performace is minimized between groups. It implies that, these regions have a very low agriculture productivity. Then, for policy, these regions should invest more in agriculture, enhancing the methods and the physical tools used by the farmers. After that, more people will be profit. We shall label this cluster as the *rural cluster*.

Cluster three is caracterized by the lower mean level of illiterates, graduated and diplomates. Also the level of population density remain low. The mean level of agricultural employment is maximized between groups, but the agricultural productivity remains relativerly low. It implies that no-specialized worker, especially in agriculture, should be present. Then the marginal product of agriculture should be low and the regional income should be at a subsitence level. Agriculture appears to be devoted to self-consumption and not to the market. Then, for policy, on one hand is necessary enhance the level of education, on the other hand may be useful dedicate more agricultural yield to the market. Indeed, to open more the economy to the market. The consumptions should be positively stimulated and with them the regional profit. We shall label this cluster as the *quasi-rural cluster*.

Labeled the cluster type, to verify if optimal clusters of regions are predictive of farmer's income we have had the necessity to fix this optimal number of clusters. Under this context, the *k-means clustering* (Macqueen (1967)) appears being performed to our case. Then, the accurancy of the clusters

in explaining the farmer's income will be tested using the $ANOVA$ analysis taking the harvested performance (s_s_y), considerated dependent, outside the analysis.

K-means clustering is one of the *unsupervised learning algorithms* which allow us to solve the cluster problem fixing *a priori* k centroids, one of each cluster. These centroids should be placed in a convinient way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take account of each point and associate it to the nearest centroid. If all points are filled the first step is completed and an early groupage is done. Now, we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same dataset observations and the nearest new centroid. So, this method can go on.
This algorithm aims at minimizing squared error objectives function:

$$G = \sum_{i=1}^{n} \sum_{j=1}^{k} ||x_i^{(j)} - c_j||^2$$

Where $||x_i^{(j)} - c_j||^2$ is a measure of distance between the ith data point $x_i^{(j)}$ and the jth centroid point[6] c_j. Obviously, k is the number of centroids and n is the number of observations.
Taking the covariate *(s_s_y)* outside the analysis, we have studied the significance of both: the two groups classification (rural-urban separation) (*Table 4*) and the optimal classification (three categories separation) in explaining the farmer's income cluster differences (*Table 5*). These tables come from a corrisponding underlying regression which treats as dependent variable the harvested performance *(s_s_y)* and as regressor the cluster classification vector. In *(Table 4)* the regressor is the two classification vector, instead in

[6]The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the clusters.

	Number of obs = 68			R-squared = 0.0249	
	Root MSE = 2.09786			Adj R-squared = 0.0102	
source	**Partial SS**	*df*	**MS**	*F*	**Prob** > *F*
Model	7.42582549	1	7.42582549	1.69	0.1985
Cluster2	7.42582549	1	7.42582549	1.69	0.1985
Residual	290.466386	66	4.40100585		
Total	297.892212	67	4.44615242		

Tabella 4: The ANOVA analysis: the no-efficiency of a rural-urban classification in explainig the cluster farmer's income differences

	Number of obs = 68			R-squared = 0.0937	
	Root MSE = 2.038			Adj R-squared = 0.0658	
source	**Partial SS**	*df*	**MS**	*F*	**Prob** > *F*
Model	27.9181418	2	13.9590709	3.36	0.0408
Cluster3	27.9181418	2	13.9590709	3.36	0.0408
Residual	269.97407	65	4.15344723		
Total	297.892212	67	4.44615242		

Tabella 5: The ANOVA analysis: the efficiency of three categories classification in explaining the farmer's income differences

(Table 5) the regressor is the three optimal classification vector. The total variance of these regressions is summarized in these tables. This is divided according its source: the variance explained by the model, the variance explained by the cluster vector and the residual variance. Since only a regressor is involved, is naturally, that the variance of the model coincides with the variance of the cluster vector.
The F statistics is a test of the null hypothesis that slopes coefficient are jointly zero. Here, we are testing differences in harvested performance levels among clusters. It appears that, clusters differ *significantly* in their average harvested performance levels only when a three cluster classification is applied (in fact, $F_{2,65} = 3.36$ is significant at 5% level). It implies that, a rural-urban classification seems to be too restrict in explaining farmer's territorial income differences.

Conclusions

The statistical analysis on available dataset for Italy, Tunisia, Morocco and Senegal, allows drawing some concluding remarks.

- Rural and urban classification of administrative areas is *important* to monitor incomes and policymakers' actions;

- Areas can be *efficiently* grouped using statistical techniques, such as cluster analysis. It implies that non-statistical territorial classification, typically baised on thresholds of residential population, might be not the best way for territorial classification;

- Territorial classification is *strictly necessary* for policy monitoring. Territorial differences can become clear through classification, specific policy advices and actions can be formulated according to the territorial clusters. It implies that gouvernment policies can be improved if countries provide more information, especially for income at disaggregated administrative levels (regional, departmental or gouvernoratal),

because *proxy* variables do not capture effectively the nature of a variable. This is especially important for income data at households' level, that are key variables for policy monitoring and analysis;

- The *relevance* of a statistical classification (the optimal number of clusters) to explain income differences, can be tested using the ANOVA approach. Non-statistical territorial classification might be useless or misleading in explaining the phenomenon of interest.

In conclusion, a good and reliable territorial classification, together with income data at a disaggregated level, is of primary importance to a clear understanding of trends and an effective policy action.

Bibliografia

[1] Agresti A. (2002): *Categorical Data Analysis*; John Wiley & Sons, 2nd edition.

[2] Alon U. , Barkai N. , Notternam D. A., Gish K., Ybarra S. Mack D. & Levine A. J. (1999): *Broad Patterns of a Gene Expression Revealed by Clustering Analysis*; Cell Biology, vol 99, pag 6754-6760.

[3] Belsley D. A. , Kuh E. & Welsch E.R (1980): *Regression Diagnostics: Identifying Influental Data and Sources of Collinearity*; Ney York, Wiley

[4] Belsley D. A. , Kuh E. & Welsch E.R (1991): *Condictioning Diagnostics: Collinearity and Weak data in Regression*; Ney York, Wiley

[5] Calinski T. & Harabasz J. (1974): *A Dendrite Method for Cluster Analysis*; Communications in Statistics vol 3, pag. 1-27

[6] Cook R. D. (1977): *Detection of Influential Observations in Linear Regression*; Thecnometrics vol 19, pag. 15-18.

[7] Duda R. & Hart P. (1973): *Pattern Classification and Scene Analysis*; Wiley.

[8] Duda R. & Hart P. (2000): *Pattern Classification*; Wiley, 2nd edition.

[9] Everitt B. S. & Bullmore (1999): *Mixture model Mapping of brain Activation in Functional Magnetic Resonance Images*; Human Brian Mapping, vol 7, pag. 1-14.

[10] Green W. H. (2006): *Econometric Analysis*; Prentice Hall, 5th edition.

[11] MacQueen J. B. & Hart P. (1967): *Some Methods for classification and Analysis of Multivariate Observations*; University of California Press, vol 1, pag. 281-297.

[12] Peracchi, F. (2000): *Econometrics*; Wiley.

[13] Pizzoli E. & Xianoning G. (2000): *How to Best Classify Rural and Urban?*; Fourth International Conference on Agriculture Statistics (ICAS-4), China.

[14] Wooldridge J. M. (2001): *Econometric Analysis of Cross Section and Panel Data*; MIT Press.