Management Theory and Studies for Rural Business and Infrastructure Development
ISSN 1822-6760 / eISSN 2345-0355
2014. Vol. 36. No. 4: 835–843
doi: 10.15544/mts.2014.078

ESTIMATION OF OLIVE OIL PRODUCTION BASED ON THE USE OF ADMINISTRATIVE DATA

Roberto Gismondi[1], Loredana De Gaetano[2], Massimo Russo[3]

[1] *Research manager. Italian National Statistical Institute. E-mail gismondi@istat.it*
[2] *Researcher. Italian National Statistical Institute. E-mail degaetan@istat.it*
[3] *Researcher. University of Foggia. E-mail massimo.russo@unifg.it*

Received 01 09 2014; accepted 08 11 2014

Italy is one of the most relevant countries in the European Union as regards crops and, in particular, olive oil production. Actually, statistical data on olive oil are provided by the 20 Italian Regions to ISTAT through estimates supplied by experts and / or local panels of influent farmers and category associations. However, more precise estimates may be derived from the administrative data collected by AGEA, which is the Italian acronym for Integrated Administration and Control System (IACS). IACS is in charge of receiving from farmers declarations of production, on the basis of which they receive subsidies. The main goals of the work are the following ones: 1) to transform the administrative IACS database into a statistical one; 2) to compare different methodologies for producing monthly estimates of olive oil productions in presence of missing declarations; 3) to aggregate the IACS yearly olive oil production, comparing these estimates with those supplied by experts estimates. The paper analyses the main outcomes concerning the Region Apulia (South Italy), where the 40.1% of Italian olive oil is produced.

Keywords: agriculture, crops statistics, economic statistics, eurostat, IACS, official statistics, oil, olives.
JEL Codes: Q01, Q15, Q56.

1. Introduction

Crops statistics are the most important production indicator in agriculture, beyond animal slaughtering as well as meat and milk derived products. Italy is one of the most relevant countries in the European Union as regards crops and, in particular, olive oil production, widely used and recommended by many food safety campaigns. Actually statistical data on crops – and olive oil as well – are provided by the 20 Italian Regions to ISTAT through estimates supplied by experts and/or local panels of influent farmers and category associations. ISTAT produces regional data on olive oil through the “estimative” technique: data are obtained multiplying surface data by the average production x hectare (EUROSTAT, 2014).

However, more precise estimates may be derived from the administrative data collected by AGEA, which is the Italian acronym for IACS (Integrated Administration and Control System). IACS is in charge of receiving from farmers declarations of production, on the basis of which they will receive EU subsidies. In the period 2011–2013 in Italy about 7000 olive pressers declared to have produced olive oil.

Given these premises, the main goals of the work are the following ones: 1) to transform the administrative IACS database into a statistical one, linking the single olive pressers to the ISTAT farm register; 2) to compare five different methodologies for producing monthly estimates of olive oil productions in presence of missing declarations in the time series analysis carried out at the single unit (olive presser) level, comparing estimates with those supplied by Regions, based on experts estimates. The overall approach adopted is founded on the same criteria suggested, for instance, by P. Falorsi et al. (2003) and M. Brodeur (2006).

The results encourage the integration of administrative IACS data into official statistics provided by ISTAT starting from the reference year 2014, substituting Regions estimates in order to increase quality of estimates (Kloek, 2013).

Topics dealt with are as follows: section 2 resumes the main features of the AGEA database; section 3 describes the various techniques applied for transforming the administrative database into a statistical tool for producing current estimates of Italian oil production; section 4 comments the result concerning the Italian Region Apulia; perspective conclusions have been proposed in section 5.

2. The AGEA database: empirical evidence

Administrative declarations available for the years 2011, 2012 and 2013 were merged into a unique database, where each unit (row) was given by a specific olive presser operator. Overall, in Italy there are 6.888 olive pressers which declared olive oil production different from zero in at least one year in the period 2011–2013 (table 1). The sector is very concentrated, because the 80% of olive oil is due to the 34.3% of operators.

Through record linkage with the 2012 ISTAT business register (which includes enterprises, but does not include agricultural holdings which are not enterprises) and with the list of agricultural holdings interviewed in the last Agriculture Census (2010), it was assessed that the 74.8% of olive pressers are pure enterprises (this share was equal to 75.9% in the Apulia region) and that the 32.4% are agricultural holdings (30.6% in Apulia).

Overall, in Apulia operate 1303 olive pressers, of which 893 declared production in 2011, 1.140 in 2012 and 915 in 2013, so that the response rate ranged from about 70% in 2011 and 2013 and raised to 87.5% in 2012. If we consider 2012 and 2013 years nearer to the “normal” situation, broadly speaking almost all the units pressed olives in a number of months ranging from 1 to 7 (table 1), since the share of olive pressers which declared production in not more than 7 months was 98.9% in 2012 and 98.4% in 2013.

Table 1. Percent of units declaring olive oil production Apulia Region

Number of months	% compositions			% cumulative compositions		
	2011	2012	2013	2011	2012	2013
Declare	100.0	100.0	100.0	100.0	100.0	100.0
	Declarations by month (1)			Cumulative declarations %		
1	5.2	13.7	5.5	5.2	13.7	5.5
2	70.7	33.1	14.6	75.9	46.8	20.1
3	23.2	22.7	28.0	99.1	69.5	48.1
4	0.3	11.8	20.8	99.4	81.2	68.9
5	0.4	11.1	12.7	99.9	92.4	81.5
6	0.1	4.8	10.4	100.0	97.2	91.9
7	0.0	1.8	6.4	100.0	98.9	98.4
8	0.0	0.6	0.8	100.0	99.6	99.1
9	0.0	0.1	0.2	100.0	99.6	99.3
10	0.0	0.3	0.2	100.0	99.9	99.6
11	0.0	0.1	0.1	100.0	100.0	99.7
12	0.0	0.0	0.3	100.0	100.0	100.0

Total number of declarations = 100%. *Source*: elaboration on IACS-ISTAT data

The most common period starts in September and ends in March next year. As a consequence, when monthly declarations are summed up in order to achieve to the whole yearly production, missing data in correspondence of months from April to August may be considered equivalent to zero production (no imputation is carried out). On the other hand, since in Apulia (as well as in Italy) three olive pressers on four are pure enterprises – which carry out this activity as unique, main or secondary – it is difficult to suppose that a missing declaration in certain year may be due to no production, while it is quite certain that the missing value is a pure non response due to a missing declaration. The most part of no declarations ("zeros") happened in 2011, which was probably the starting year of administrative data collection and paid the gap derived from the progressive compliance of olive pressers with their administrative duties.

3. Estimation procedures

The use of IACS data for estimation olive oil production in presence of non responses and other non sampling sources of errors may be tackled according to 2 approaches:

- macro approach: corrections are applied on total figures derived from the database;
- micro approach: corrections are applied to single micro-data referred to specific olive pressers.

On the basis of approach 1, at the macro level there are two general models which may be implemented in order to achieve to final estimates of the amount of production *y* referred to the period *T*. The first one is a level estimate based on transformation of the original macro data (labeled as *IACS*) according to the following formula:

$$\hat{y}_T = \vartheta_{Def,T}\left[y_{IACS,T}\left(1 + \alpha_{Under,T}\right) + \beta_{Hidden,T}\right] + \varepsilon_T \qquad (1)$$

The formula (1) is based on the following assumptions regarding potential causes of difference between the IACS macro-data $y_{IACS,T}$ (obtained summing up all the available micro-data referred to the period T) and the "true" amount $\hat{y}_T$ to be estimated:

a) different definitions and/or observation fields: their effect may be resumed by means of the parameter $\vartheta_{Def,T}$, which may be higher or lower than one (for instance: if the IACS data do not cover cooperatives as well, then the parameter will be > 1);

b) under-declaration of respondents, due to the fact that olive pressers which declared production did not declare the whole true production but only a part of that. As a consequence, the parameter $\alpha_{Under,T} > 0$, otherwise it will be < 0 if over-declarations are more relevant than under-declarations;

c) total "hidden" production, which may happen if some olive pressers do not declare at all. The effect may be resumed through the parameter $\beta_{\mathrm{H}idden,T}$;

d) other errors (potential duplications, measurement errors, etc.), which may be expressed by means of the residual error term ε_T.

The model (1) implies the estimation of 3 parameters, of which 2 may change monthly. In addition, the effect of the residual error term ε should also be evaluated. Its complexity hampers the effective possibility to implement that. This risk is quite common when administrative data are dealt with (Daas, 2008).

The second method tackles the implementation problems of method (1). It is based on the formula:

$$\hat{y}_T = \hat{y}_{T-1}\left(\frac{y_{IACS,T}}{y_{IACS,(T-1)}}\right) \qquad (2)$$

The main rationale of formula (1) is founded on the logic underlying calculation of index numbers. We suppose that for a certain previous time (*T*–1) a reliable estimate $\hat{y}_{T-1}$ of the true amount of production is available. Afterwards, the ratio between the amounts of production derived from IACS data referred to the periods *T* and (*T*–1) may be used for updating the level $\hat{y}_{T-1}$ to the new level $\hat{y}_T$. This simple technique avoids the estimation problems of method (1), but requires the estimate of the „true" base level and loses precision as *T* grows and become quite far from (*T*–1). The periodic new estimate of a „true" level (*T*–1) will imply periodic revisions of previous estimates calculated before the update.

The micro approach supposes to apply corrections to single micro-data and was the one used in the application to real data. The choice in favour of the micro approach was due to the not possibility to apply the macro approach at this stage and to the possibility to use a wide set of single pressers micro-data. In this framework, we suppose that a certain unit *i* did not respond in the period *T*. For the estimation of the unknown amount $\hat{y}_T$, five methods have been defined and compared.

Method 1.

No imputation: totals for any period *T* are obtained summing up data concerning all the *n* respondent units at time *T*. This method can be used as benchmark for evaluating effects of the methods and can be expressed through the formula:

$$\hat{y}_T = \sum_{I=1}^{n} y_{T,i} \qquad (3)$$

Method 2.

No imputation and use of "panel" units only: totals for any coupled of periods *T* and (*T*+1) are obtained summing up data concerning all the n_{P} units respondent at both times *T* and (*T*+1). We will label as Panel (*P*) these units. The formulas are:

$$\hat{y}_T = \sum_{1=1}^{n} y_{T(P),i} \;\; ; \;\; \hat{y}_{T+1} = \sum_{1=1}^{n} y_{(T+1)(P),i} \qquad (4)$$

It is worthwhile to note that using in a recursive way formulas (4), for the same year there will be two different estimates, based on a different number of panel units.

Method 3.

Imputation of each non response using the amount available, for the same unit *i*, in the previous or in the next period, corrected by means of the average ratio of amounts calculated on the only panel units *P*. The main rationale which justifies this ratio-method approach is founded on the hypothesis of a linear super-population model which explains levels at time *T* on the basis of linear regression through the origin with respect to levels at times (*T*–1) of (*T*+1) and model variances proportional to levels (Cicchitelli, 1992). The related formulas are:

$$\hat{y}_{T,i} = y_{T-1,i}\frac{\bar{y}_{(P),T}}{\bar{y}_{(P),T-1}} \;\; \text{(5a) or } \;\; \hat{y}_{T,i} = y_{T+1,i}\frac{\bar{y}_{(P),T}}{\bar{y}_{(P),T+1}} \qquad \text{(5b)}$$

where $\bar{y}_{(P)}$ is an average amount calculated on panel units. The second formula can be used if the amount at time (*T*–1) is not available. Of course, in real practice only the first formula can be used (since at time *T* we cannot know future amounts at time (*T*+1)).

Method 4.

Estimates obtained with the method (3) are corrected through the ratio between turnover *x* of unit *i* at time *T* and the average turnover calculated on the same panel units *P* used in formulas (4). This method can be applied only for units whose turnover is available, otherwise method 4 will reply method 3. We have the formula:

$$\hat{y}_{T,i*} = \hat{y}_{T,i}\frac{x_{T,i}}{\bar{x}_{(P),T}} \qquad (6)$$

Method 5.

Estimates obtained through method 4 are corrected if the percent change between times *T* and (*T*–1) is outside the acceptation range [–100%;+100%]. If the check is not satisfied, the amount at time T corresponding to –100% or +100% changes is imputed, according to the Winsorization criterion (Hasings, 1947).

The imputation procedure applied for the 3 years 2011, 2012 and 2013 has been resumed in the table 2.

Table 2. Imputation procedure used for the years 2011–2013. Apulia Region

Cases (No = non response)			Number	Imputation procedure		
2011	2012	2013		2011	2012	2013
Yes	Yes	Yes	524	–	–	–
No	Yes	Yes	333	Formula (5b) T = 2012	–	–
Yes	No	Yes	3	–	Formula (5a) T = 2013	–
Yes	Yes	No	266	–	–	Formula (5b) T = 2013
No	No	Yes	57	Formula (5a) T = 2012 [2]	Formula (5a) T = 2013 [1]	–
Yes	No	No	101	–	Formula (5b) T = 2012 [1]	Formula (5b) T = 2013 [2]
No	Yes	No	19	Formula (5a) T = 2011	–	Formula (5b) T = 2013

Note: the figures in square brackets indicates the priority of imputation

Overall, in 524 cases no imputation was needed (units responded in all the 3 years). In the 333 cases when an olive presser did not declare in 2011, but declared in 2012 and 2013, the production in 2011 was estimated using formula (5b) putting T = 2012. In the 3 cases when an olive presser declared in 2011 and 2013, but did not declare in 2012, the production in 2012 was estimated using formula (5b) putting T = 2012: this formula was preferred to the potential use of formula (5a) with T = 2012, since we have supposed that data reliability improved along time, so that 2013 date were considered more suitable for carrying out imputation rather than 2011 data. The same rationale was applied in the other cases in the table. At the end of the imputation process, the database included only not empty cells for any unit and any year and was used for carrying out final estimates.

4. Main results

All methods described so far have been applied separately within specific estimation domains. Given a certain Region, estimation domains have been obtained crossing each Province belonging to the Region with 3 strata inside the Province. These strata have been calculated using the variable "average olives production in the period 2011–2013", according to which olive pressers inside the same Province were sorted and split in 3 subgroups, whose lower and upper limits have been identified in order to have that olive production inside each stratum was equal to one third of the overall province production. The exercise commented concerns the Region Apulia, which produces the 40,1% of Italian olive oil (table 3). It is quite clear how the first method (use of respondents' data only) implies biased estimates for 2011.

Table 3. Results of estimates obtained through different methods Apulia Region

Year	Production (Tons)			% changes	
	IACS	Crops statistics		IACS	Crops statistics
Method 1. All respondents without imputation					
2011	148.440	185.072			
2012	200.079	190.160		34,8	2.7
2013	194.859	184.826		-2,6	– 2,8
Method 2. Panel units (couples of years) (*)					
2011	143.329	185.072			
2012	155.979	190.160		8,8	2,7
2012	180.684	190.160			
2013	190.210	184.826		5,3	– 2,8
Method 3. Imputation using historical data					
2011	208.921	185.072			
2012	206.854	190.160		-1,0	2.7
2013	218.125	184.826		5,4	-– 2,8
Method 4. Imputation using historical data and number of persons employed					
2011	206.793	185.072			
2012	211.067	190.160		2,1	2.7
2013	219.009	184.826		3,8	– 2,8
Method 5. Imputation and outliers correction					
2011	207.955	185.072			
2012	211.067	190.160		1,5	2.7
2013	221.063	184.826		4,7	– 2,8

(*) On the basis of method 2, two different estimates for 2012 will be available.

Source. Elaboration on IACS-ISTAT data.

That is confirmed by comparison with the current crops statistics elaborated by ISTAT on the basis of regional experts estimates; however, in 2012 and 2013 the two sources (IACS and ISTAT) converge towards more aligned figures, especially in 2013, when the percent changes with respect to 2012 are almost the same (respectively, – 2,6% and – 2,8%, table 2).

The use of panel units leads to larger differences between the percent yearly changes obtained with the two sources, and the recourse to method 3 (imputation using historical data as explained in the previous section) does not change things too much. The most significant improvements have been obtained using method 4 (imputation corrected through the yearly turnover auxiliary variable), since in this case the percent change 2012 / 2011 is equal to +2,1% with the IACS source and to +2,7% with the ISTAT source, so that the difference among them is quite small. Finally, method 5 (based on method 4 plus outlier changes correction) does not seem to provide clear improvements.

Even though comparison between the two sources is only one of the tools for assessing reliability of estimates, the application provides encouraging outcomes, which of course need to be replied in other regions and along a longer period.

5. Perspective conclusions

1. The results show that administrative data collected by the Italian agency for payment in agriculture can be used for statistical purposes. Even though the analysis was carried out along 3 years only and the empirical attempts carried out so far have been limited to Apulia, the overall reliability of the database seemed to be quite satisfactory. Moreover, administrative data show more regular seasonal patterns and yearly production fluctuations than the Regions estimates actually used.

2. The most reliable estimation methodology was based on imputation of monthly missing declarations based on the ratio estimator coupled with an outlier detection procedure. The work should be continued on the basis of further steps, among which:

- extension of the database length to 2014;
- replication of the above mentioned simulations to all the Italian Regions;
- estimation of the hidden production, as formalized in the equation (1);
- elaboration of indicators able to measure quality of estimates obtained through administrative sources.

References

Brodeur, M. (2006). Use of Tax Data in the Unified Enterprise Survey (UES). – http://millenniumindicators.un.org/unsd/economic_stat/Web/PDF/Canada%20-%20Use%20of%20tax%20data%20in%20the%20UES-E.pdf [2014 10 01].

Cicchitelli, G., Herzel, A., Montanari, G. E. (1992). Il campionamento statistico. Bologna: Il Mulino. 385 p.

Daas, P. J. H., Ossen, S. J. L., Vis-Visschers, R. J. W. M., Arends-Toth, J. (2008). Quality Framework for the Evaluation of Administrative Data. Q2008 European Conference on Quality in Official Statistics, Roma.

EUROSTAT. (2014). Eurostat Handbook for Annual Crop Statistics (Regulation 543/2009). Luxembourg: Eurostat.

Falorsi, P. D., Pallara, A., Russo, A. (eds.). (2003). Temi di ricerca ed esperienze sull’utilizzo a fini statistici di dati di fonte amministrativa. – Milano: Franco Angeli. 30 p.

Hasings, C., Mosteller, F., Tukey, J. W., Winsor, C.P. (1947), Low moments for small samples: a comparative study of order statistics // *Annals of Mathematical Statistics*. 413–426 p.

Kloek, W., Vaju, S. (2013), The Use of Administrative Data in Integrated Statistics. – http://www.google.it/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCEQFjAA&url=http%3A%2F%2Fessnet.admindata.eu%2FDocument%2FGetFile%3FobjectId%3D5821&ei=FpJgVKqUENfZat7HgsgH&usg=AFQjCNFCNc7MYuiCugbgs84FWnOOF3Ppog&bvm=bv.79189006,d.d2s [2014 09 09].

ALYVUOGIŲ ALIEJAUS GAMYBOS APSKAIČIAVIMAS REMIANTIS ADMINISTRACINIAIS DUOMENIMIS

Roberto Gismondi[1], Loredana De Gaetano[2], Massimo Russo[3]
Italijos nacionalinis statistikos institutas

Įteikta 2014 09 01; priimta 2014 11 08

Santrauka

Italija yra viena Europos Sąjungos šalių, kur alyvuogių aliejaus gamyba yra labiausiai išvystyta. Straipsnyje analizuojami Apulia (Pietų Italija) rezultatai; šiame regione pagaminama 40,1 procentų viso Italijoje gaminamo alyvuogių aliejaus. Pagrindiniai tyrimo tikslai yra: 1) transformuoti administracinę IACS (integruota administravimo ir kontrolės sistema) duomenų bazę į statistinę duomenų bazę; 2) palyginti skirtingas metodikas alyvuogių aliejaus gamybos mėnesiniam apskaičiavimo atlikimui; 3) kaupti metinius alyvuogių aliejaus gamybos duomenis IACS, lyginant šiuos apskaičiavimus su pateiktais ekspertų. Statistinius duomenis apie alyvuogių aliejaus gamybą teikia Italijos Nacionalinės Statistikos Institutas (ISTAT) iš viso 20 Italijos regionų; jie ruošiami pagal ekspertų ir / ar vietinių specialistų grupių apskaičiavimus gavus duomenis iš ūkininkų ir asociacijų. Tačiau tikslesni skaičiavimai gaunami iš AGEA (tai yra Itališkas IACS akronimas) surinktų administracinių duomenų.

Raktiniai žodžiai: žemės ūkis, pasėlių statistika, ekonominė statistika, Eurostatas, IACS, oficiali statistika, aliejus, alyvuogės.

JEL kodai: Q01, Q15, Q56.