

Optimal design for area crop insurance: a model-based area estimation approach

Disegno Ottimo per Assicurazioni ad Area in Agricoltura: un Approccio di Stima basato su Modello

Carlo Russo

Department *Istituzioni, Metodi Quantitativi e Territorio*, University of Cassino, Italy
E-mail: russocar@unicas.it

Renato Salvatore

Department *Istituzioni, Metodi Quantitativi e Territorio*, University of Cassino, Italy
E-mail: rsalvatore@unicas.it

Riassunto: Le assicurazioni ad area sono contratti finalizzati ad offrire agli agricoltori uno strumento di difesa contro la variabilità dei raccolti a costi estremamente contenuti. Tuttavia, affinché la copertura offerta da questi contratti sia efficace è necessario che vi sia un'elevata correlazione fra le rese individuali dei singoli agricoltori e la resa aggregata dell'area. In questo lavoro proponiamo una metodologia per la ripartizione di un universo di produttori in aree che massimizzino la correlazione fra le rese individuali e dell'area, garantiscano la contiguità territoriale e impediscano fenomeni di collusione fra i produttori. Il metodo si caratterizza per l'estrema parsimonia nella richiesta di dati. La metodologia si basa per la prima parte su una stima multivariata per piccole aree, e successivamente su un procedimento di aggregazione delle aree basato sulla massimizzazione della variabilità relativa between-clusters, misurata attraverso le stime delle componenti di varianza del modello lineare ad effetti misti che lega la resa delle aree con dei predittori di quest'ultima. A titolo illustrativo, il metodo è stato applicato al mercato assicurativo sulla produzione di grano della regione Lazio.

Keywords: area crop insurance, multivariate model-based small area estimation, linear mixed effects models, clustering techniques.

1. Introduction

Risk management is one of the main concerns for farmers. Agricultural firms, in fact, are exposed to a remarkable variation in income, due to a broad set of sources: weather, biological factors, price fluctuation, policy change are just few examples of the many determinants of farm risk. As a consequence, the farm income depends on many factors that are not under the direct control of the entrepreneur and traditionally farmers look out for viable risk management tools.

Despite of the relevance of risk, farmers' participation to insurance schemes is still considered a problem (European Commission, 2001). Without heavy public subsidies many European farmers are unwilling to subscribe insurance contracts, because of the high expenses for the premium and of the limited coverage of the programs. On the supply side, insurance companies are often unwilling to offer multi-

risk contracts because of information asymmetries and potentially opportunistic behavior by farmers.

In order to boost the insurance market, in early 1990s it was proposed a new contract design aimed to reduce the effects of information asymmetries: the so-called area insurance. In an Area Contract, the insurance company offers a standard contract to all the farmers located in a given area. The insurance company will pay the farmers subscribing the policy if the average yield of the area falls below a threshold, regardless of the yield obtained by the individual farmers (Miranda 1991; Skees, Black, Barnett 1997). Unlike standard insurance contracts, the trigger for the payment is not the individual yield, but it is the aggregate production for the whole area. Similarly, the amount of the payment depends on the aggregate production. As a consequence, the insurance contract covers the systematic risk only, and does not affect the idiosyncratic risk.

A key feature of the area contracts is that they can reduce significantly the insurance costs by lowering the transaction costs due to moral hazard. In fact, in an individual insurance contract, the farmer may have low incentive to avoid the loss of production, because of the insurance. If production protection is costly, the farmer may prefer to lose the yield and cash the insurance payment. As a consequence, insurance companies are forced to monitor farmers in order to make sure that they make all the possible effort to avoid the production loss. This monitoring costs are so high that they raise the premium of the insurance considerably, further discouraging farmers from subscribing. By cutting the monitoring costs, the area contract can offer cheaper risk management tools.

Summarizing, area contracts greatly reduce the transaction costs in the insurance markets. As a consequence insurance companies may offer cheaper contracts. However, this result is achieved at the cost of a less effective risk reduction for the farmer. Miranda (1991) and Mahul (1999) showed that the benefit for farmers of the area contract increase with the degree of correlation between the yield of the farmer and the area yield. Thus, in order to have effective insurance contracts, it is necessary that the overall level of correlation within the area is high, i.e. that the farmers are homogeneous.

Area contracts performed extremely well in the central United States (the so-called "corn belt"), but their implementation in Italy and Europe is considered quite difficult. In order to be efficient area contracts requires a strong correlation between the yield of the individual farmers and the average yield of the area. Thus, the remarkable heterogeneity of European farms at local level can prevent the success of these program. In fact, in Europe, area contract are usually offered at NUTS 3 level (the Italian Provincia), because this is the most disaggregated level of time series data for agricultural production. However, at this level of aggregation farms are too heterogeneous and the farmers have little use for these contracts.

In this paper we propose a methodology, based on the joint use of small area estimation and clustering, to identify smaller contract areas that maximize the correlation between the yield of the individual farmers and the aggregate area. Using this approach insurance companies can offer more efficient risk management tools and farmers may be willing to pay higher premium, increasing the participation to the insurance market. An additional feature of the model is that it is extremely parsimonious, requiring extremely short time series. In this paper we propose an

application using only cross-section data. Thus, an accurate identification of the areas could be obtained after one sample survey.

In the field literature, usually the distribution of the area yield is assumed exogenous, because the borders of the area are given. In our approach, instead, the area is determined endogenously by a planner willing to maximize the sum of the yield correlation. This approach changes radically the nature of the problem, because now the social welfare from the insurance contract may increase if the area is identified optimally. The identification of the area is a relevant estimation problem, because the universe of the producers must be broken down in areas that have high correlation across the yields, but at the same time must contain a high number of farmers in order to prevent moral hazard and collusion. Furthermore the areas should be spatially contiguous, in order to lower the insurance companies' administration costs and to grant farmers the possibility of monitoring the aggregate yield.

The problem is even more complicated if we consider that time series about individual or small area yields are usually unavailable or unreliable. Moreover large scale surveys are extremely costly and will take multiple years to build the necessary datasets. Thus we cannot easily identify the optimal areas using a traditional time series approach. In section 2 we describe a methodology to achieve the stated objectives under data scarcity using small area estimation and clustering. In section 3 we will provide an application of the model, identifying the optimal areas for an area insurance contract covering the variation of wheat yield in the Lazio Region.

2. The Estimation of local yields and clustering of areas

In this paper the variable yield is the simple ratio between the wheat production and the acreage. In order to get estimation of variable yield at the township level of aggregation, we use small area estimation models, in particular a multivariate model that achieve effective estimates of totals of production and acreage, jointly considered in a bivariate vector of area level estimates.

In recent years one of the most difficult problem to solve by the official statistical institutes is the information delivery about small geographical areas. In the context of agriculture, there is a growing demand for reliable small area statistics in order to assess or to put into effect agricultural policies and programs. Data collected from some well-organized sample surveys, like the farm structure surveys, can be used only to derive direct estimates for large areas, such as administrative districts or regions of national territories. Sample sizes in small domains of interest are typically reduced by the survey design itself, if the objective is to provide accurate estimates at higher level of aggregation than that of small areas. In any case, direct estimates in specific areas are not reliable, because the smallness of sample sizes in the areas can drive to unacceptably large standard errors. Objectives of small area statistics provided by agricultural sample surveys are generally related to parameters of interest such as totals or means of tillable areas, amounts of specific area crops, and so on, based on arrays of some questionnaire response variables. The estimates, derived from synthetic, composite and model-based estimators, are connected with important auxiliary data. In particular, small area models, such as linear mixed models, involve random area-specific effects, and the related composite models was studied through *EBLUP* (empirical best linear unbiased prediction), *EB* (empirical Bayes), and *HB* (hierarchical

Bayes) estimation procedures. The area estimation approach used in this paper is the multivariate linear mixed model and (empirical) linear unbiased prediction estimator of area totals. The *EBLUP* estimator is multivariate (*M-EBLUP*).

We now consider the linear mixed model with block diagonal covariance structure, that for the i -th area (Rao, 2003) is as follows:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{v}_i + \mathbf{e}_i \quad i = 1, \dots, m \quad (3.1)$$

with

$$\mathbf{y} = \text{col}_{1 \leq i \leq m}(\mathbf{y}_i) = (\mathbf{y}'_1, \dots, \mathbf{y}'_m)', \quad \mathbf{X} = \text{col}_{1 \leq i \leq m}(\mathbf{X}_i)$$

$$\mathbf{Z} = \text{diag}_{1 \leq i \leq m}(\mathbf{Z}_i), \quad \mathbf{v} = \text{col}_{1 \leq i \leq m}(\mathbf{v}_i), \quad \mathbf{e} = \text{col}_{1 \leq i \leq m}(\mathbf{e}_i)$$

where m is the total amount of areas, $\mathbf{X}_i$ ($n_i \times p$) is the matrix of area-level (p) covariates, $\mathbf{Z}_i$ ($n_i \times m$) is the design matrix, $\mathbf{v}_i$ are *iid* and independent of the sampling error $\mathbf{e}_i$, with zero means and covariance matrices $\mathbf{G}_i \in \mathbf{R}_i$, respectively. The m models come together in the general linear mixed model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}$, and this model, in the case of a single parameter of interest (with $\mathbf{Z}_i = \mathbf{1} \rightarrow \mathbf{Z} = \mathbf{I}_m$) is of the following general form:

$$\hat{\boldsymbol{\theta}} = \mathbf{z}'\boldsymbol{\beta} + \mathbf{v} + \mathbf{e} \quad (3.2)$$

(where, for the i -th area: $\hat{\boldsymbol{\theta}}_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i + e_i$). The multivariate small area model (as generalization of the basic Fay-Herriot model) extends the area-level model to a vector of r survey estimators. The multivariate model is $\hat{\boldsymbol{\theta}}_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i + e_i$, and can lead to more efficient estimators, because it takes advantage of the correlation between the components of the vector $\hat{\boldsymbol{\theta}}_i$ unlike the univariate model. More precisely, the value of $MSE(\hat{\boldsymbol{\theta}}_i^H)$ is related to the level of correlations between the vector components $\hat{\boldsymbol{\theta}}_i^H$ (the *M-BLUP* estimator) of the parameters of survey interest. This linear mixed model is as the repeated measurement small area estimation model in which the vector $\hat{\mathbf{y}}_i$ of the n_i repeated measures for the i -th area ($\mathbf{Z}_i = \mathbf{1}_{n_i}$) are replaced by the observed values of the r -dimensional vector $\hat{\boldsymbol{\theta}}_i$, and so in the model by the estimators of the parameters at area level. The variance components of the general model are represented by the matrices $\mathbf{G}_i = \boldsymbol{\Sigma}_v$ (diagonal or symmetric) and $\mathbf{R}_i = \boldsymbol{\Psi}_i$ (symmetric), namely the random-area error multivariate vector covariance matrix and sampling covariance matrix. The variance-covariance matrix of the estimate $\hat{\boldsymbol{\theta}}_i^H$ is of the form $\mathbf{V}_i = \boldsymbol{\Psi}_i + \mathbf{Z}_i \boldsymbol{\Sigma}_v \mathbf{Z}_i' = \boldsymbol{\Psi}_i + \mathbf{I}_r \boldsymbol{\Sigma}_v \mathbf{I}_r' = \boldsymbol{\Psi}_i + \boldsymbol{\Sigma}_v$ (Rao, 2003). The *M-EBLUP* estimator $\tilde{\boldsymbol{\theta}}_i^H$ is the generalization of the univariate case, and is as follows:

$$\tilde{\boldsymbol{\theta}}_i^H = \hat{\boldsymbol{\Sigma}}_v(\boldsymbol{\Psi}_i + \hat{\boldsymbol{\Sigma}}_v)^{-1}\hat{\boldsymbol{\theta}}_i + \boldsymbol{\Psi}_i(\boldsymbol{\Psi}_i + \hat{\boldsymbol{\Sigma}}_v)^{-1}\mathbf{Z}_i\tilde{\boldsymbol{\beta}} \quad (3.3)$$

Under normality of the errors, the estimation of the matrix $\boldsymbol{\Sigma}_v$, that brings the variance component of the multivariate random-area effect (the r dimensions of the vector $\hat{\boldsymbol{\theta}}_i$), is based on maximum likelihood (*ML*) or restricted (residual) maximum likelihood estimation (*REML*). Method of moments, or Rao's minimum norm quadratic unbiased estimation (*MINQUE*) are also possible. The *M-EBLUP* estimator, as in the univariate case, attach more weight to the regression-synthetic estimator $\mathbf{Z}_i\tilde{\boldsymbol{\beta}}$ when random-area effect variances are relatively small. The estimated *MSE* of $\hat{\boldsymbol{\theta}}_i^H$ is as follows:

$$\begin{aligned} MSE(\hat{\boldsymbol{\theta}}_i^H) &= E(\hat{\boldsymbol{\theta}}_i^H - \boldsymbol{\theta}_i)(\hat{\boldsymbol{\theta}}_i^H - \boldsymbol{\theta}_i)' = (\boldsymbol{\Psi}_i^{-1} + \boldsymbol{\Sigma}_v^{-1})^{-1} + \\ &+ (\boldsymbol{\Psi}_i^{-1} + \boldsymbol{\Sigma}_v^{-1})^{-1}\boldsymbol{\Sigma}_v^{-1}\mathbf{Z}_i \times \left[\sum_{i=1}^m \mathbf{Z}_i'(\boldsymbol{\Psi}_i + \boldsymbol{\Sigma}_v)^{-1}\mathbf{Z}_i \right]^{-1} \mathbf{Z}_i' \boldsymbol{\Sigma}_v^{-1}(\boldsymbol{\Psi}_i^{-1} + \boldsymbol{\Sigma}_v^{-1})^{-1} \end{aligned}$$

More precisely, if we have $(tr(\boldsymbol{\Sigma}_v', \boldsymbol{\Sigma}_v))^{1/2} < (tr(\boldsymbol{\Psi}_i', \boldsymbol{\Psi}_i))^{1/2}$, it is easy to see that, for the i -th area, if $\boldsymbol{\Sigma}_v = diag(\sigma_{11}^2, \dots, \sigma_{rr}^2)$, $\sigma_{jj}^2 \downarrow$, and so $(\boldsymbol{\Psi}_i^{-1} + \boldsymbol{\Sigma}_v^{-1})^{-1} = \boldsymbol{\Sigma}_v^* < \boldsymbol{\Sigma}_v$, for the *M-BLUP* estimator $\hat{\boldsymbol{\theta}}_i^H$ the approximated inequality holds:

$$MSE_1(\hat{\boldsymbol{\theta}}_i^H) \approx \boldsymbol{\Sigma}_v^* \boldsymbol{\Sigma}_v^{-1} (\sum_{i=1}^m \boldsymbol{\Psi}_i^{-1})^{-1} \boldsymbol{\Sigma}_v^{-1} \boldsymbol{\Sigma}_v^* < (\sum_{i=1}^m \boldsymbol{\Psi}_i^{-1})^{-1}$$

In that expression, approximations are valid for matrices $\hat{\boldsymbol{\theta}}_i^H$, $\boldsymbol{\Psi}_i$, $\boldsymbol{\Sigma}_v$, $\boldsymbol{\Sigma}_v^*$ elements. On the other hand, with constant elements of matrix $\boldsymbol{\Psi}_i$ (the sampling area level variance-covariances) and $\sigma_{jj}^2 \uparrow$, we have $(\boldsymbol{\Psi}_i^{-1} + \boldsymbol{\Sigma}_v^{-1})^{-1} \approx \boldsymbol{\Psi}_i$, and so the mean squared error of the *M-BLUP* is:

$$MSE_2(\hat{\boldsymbol{\theta}}_i^H) \approx \boldsymbol{\Psi}_i + \boldsymbol{\Psi}_i \boldsymbol{\Sigma}_v^{-1} (\sum_{i=1}^m (\boldsymbol{\Psi}_i + \boldsymbol{\Sigma}_v)^{-1})^{-1} \boldsymbol{\Sigma}_v^{-1} \boldsymbol{\Psi}_i > \boldsymbol{\Psi}_i$$

Clearly we have $MSE_1(\hat{\boldsymbol{\theta}}_i^H) < MSE_2(\hat{\boldsymbol{\theta}}_i^H)$, in consequence of that the appropriateness of the model is crucial to achieve small values of variance of estimators (Särndal et al., 1992).

After the first step, consisting of the estimation of the yields at the township level through the multivariate small area estimation of wheat production and acreage, the main problem was related to the possible more efficient method for clustering local territories in similar areas, in such a way that the estimated wheat yields of local territories in each cluster seems to be very similar. In the methodology that we propose, clustering techniques are jointly employed with linear mixed effects models methodology (Demidenko, 2004) in order to achieve high levels of relative between-clusters variation of yields throughout the areas. More precisely, we optimize cluster membership after a first aggregation via classic clustering algorithms, namely by using hierarchical methods, by substituting the classic between and total inertia ratio maximization

criterion with the maximization of the ratio between the variance estimate of the random effect and the model overall variance estimate. In the mixed model

$$y_{ij} = \alpha + \beta x_{ij} + \eta_{ij}, \quad i = 1, \dots, N, \quad j = 1, \dots, n_i$$

where y_{ij} is the j -th observation in the i -th cluster, $\eta_{ij} = \varepsilon_{ij} + b_i$ is the composite random error, observations in the same area (within cluster) correlate with the correlation coefficient:

$$\rho = \frac{\text{var}(b_i)}{\text{var}(b_i + \varepsilon_{ij})} = \frac{\sigma_b^2}{\sigma^2 + \sigma_b^2}$$

There are two sources of variation: the within-cluster variation, σ^2 , and the between-cluster variation, σ_b^2 , and so the larger is the variation between clusters, the higher is the correlation within each cluster. In general, the linear mixed effects model is written as:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, N$$

where $\mathbf{b}_i$ is a vector of random effects such that $\text{cov}(\mathbf{b}_i) = \sigma_b^2 \mathbf{G}$, and $\mathbf{Z}_i$ is the design matrix. The variance parameters, σ_b^2 and $\mathbf{G}$, are unknown and are subject to estimation (via *ML*, *REML*, under assumption of normality, and also by method of moments or *MINQUE*) along with the population-averaged parameter $\boldsymbol{\beta}$. We firstly assumed in this paper the normality of random errors, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ and $\mathbf{b}_i \sim N(\mathbf{0}, \sigma_b^2 \mathbf{G})$, so that the model is:

$$\mathbf{y}_i \sim N(\mathbf{X}_i \boldsymbol{\beta}, \sigma^2 (\mathbf{I} + \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i')) \quad i = 1, \dots, N$$

and then we have reassigned each element in predefined clusters in order to get, at each iteration, the best value of the ratio:

$$\hat{\rho} = \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' (\mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' + \mathbf{I})^{-1} = \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' \mathbf{V}_i^{-1} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}^2 + \hat{\sigma}_b^2} \quad (3.4)$$

with $\text{var} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\boldsymbol{\varepsilon}} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{G}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$. The estimation method used is, under assumption of normality of $\boldsymbol{\varepsilon}_i$ and $\mathbf{b}_i$, the *REML* estimation.

3. Application: the identification of optimal areas for a wheat area insurance in the Lazio Region

The data of the application are those from the 1999 Italian Farm Structure Sample Survey for the Lazio administrative Region. We utilize two predictors, the *tillable area* and the *irrigated farmland*, in order to obtain joint estimates of wheat production and acreage at area level (the township level). The multivariate small area estimation model is the linear mixed model with block diagonal covariance structure and normal distributed errors, and so we use the best linear unbiased predictor at multivariate level in order to estimate the values of the variable *wheat yield* (the ratio between the wheat production and acreage). The estimated values are the totals of wheat production and acreage in each of the small areas. The values of area-based direct estimators are affected by (estimated) large values of the *MSE*. The approximate value of correlation between wheat production and acreage is 0.97. This value allows *M-EBLUP* to perform very well. As we see below, the ratio between the average estimate of *MSE* of direct area-level estimators and the average estimate of *MSE* of *M-EBLUP* estimators is very high, especially for the variable *acreage*. This because linear models that explain acreage with *tillable area* and *irrigated farmland* covariates perform better than that wheat production with the same predictors. The variance components are the multivariate random-area effect variance estimates, the diagonal matrix $\mathbf{G}_i = \Sigma_v$, and the symmetric unstructured matrix of multivariate residual model variance with correlated error components $\mathbf{R}_i = \Psi^R$, where *R* here means “residual”. The estimated matrices of variance components were obtained through Sas[®] Proc Mixed software, and fixed and random effects tested with Prasad-Rao-Jeske-Kackar-Harville method, through Kenward-Roger denominator degrees of freedom method. The *REML* estimation of variance components was obtained through the Sas[®] modified Newton-Raphson maximization algorithm. In this case, the second derivative matrix of the objective function is the asymptotic variance-covariance matrix of the estimated parameters $\mathbf{G}$ and $\mathbf{R}$.

The first step of clustering is a simple hierarchical nearest neighbourhood algorithm (the Ward criterion), and clustering consolidation and re-allocation of local small areas was obtained performing the best value of the objective function stated by (3.4). The algorithm reiterates this overall-clusters measure until the relative increase is less than 10^{-6} . Firstly, we obtain the measure in (3.4) for the administrative mapping (5 clusters) of the Lazio Region, that explain only a relative between-cluster variation estimated in 7×10^{-4} . Then we perform the above explained consolidation iterative procedure with 6, 12, 6, and 5 clusters respectively. The optimized consolidation achieve a relative between-cluster variation estimated in 0.997 (6 clusters). The last procedure in the application perform the best clustering subjected to two different constraints: spatial contiguity and presence of at least 300 producers. The spatial contiguity constraint worked in such a way that the procedure is stopped when the relative between-cluster variation is smaller than 0.9 (the error is less than 10^{-2}).

In tables and figures below we furnish the summary statistics of the area estimation, the values of the objective function in each clustering procedure and, finally, five different mappings of the Lazio Region at each different clustering.

Table 1: *Estimated wheat yields statistics*

Estimated yields summary statistics - Hill		
mean	std err	coeff. of var.
32.19577237	83.48996995	2.593196678

Estimated yields summary statistics - Mountain		
mean	std err	coeff. of var.
30.87643893	77.96048261	2.5249182

Estimated yields summary statistics - Plain		
mean	std err	coeff. of var.
30.31421781	40.32870482	1.330356108

Table 2: *Average estimated MSE (AEM) ratio - (M-EBLUP / Direct estimator)*

Ratio AEM (M-EBLUP) / AEM(direct)	
wheat production	126.54796
acreage	4402.09214

Table 3: *Estimated values of relative between-cluster variation in several clustering procedures (in brackets the number of clusters)*

Estimated values of correlation (between-cluster / total variation ratio)				
Administrative (5)	Univariate (6)	Trivariate (12)	Optimized (6)	Constrained optimized (5)
0.000748	0.992765	0.996117	0.997028	0.898343

Figure 1: NUTS 3 Administrative Areas of Lazio

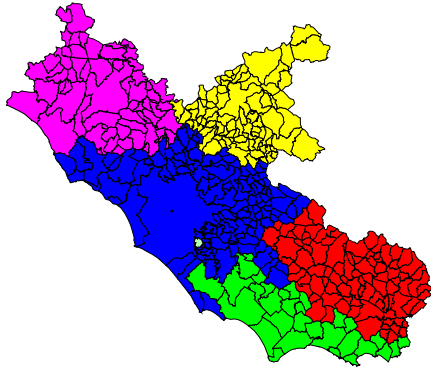


Figure 2: Univariate Clustering (6 Clusters)

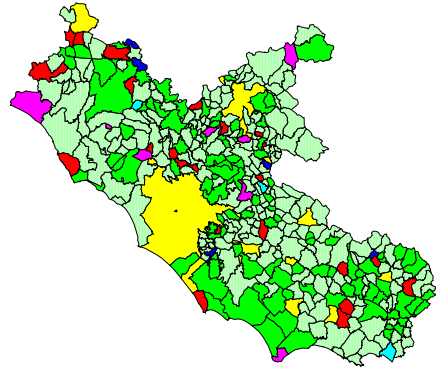


Figure 3: Trivariate Clustering (12 Clusters)

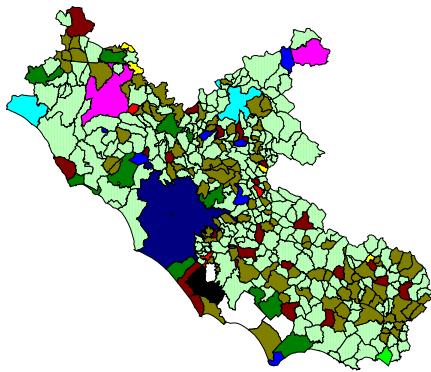


Figure 4: Optimized Clustering (6 Clusters)

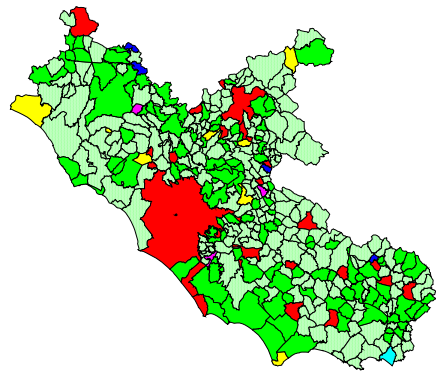
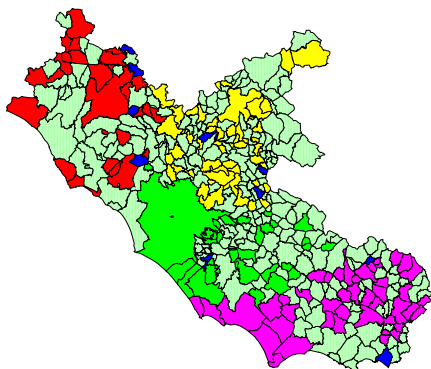


Figure 5: Constrained Optimized Clustering (5 Clusters)



4. Conclusions

The methodology described in the previous section broke down the Lazio Region into five homogeneous areas. Each area has at least 300 producers, so that there is low chance for opportunistic behavior and moral hazard. Moreover four clusters out of five are spatially contiguous so that the administration costs for the insurance companies are lower and farmers can monitor the area yields.

Figure 5 reports the results on a map. At first glance it is apparent that the administrative borders (the NUTS 3 areas, showed in Figure 1) are only vaguely maintained. The inclusion of spatial contiguity in the objective function grouped the observation in four main geographical areas approximately corresponding with the Southern Lazio and the Province of Viterbo, Rieti and Rome. The fifth cluster is composed of a group of observation that have extremely high yield and are strongly correlated. Thus, the optimal insurance scheme is based on four contiguous areas and one additional group containing the “yield champions”, which can be considered outliers from the other four clusters. Compared to the current organization, based on the administrative borders, the new mapping greatly increases the expected correlation between the individual yields and the areas. As it was shown in section 2, the higher correlation increases the expected utility of the insurance contract for farmers, who should exhibit a higher willingness to pay for these schemes. The actual effect of the clustering is that the participation of farmers to the insurance market should increase and, at the same time, the policy maker should be able to lower the financial incentive for farmers.

In the application the gain in correlation is quite remarkable, going from 0.007 of the administrative regions to the 0.89 of the constrained clustering. This result suggests that the optimization of the insurance mapping potentially can make the difference between an existing and a missing market. If the insurance offers a contract based on the administrative borders, the low correlation will discourage a large number of farmers from participating. Instead using the optimized mapping the number of subscribers is expected to increase significantly.

References

- Arrow K.J. (1970): *Essays in the Theory of Risk Bearing*, Markham, Chicago 1970
- Demidenko E. (2004), *Mixed models: Theory and Applications*, Wiley, New York
- European Commission (2001): Risk Management Tools for EU Agriculture, with a special focus on insurance, *Working Document*, January 2001
- Mahul O. (1999): Optimum Area Yield Crop Insurance, *American Journal of Agricultural Economics*, n. 81, pp. 75–82.
- Mahul O., Vermersch D. (2000): Hedging crop risk with yield insurance futures and options, *European Review of Agricultural Economics*, vol.27, pp. 109–126.
- Miranda M. (1991): Area-yield Crop Insurance reconsidered, *American Journal of Agricultural Economics*, n. 73, pp. 233–242.
- Rao, J. N. K. (2003), *Small Area Estimation*, Wiley, New York, 153-155
- Sarndal, C. E., Swensson B., Wretman J., (1992), *Model Assisted Survey Sampling*, Springer, New York, 238-239
- Skees J. R., Black J. R., Barnett B. J. (1997): *Designing and Rating an Area Yield Crop Insurance contract*, *American Journal of Agricultural Economics*, 79, 430-438