Administrative Data and Model Based Estimation in Italian Agriculture Statistics

Roberto Gismondi[1], Loredana De Gaetano[2]

[1] *ISTAT, Roma, Italy; gismondi@istat.it*
[2] *ISTAT, Roma, Italy; degaetan@istat.it*

Abstract
In the European Union, the Regulation (EC) 543/2009 requires that each Member State produces estimates on agricultural surfaces by kind of crops and early estimates concerning the forthcoming agricultural year. Actually, agricultural surfaces are estimated by the Italian National Institute of Statistics (ISTAT) through regional experts evaluations: it is difficult to assess quality and several delays occur. However, IACS administrative data are now available within a shorter time lag, they cover a broader set of crops and may substitute actual estimates gradually. As regards early estimates, the usual design based estimation strategy (stratified random sample and Horvitz Thompson estimator) has been improved through double sampling and model based regression estimation using relationship between crop data of two consecutive years. Results show decrease of estimates model variances and higher degree of coherency between land use in following years.

Keywords: Administrative data, Agriculture, Crop, Double sampling, IACS.

1. Quality issues for Italian crop statistics

Actually ISTAT is producing data on agricultural areas and production at regional and national levels on the basis of the survey "Crop statistics". The basic methodology is mainly founded on the "estimative" technique. For any particular cultivation, data derive from the product between the estimation of Utilized Agricultural Area (UAA) and the average yield per hectare. Data are provided by local authorities[1]. that collect experts evaluations on area and yield of different crops. Auxiliary information may be added to experts' estimates (estimates by associations of producers, administrative data). Crops under investigation are different for

[1] They are mainly given by Italian Regions and Autonomous Provinces. In Italy there are 19 Regions and 2 Autonomous provinces. Overall, the number of provinces is 110.

each month and take into account the phenological stage of cultivation. For this reason more than one estimate can be determined for each crop during the same year. Data are provided monthly at the province level. ISTAT checks and validates them, then province data are summed up at the regional and national level. Along last years, serious sustainability problems arose as regards data quality and timeliness. Recently some attempts aimed at using alternative data sources were carried out, in order to gradually substitute the estimative technique. Moreover, every year ISTAT carries out the sampling survey "Crops early estimates", with the goal of producing anticipated estimates as regards agricultural land use. The quality of the forecasts is founded on the degree of difference between forecasts and true areas(derived from crop statistics), available one year later. In this context, methodological improvements concerning both crop statistics and early estimates are presented. While section 2 deals with the use of administrative data in official statistics (Kloek and Vaju, 2013), section 3 shows a new strategy for early estimates. Perspective conclusions have been drawn in section 4.

2. Administrative data on land use for agricultural purposes

2.1. Statistical use of IACS data in Italy.

IACS (*Integrated Administration and Control System*) is the most important system for the management and control of payments to holders made by the Member States in application of the Common Agricultural Policy. IACS is operated in the Member States by accredited paying agencies. It covers all direct payment support schemes as well as certain rural development measures. The legal requirements concerning IACS are laid down in Council Regulation (EC) No 73/2009 establishing common rules for direct support schemes for holders and implementing rules are given in Commission Regulation (EC) No 1122/2009. IACS is a system of interconnected databases used to receive and process aid applications. The IACS databases is updated by the Member States and the holders' historical data must be saved.

As regards the Italian IACS authority (AGEA), obligation by law should limit the risk of cases for which agriculture producers or traders do not subscribe. On the other hand, the logic underlying the IACS register is based on self-declarations as regards area used for agricultural purposes: this feature may hamper data reliability, since under-declarations (or over-

declarations in cases when a specific EU financial contribution system is operational) may happen. Other potential causes of errors may be due to the following factors:

- mistakes due to producers' declarations;
- duplications derived from double counting of some productions: for instance, the production concerning an olive presser may be duplicated if it is declared also by the packaging/trading enterprise which received the oil from the same olive presser.

Broadly speaking, all the previous risks may be addressed to the "population coverage" problem which must be tackled whenever an administrative source is intended to be used for statistical purposes. Moreover, limitations to the use of IACS data within current crop statistics mainly derive from: a) periodicity of declarations (data are available after 6 months from the end of the reference year, while current crop statistics must release estimates on a monthly basis depending on the cultivation); b) the need to manage properly and gradually the overlapping between this data source and estimates carried out by Italian Regions. On a lesser extent, it is also needed further effort for achieving deeper comparison between concepts and definitions adopted within the IACS and the ISTAT current crops statistics frameworks.

2.2. Comparison among sources

On the basis of the last data, referred to 2014, comparison among IACS data and the ISTAT crop statistics have been carried out. The main outcomes have been resumed in table 1. The kind of cultivations analyzed cover the 20% of Italian agricultural area: they are rice, olives, grapes, fruit and citrus fruit. As regards fruit, additional details are presented in table 2. The first outcome is that IACS data are aligned with crop statistics and are not systematically higher or lower, both at the whole Italy and at the geographical area levels. If we exclude rice – whose statistical data derive from another administrative source (*Ente Risi*) – crop statistics are a bit higher than IACS data: that could be due to multiple uses of the same agricultural land occurring during the same agrarian year. On average, IACS data are 1,9% lower than crop statistics, and this evidence occurs in the Centre and in the South as well. The largest difference concerns citrus fruit (25%), while discrepancies are quite low especially for olives (0,5%) and fruit (1,3%). On the other hand, larger differences characterize rice and citrus fruit.

Table 1 – *Agricultural land use in 2014 - Comparison among sources (hectares)*

Source/Cultivation	Rice	Olives	Grapes	Citrus	Fruit	Total
IACS						
Italy	**234.813**	**1.119.633**	**653.697**	**106.476**	**377.557**	**2.492.176**
North	229.981	17.879	253.983	17	159.437	661.298
Centre	422	176.959	101.243	313	62.238	341.175
South	4.410	924.795	298.471	106.145	155.883	1.489.703
Crop statistics						
Italy	**219.532**	**1.125.183**	**682.183**	**142.011**	**372.582**	**2.541.491**
North	215.342	23.343	230.959	55	133.559	603.258
Centre	378	201.986	107.984	653	37.893	348.894
South	3.812	899.854	343.240	141.303	201.130	1.589.339
FSS 2013						
Italy	**212.238**	**1.073.324**	**635.979**	**129.155**	**388.808**	**2.439.504**
North	209.960	20.121	246.962	16	164.886	641.945
Centre	0	182.122	103.056	2.286	51.834	339.298
South	1.834	871.081	285.961	126.853	172.088	1.457.817
% Difference (Italy)						
IACS vs crop statistics	**7,0**	**-0,5**	**-4,2**	**-25,0**	**1,3**	**-1,9**
IACS vs FSS 2013	10,6	4,3	2,8	-17,6	-2,9	2,2
Crop statistics vs FSS 2013	3,4	4,8	7,3	10,0	-4,2	4,2

Source: elaboration on ISTAT and IACS data.

Table 2 – *Fruit area in 2014 - Comparison among sources (hectares)*

Source/Cultivation	Nuts*	Peers	Peaches	Other fruit	Total Fruit
IACS					
Italy	**136.531**	**28.278**	**59.141**	**153.607**	**377.557**
North	21.191	26.098	24.323	87.825	159.437
Centre	32.346	576	2.829	26.487	62.238
South	82.995	1.604	31.988	39.295	155.883
Crop statistics					
Italy	**125.558**	**30.145**	**63.733**	**153.146**	**372.582**
North	15.598	23.756	20.823	73.382	133.559
Centre	19.665	907	4.088	13.233	37.893
South	90.295	5.482	38.822	66.531	201.130
% Difference (Italy)					
IACS vs crop statistics	**8,7**	**-6,2**	**-7,2**	**0,3**	**1,3**

Source: elaboration on ISTAT and IACS data. *Hazelnut, almond, pistachio.

Very similar results have been obtained comparing IACS data with the FSS (Farm Structure Survey) 2013. Structural FSS data derive from a sample of about 38.000 holdings and are characterized by sampling error not larger than ±5%. IACS data are 2,2% higher than FSS and differ from FSS especially as regards rice and citrus fruit, as already seen for crop statistics.

The discrepancy between IACS data and crop statistics are larger as regards specific kinds of fruit (table 2): the largest difference concerns nuts (8,7%), followed by peaches (7,2%) and peers (6,2%). However, comparability is not ensured, because nuts and peaches definitions adopted by IACS are not fully coherent with crops (some fruits may be included or excluded).

3. New estimation strategy for crop early estimates

3.1. Crop early estimates in Italy: from probabilistic to deterministic sampling

The last "Crop early estimates survey" (*Cees*) has been carried out between November 2015 and January 2016 through the CATI technique. It was aimed at interviewing a sample of 12.000 agricultural holdings for collecting early estimates regarding land use for agricultural purposes in the agrarian year (*ay*) 2015-16. Estimates concern the percent changes of land use between two agrarian years; they have been released at February 2016 and concerned the 5 categories requested from the EU Regulation 543/2009: common wheat, durum wheat, rye, barley, rape and turnip rape. The survey also included other kinds of crops. Since the main survey target is to produce estimates of changes between two following years, information on agricultural land use in the *ay* 2014-15 has been asked as well. The reference population is given by the agricultural holdings which had arable land at the end of 2015. Until 2015 (early estimates referred to the *ay* 2014-15) the estimation strategy was based on the two pillars: a) stratified random sample selected from the 2010 agriculture census list; b) the design-based Horvitz-Thompson estimator, with sampling weights adjusted for non responses.

Experimental methodological changes have been introduced in the last survey edition. Beyond the simplified questionnaire, they concern the sample selection and the estimation procedure.

Ad regards sampling, ISTAT switched from probabilistic to deterministic sampling. Instead of random selection from the not updated census list, two sub-samples including 6.000 units have been drawn from the subsets of respondents in the following surveys: *Cees* 2015 and FSS 2013. The samples were selected choosing the largest holders in each Italian Region which guaranteed at least the 80% of agricultural area surveyed in *Cees* 2015 and FSS 2013, with the additional constraint to guarantee at least 20 units for each combination of Region and kind of crop. This selection process simplified the further link between each sampling unit and its

certified electronic postal address (which in Italy is required by law in order to contact holdings), since the sample units had been already linked in the two previous surveys frame. As a matter of fact, The *Cees* 2016 response rate was 74,5%, against the 65,8% obtained in the *Cees* 2015. Another advantage consisted in shorter time needed for the complex data editing process: in the *Cees* 2016 it took about 4 weeks, against the 6 weeks spent in the *Cees* 2015 (as regards weights adjustment for tackling data editing see Gismondi and De Gaetano, 2015).

3.2. New estimation methodology

Let's suppose that Y_1 and Y_2 are the population totals at times 1 and 2. At time 1 a sample of *n* units is drawn from the population of size *N*. At time 2, $n\lambda$ units are kept into the sample, while $n(1-\lambda)$ are rotated. At time 1 we have the estimator $\bar{y}_1$, that is the sampling mean of units observed at time 1. We can define as $\bar{y}_1'$ the mean of the $n\lambda$ units that remain in the sample at time 2. We can derive two estimators of the mean at time 2, that are $\bar{y}_2'$ (units that responded also at time 1) and t_2'' (units that did not belong to the sample at time 1). According to double sampling, we can define the regression estimator of the total at time 2:

$$\hat{Y}_{2r} = N[\bar{y}_2' + \hat{\beta}(\bar{y}_1 - \bar{y}_1')] \quad (1)$$

where $\hat{\beta}$ is the regression coefficient estimate calculated on the $n\lambda$ units that remain in the sample at time 2. Its sampling variance is $V(\hat{Y}_{2r})$. Furthermore, if t_2'' is an estimator of the mean at time 2, with sampling variance $V(\bar{y}_2'')$, we can use the combined estimator given by:

$$\hat{Y}_{2c} = \phi \hat{Y}_{2r} + (1-\phi) t_2''. \quad (2)$$

If $\hat{Y}_{2r}$ and t_2'' are unbiased, then (2) is unbiased as well. Model (2) is widely used in Small Area Estimation (Rao, 2003, 2010) and in the estimation of a total deriving from a multiple frame survey (Lohr and Rao, 2006). Since the two combined estimators are independent (they are based on different sub-groups of units), the optimal choice of the shrink factor is $\phi_0 = V(t_2'')/[V(\hat{\bar{Y}}_{2r}) + V(t_2'')]$ (Rao, 2003). As a consequence, the minimum variance unbiased combined estimator and its variance will be given by, respectively:

$$\hat{Y}_{2c}{}^{*} = \phi_0 \hat{Y}_{2r} + (1-\phi_0) t_2{}'' \quad \text{and} \quad V(\hat{Y}_{2c}{}^{*}) = \frac{V(t_2{}'')V(\hat{Y}_{2r})}{V(\hat{Y}_{2r}) + V(t_2{}'')}. \tag{3}$$

If the sample is deterministic, the estimators (1), (2) and (3) can derive from a model based approach, as well as the related "model" variance *V*.

A more complex version of the estimation strategy (3) has been used by Preston (2015). The strategy can be adapted to *Cees*. Let's indicate as *Y* the total surface used for a certain cultivation, while *m* is the overall sample size at time 2. Since respondent units were asked to provide data as regards time 1 as well, *m* is the sample size at time 1 as well. For each holding, "surface" is the sum of surfaces used for any kind of crop surveyed. Furthermore we define as:

1) $n\lambda$: the number of units which declared positive surface at both times 1 and 2;
2) $n(1-\lambda)$: the number of units with positive surface at time 2 and surface equal to zero at time 1; therefore, the overall number of units which declared positive surface at time 2 is *n*;
3) *m*-*n*: the number of units which declared surface equal to zero at time 2.

As regards the β estimation, for any agricultural holding *i* among the $n\lambda$ which declared positive surface *y* at both times 1 and 2, we assume that the observed data follow a model ξ:

$$y_{2i} = \beta y_{1i} + \varepsilon_i \qquad \text{where:} \qquad \begin{cases} E_\xi(\varepsilon_i) = 0 & \forall i \\ V_\xi(\varepsilon_i) = \sigma^2 y_{1i} & \forall i \\ Cov_\xi(\varepsilon_i, \varepsilon_j) = 0 \quad \text{if} & i \neq j \end{cases} \tag{4}$$

where expected values, variances and covariances refer to the model ξ, with β and σ^2 unknown parameters. The properties of the estimates are thus analyzed under a super-population approach. The *BLUP* of β (Cicchitelli *et al.*, 1992, 385-390) is the ratio estimator:

$$\hat{\beta}^{*} = \bar{y}_2^{'} / \bar{y}_1^{'}. \tag{5}$$

As regards the estimator $t_2{}''$, it may be the sample mean $t_2{}'' = \bar{y}_2{}''$. Another approach consists in the use of a different model as regards the $n(1-\lambda)$ agricultural holdings which declared zero surface at time 1 (García and Labeaga, 1996). We can suppose the alternative model φ:

$$y_{2i} = \gamma z_i + \delta_i \qquad \text{where:} \qquad \begin{cases} E_\varphi(\delta_i) = 0 & \forall i \\ V_\varphi(\delta_i) = \theta^2 z_i & \forall i \\ Cov_\varphi(\delta_i, \delta_j) = 0 \quad \text{if} & i \neq j \end{cases} \tag{6}$$

where ***z*** is a not null auxiliary variable available for all the population units. According to (1), (5) and (6), we can calculate the estimator:

$$\hat{t}_2'' = N[\bar{y}_2'' + \hat{\gamma}(\bar{z} - \bar{z}'')] \qquad \text{where} \qquad \hat{\gamma}^* = \bar{y}_2'' / \bar{z}''. \tag{7}$$

In the *Cees* framework, *z* is given by agricultural surface referred to 2010 as derived from the last agriculture census. The table 3 resumes the five estimations strategies compared in the empirical attempt whose results have been resumed in section 3.3. The strategy used until the *Cees* 2015 is (I). The new strategy definitively applied in *Cees* 2016 is (IV).

Table 3 – *Compared estimation strategies for crop early estimates*

Code	Methodology	Estimator time 1	Estimator time 2
(I)	Sample mean expansion	$N\,\bar{y}_1$	$N\,\bar{y}_2$
(II)	Sample mean expansion using only units with positive surfaces at both times	$N\,\bar{y}_1'$	$N\,\bar{y}_2'$
(III)	Use of (2) where ϕ=1	Crop statistics	$N[\bar{y}_2' + \hat{\beta}(\bar{y}_1 - \bar{y}_1')]$
(IV)	Use of (2) where $t_2'' = \bar{y}_2''$, $\phi = \phi_0$	Crop statistics	$\phi_0 \hat{Y}_{2r} + (1-\phi_0)\,\bar{y}_2''$
(V)	Use of (2) where t_2'' is calculated as defined in (7), $\phi = \phi_0$	Crop statistics	$\phi_0 \hat{Y}_{2r} + (1-\phi_0)\hat{t}_2''$

3.3. Main results

The five estimation strategies have been applied to the *Cees* 2016. Even though estimates refer to the overall surface, the main target of *Cees* is the estimation of % change of surfaces used in two following agrarian years, as shown in the first row of table 4; for each strategy, the second row (figure in brackets) displays the Coefficient of variation (*Cv*) of estimates. If we suppose bias equal to zero for any strategy, the relative estimation error is given by $Cv = 100\sqrt{\hat{MSE}}/\hat{T}$, where *MSE* is the Mean Squared Error and $\hat{T}$ is the agricultural area estimate. Results concern the 5 most relevant cereals in Italy, which explain the 30% of the arable land.

Table 4 - *Main results of compared estimation strategies (agrarian year 2015-16) – Crops from arable land area % changes and coefficient of variation (Cv) of estimates*

Strategy	Arable land	Common wheat	Durum wheat	Barley	Oat	Grain Maize	Sum of 5 crops
(I)	-0,3	-1,6	-0,5	2,1	7,4	-3,0	-0,8
	(3,6)	(8,9)	(15,7)	(14,5)	(12,6)	(17,7)	(7,9)
(II)	0,9	2,5	2,3	3,3	9,1	-5,1	1,0
	(4,4)	(9,5)	(11,6)	(15,0)	(11,9)	(17,5)	(7,3)
(III)	0,5	1,5	0,7	0,8	4,2	-2,5	0,3
	(4,8)	(9,5)	(14,8)	(15,3)	(15,1)	(16,7)	(7,9)
(IV)	**2,4**	**5,6**	**6,2**	**6,9**	**11,2**	**-3,9**	**3,8**
	(2,7)	**(7,8)**	**(9,2)**	**(9,5)**	**(8,4)**	**(13,4)**	**(5,4)**
(V)	2,9	6,2	7,1	9,5	10,0	-4,3	4,6
	(2,8)	(8,3)	(10,1)	(9,3)	(9,0)	(15,5)	(5,8)

Source: elaboration on ISTAT data. CVs are into squared brackets.

Since the sample 2016 was not selected under a probabilistic approach, the first two strategies (based on the sample means) have sense only if we use them (surface estimates at times 1 and 2) for calculating the year to year % change (the expansion factor *N* disappears). The use of strategies 1 (summing up all available responses), 2 (summing up data of units with positive surface at both times) and 3 (regression estimator) lead to small changes: estimates are near to zero, with the only exception for oat. Strategies 4 and 5 are the only ones which use data on surfaces larger than zero at time 2 (forecasts for the agrarian year 2015-2016) declared by the agricultural holdings which had zero surface at time 1 (agrarian year 2014-2015), through combination of these data with the regression estimator. As a matter of fact, strategies 4 and 5 lead to larger % changes estimate just for this reason; on average strategy 4 is characterized by the smallest MSE (3,8% for the sum of 5 cereals and 2,4% for the whole arable land).

4. Conclusions

As regards crop statistics, results show that administrative data collected by the Italian agency for payment in agriculture can be used for statistical purposes. Even though the analysis does not concern yield, was carried out along 2 years only and the empirical attempts have been limited to some kinds of crops, the overall reliability of the database is satisfactory. Further work should concern: a) extension of the database to 2015 and to other cultivations; b)

methods for producing estimates based on IACS early declaring holdings, in order to satisfy the time deadlines imposed by the EU legislation. As regards early estimates, the sampling design may be based on a deterministic approach, coupled with a model based estimation technique. The presence of many zeroes implies the use of specific models whenever the traditional regression model may fail. Quality of both administrative data and model based estimates must be evaluated according to agreed international standard indicators.

5. References

Cicchitelli, G., Herzel, A. and Montanari, G.E. (1992), Il campionamento statistico, Il Mulino, Bologna.

García, J. and Labeaga, J. M. (1996), Alternative Approaches to Modelling Zero Expenditure: An Application to Spanish Demand for Tobacco, Oxford Bulletin of Economics and Statistics, 58(3), pp.489–506.

Gismondi, R. and De Gaetano L.(2015), Sampling Weights Adjustment for Improving Crops Early Estimates Precision, Proceedings of the ISI2015 Congress, www.isi-web.org.

Kloek, W. and Vaju, S. (2013), The Use of Administrative Data in Integrated Statistics, Http://essnet.admindata.eu.

Lohr, S. and Rao, J. N. K. (2006), Estimation in Multiple-Frame Surveys, Journal of the American Statistical Association, Vol. 101, No. 475, pp.1019-1030.

Preston, J. (2015), Modified Regression Estimator for Repeated Business Surveys with Changing Survey Frames, Survey Methodology, 41-1, pp.79-97.

Rao, J. N. K. (2003). Small Area Estimation. Wiley, Hoboken, New Jersey.

Rao, J. N. K. (2010) Small-area Estimation with Applications to Agriculture, in Agricultural Survey Methods (eds R. Benedetti, M. Bee, G. Espa and F. Piersimoni), John Wiley & Sons, Ltd, Chichester, UK. doi: 10.1002/9780470665480.ch9.