Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (i): Selective and macro editing

Model-based selective editing procedures for agricultural price indices

Prepared by Pichiorri T., Ichim D., Ferraro M.L., Guarnera U.
Istat, Italy

I. Introduction

1. In this paper we address the problem of estimating agricultural prices indices of products and means of production. The main use of agricultural prices is to compare the prices levels, to analyse the price developments and trends and to study their effect on agricultural income. They also allow agricultural policies to be determined and monitored.

2. In Italy, data on the agricultural prices is monthly collected by the Chambers of Commerce in 85 provinces and transmitted to the Italian National Statistical Institute (Istat) which is in charge with the remaining data treatment and processing phases, including data validation. Big efforts are spent in the editing and imputation phase due to the presence of measurement errors in the elementary price indices. In particular, outlying observations are carefully inspected by manual reviewing in order to avoid that the quality of the output target estimates is affected by gross errors. Since interactive editing is an expensive activity in terms of time and resources involved, the implementation of more automatic procedures becomes mandatory, given the budget restrictions generally faced by statistical authorities. The aim of the current Istat project is to evaluate the implementation of automatic editing procedures able to identify the cases with highest expected benefit.

3. In this work we present the results obtained by the application of model-based selective editing strategies, as implemented in the R package SeleMix (Di Zio and Guarnera, 2013). In section II, the data collection process is described, together with several significant features of the agricultural indices. Among these features, one can emphasize the reduced number of observations, the range of variability together with the non-negligible percentage of missing values. The contruction of agricultural price indices is very detailed since, in our opinion, this problem is approached for the first time in automatic data editig framework. In section III, the theoretical foundations of the selective editing methodology based on contamination models are illustrated. Moreover, the link betwween the complex indices and the methodology implemented in the R package SeleMix is fully justified. In section IV several implementation strategies are introduced; they are based on the exploitation of different temporal lags or hierarchical levels. The results obtained in case of agricultural products and means of production are reported. For comparison, results from an application of a simpler method based on the assumption of stability of the monthly price variation over consecutive years are also included in the paper. The criteria used for comparing the editing procedures are based on the convergence of the algorithm, on the

number of identified influential values and the comparison of the impact on official published (aggregated) agricultural price indices. Finally, in Section V, some conclusions and directions for further developments are discussed.

II. Agricultural price indices: the data

4. The aim of the Istat project is to implement and test the automatic selective editing procedure proposed in (Di Zio and Guarnera, 2013). The application concerns the data referring to the composite agricultural price indices. In this section we describe the data collection method, the rules governing the computation of the composite agricultural indices and some particular features of the data. During the project, the data reference period 2013 was used.

5. About 200 agricultural products and means of production are constantly monitored by Istat as this information is used for evaluating and determining agricultural policies. 85 Chambers of Commerce transmit to Istat the monthly prices p corresponding to agricultural products. For each agricultural product, the elementary price index ${}_{prov,v}I^{t}_{r,a}$ is derived by dividing the current price by the average price base year, where t is the observation period, a denotes the product, r denotes the geographical region (NUTS2 classification), while $prov,v$ represent the province and the variety of the product observation respectively. Of course, not all the products (and their varieties) are observed in all the provinces/regions. Overall, there are 3081 and 3359 observations monthly registered by Istat for means of production and agricultural products, respectively. In this work the base year for the computation of indices is 2010, according to the European regulations. After an initial deterministic check and validation of the quantities and measurement units, the elementary indices are computed for each product variety at province level. The elementary price indices comprise the indices of producer prices of agricultural products (output products) and the indices of purchase prices of the means of agricultural production (input products). At the most detailed level, Istat monthly registers the elementary price indices for 114 means of production products (input) and 98 agricultural products (output). According to their natural characteristics, the agricultural products may be grouped in broader categories. At European level, this nested hierarchical classification of agricultural products is defined by Eurostat in cooperation with the Member States. In this work, the used hierarchical levels will be denoted by synt_3 and synt_4, the first being less detailed than the latter. Additionally, a further natural characteristics of the agricultural products is their seasonality. Indeed, 53 products among the 98 monitored by Istat are registered only in some months. Indeed, there are two products, namely cheeries and plants of Poinsettia which are registered in a single month. In average, the products are registered in about 5 months.

Products	**Number of products**	**Seasonal products**	**N. elementary monthly indices *not seasonal***	**N. elementary monthly indices *seasonal* (min - max)**
Means of production (synt_4)				
Seeds	20	-	491	-
Energy; lubricants	6	-	262	-
Fertilisers. Soil improvers	13	-	577	-
Plant protection products and pesticides	19	-	469	-
Veterinary expenses	1	-	122	-
Animal feeding stuffs	22	-	477	-

Other goods and services	15	-	402	-
Maintenance of materials	18	-	281	-
Total	**114**	**-**	**3.081**	**-**
Agricultural product (levels synt_3)				
Cereals	6	-	288	-
Industrial crops	3	-	41	-
Forage plants	3	-	100	-
Vegetables and horticultural products	41	32	184	313 - 458
Potatoes	2	1	38	0 - 14
Fruit	23	20	101	26 - 457
Wine	3	-	329	-
Olive oil	1	-	61	-
Animals	12	-	632	-
Animal products	4	-	97	-
Total	**98**	**53**	**1.871**	**361 - 909**

Table 1. Number of products and monthly elementary indices for agricultural product and means of production.

6. The elementary indices may be used to derive different indices. Such indices are generally used to perform temporal (annual or quarterly), regional, national or international comparisons. The monthly ($t, t = 1, ..., 12$) regional ($r, r = 1, ..., R = 20$) index is defined, for each product, as the arithmetic mean (a linear combination) of the elementary indices ${}_{prov,v}I^t_{r,a}$, i.e.

$$I^t_{r,a} = \frac{\sum_{prov}^{N.prov} \sum_{v}^{V} {}_{prov,v}I^t_{r,a}}{n^t_{r,a}} \tag{1}$$

where $N.prov$ denotes the number of provinces, V denotes the number of varieties of a certain product and $n^t_{r,a}$ indicates the cardinality of elementary indices for a certain product a. Subsequently, the monthly national index of a certain product a is computed, using the Laspeyres formula, as a linear combination of monthly regional indices:

$$I^t_a = \frac{\sum_r w_{r,a} I^t_{r,a}}{\sum_r w_{r,a}} \tag{2}$$

where $w_{r,a}$ is a regional weight a-priori assigned by Istat to each product in each Italian region.
They are fixed-base index numbers calculated using weighting coefficient referred to the base year.
Consequently, it is straightforward to obtain that the monthly national index of a certain product a is a linear combination of elementary price indices, i.e.

$$I^t_a = \frac{\sum_r w_{r,a} I^t_{r,a}}{\sum_r w_{r,a}} = \frac{\sum_r w_{r,a} \frac{\sum_{prov} \sum_{v} {}_{prov,v}I^t_{r,a}}{n^t_{r,a}}}{\sum_r w_{r,a}} = \frac{\sum_r \frac{w_{r,a}}{n^t_{r,a}} \sum_{prov} \sum_{v} {}_{prov,v}I^t_{r,a}}{\sum_r w_{r,a}} = \sum_r \sum_{prov} \sum_{v} \tilde{w}^t_{r,a} * {}_{prov,v}I^t_{r,a} \tag{3}$$

where $\tilde{w}^t_{r,a} = \frac{w_{r,a}}{n^t_{r,a} \sum_r w_{r,a}}$. The linear relationship between the national composite index and the elementary price indices is valid as the system of weights is designed according to the

recommendations mentioned in the Handbook for EU Agricultural Statistics, Eurostat (2008). The weights are computed by Istat in order to guarantee the coherence of the entire system of composite indices and the comparability at European level. Moreover, the regional weights are constant, independently on the considered month. The monthly, quarterly and annual indices are computed using another system of weights which is not discussed in this paper.

7. A significant issue of the data collection and validation process is represented by the extremely reduced number of observations concerning each agricultural product or means of production. Indeed, with respect to the 2013 reference period, the number of non-missing elementary indices monthly registered varies from 1 to 108 for means of production, while for the agricultural products the corresponding range of variation equals (1-193). In Table 1, some descriptive statistics of the number of useful observations are illustrated. The first column shows the number of levels of each hierarchical level of the agricultural products and means of production. The next three columns show the minimum, mean and maximum number of real-valued elementary indices registered for each hierarchical level of the agricultural products classification. The annual average percentage of missing values equals 14% for means of production and 20% for agricultural products. The missing values are not uniformly distributed among the products, as shown in Figures 1 and 2, which illustrate the monthly percentage of missing values for agricultural products and means of production respectively. The synt_3 level (Figure 1) and the synt_4 level (Figure 2) of the classification of products were used. The aggregates show very different trends and levels of non-response. Obviously, the composite national or regional indices cannot be computed in presence of missing values.

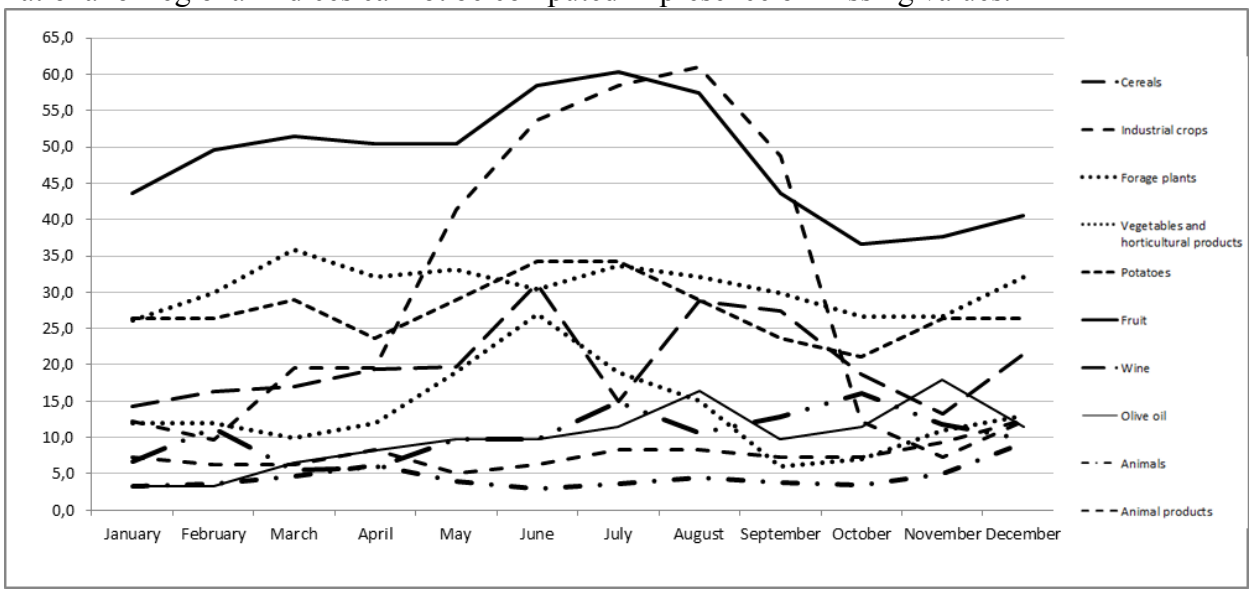

Figure 1. The monthly percentage of missing values for agricultural products.

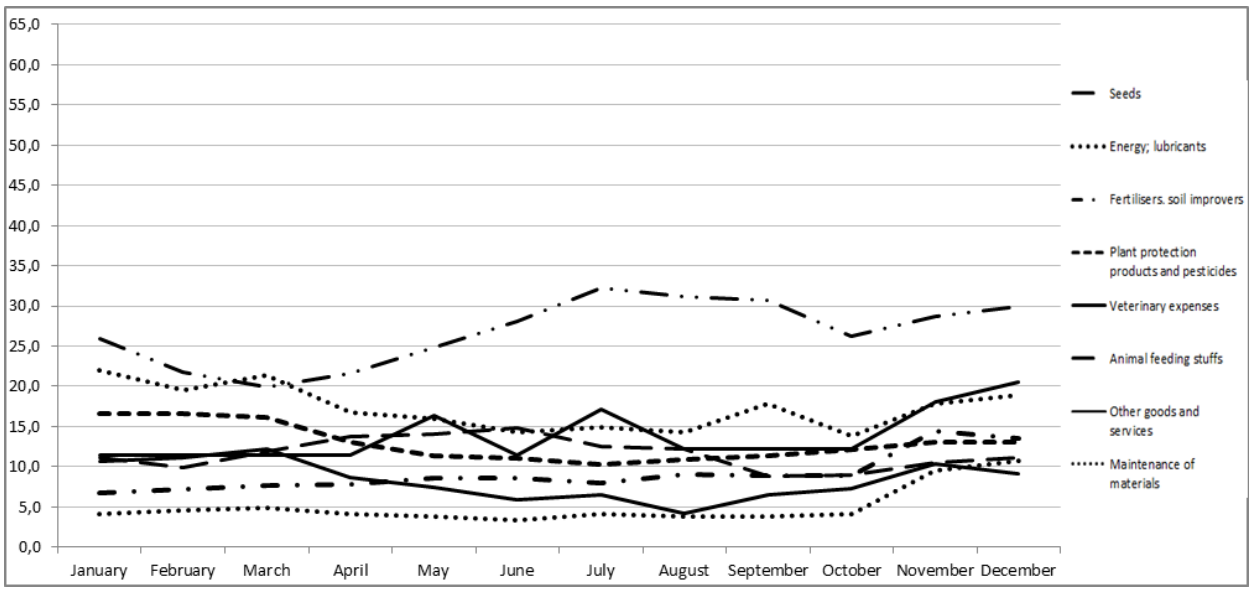

Figure 2. The monthly percentage of missing values for means of production.

8. Another striking feature is the variability of the underlying elementary indices. In Table 1, the minimum, mean and the maximum of the variance of the elementary indices computed for each month is shown, denoting the absolute need to apply robust data validation procedures. A manual validation procedure is implemented at Istat. It is an interactive procedure consisting in checking each elementary index whose absolute relative variation with respect to the previous month exceeds a given threshold. By contacting again the Chambers of Commerce, the elementary indices are changed, leading to more robust values. For example, the manual changes introduced for the agricultural product corresponding to the largest variation in Table 1 would correspond to a reduction in variance to 390. With respect to 2013 reference period, 1730 agricultural products and 175 means of production elementary indices were manually checked and validated. Unfortunately, this procedure is very expensive, thus Istat requires the implementation of a more automatic procedure. In this application, we take advantage of the "corrected" indices in order to evaluate the discussed selective editing methodology.

Means of production							
	#levels	**#observations**			**Variance**		
		Min	*Mean*	*Max*	*Min*	*Mean*	*Max*
Prod	114	1	23.45	108	0.125	406.56	26405.64
synt_3	2	219	1334.08	2465	107.55	476.39	3798.18
synt_4	8	97	333.52	538	107.55	466.97	3798.18

Agricultural products							
	#levels	**#observations**				**Variance**	
		Min	*Mean*	*Max*	*Min*	*Mean*	*Max*
Prod	98	1	29.52	193	0.32	1154.94	40818.76
synt_3	10	16	197.32	613	128.02	1226.21	2876.15
synt_4	26	4	81.34	432	23.14	963.28	5702.73

Table2. Descriptive statistics of the number of observations and the variance of the elementary indices.

III. SeleMix methodology

A. Selective Editing based on contamination model: SeleMix

9. The methodology adopted in the selective editing procedure for agricultural price indices has been recently developed in Istat and implemented in the R-package SeleMix. According to this methodology, both true data and error mechanism are modelled in a parametric framework. Specifically, given a dataset composed of n units, true data (or logarithms of true data) $\mathbf{Y}$ are assumed to be realizations from n Gaussian distributions with the same variance-covariance matrix $\mathbf{\Sigma}$ and mean vectors linearly dependent on a set of covariates $\mathbf{X}$ through some unknown set of parameters $\mathbf{B}$ (standard linear regression model of $\mathbf{Y}$ on $\mathbf{X}$ with normal residuals). The measurement process is supposed to be associated with an *intermittent* error mechanism, modelled through a Bernoullian random variable $\mathbf{I}$, with parameter w, assuming value 1 or 0 depending on whether an error occurs in data or not respectively. The parameter w can be interpreted as the *a priori* probability of an observation of being affected by an error. Conditional on $\mathbf{I}$=1 (presence of error), we assume a zero mean Gaussian additive error with covariance matrix proportional to $\mathbf{\Sigma}$, the proportionality constant being some positive number λ. This model is sometimes referred to as *contamination model* and implies that the distribution function for the observed data which is a mixture of Gaussian distributions having same means but proportional covariance matrix (the proportionality constant being $(\lambda + 1)$. Using the function *ml.est* of the R-package SeleMix, the model parameters $\theta \equiv (\mathbf{B}, \mathbf{\Sigma}, w, \lambda)$ can be estimated via maximum likelihood estimation (MLE) through an ad hoc EM algorithm.

10. From a selective editing perspective it is natural using the above model to derive, via Bayes formula, the conditional distribution of the true value y^*_i given the observed value y_i for $i=1, .. n$. It results that this distribution is a mixture of a singular distribution $\delta(\cdot\,; y^*_i)$ with mass on the observed value y_i and a Gaussian distribution N_i^{err} with suitable parameters corresponding to the presence of error on y_i. Expectations from the conditional distribution of true data given observed data can serve as predictions (*anticipated values*) to be used in the selective editing procedure.
It can be easily verified that for each unit (and each analysed variable), the anticipated value is a weighted mean of the observed value y_i and the expectation from N_i^{err}, the latter being in turn a weighted mean of the observed value and the *unconditional* expectation of the true value.
Once predictions are available for all the observations, they can be used to define a *score function* in terms of discrepancies between predictions and observed values (expected errors).
Once all the observations have been ordered according to this score function, we are able to estimate the residual error remaining in data after the correction of the first k units (k=1, .., n), if the target population quantity is a linear function of the observed values, for instance a total or a (possibly weighted) mean. In fact, one can chose the first k^* units to be interactively revised, such that the estimate of the residual error is below a prefixed threshold so that the efforts spent in interactive editing (number of units to be revised) is directly related to the accuracy requirements.

B. Applying Selective Editing to agricultural price indices

11. As mentioned in the previous paragraph, the selective editing procedure implemented in SeleMix is appropriate in case of target estimates, such as means or totals, that are linear functions of the data. In this case in fact, it is possible to interpret the score function in terms of contribution to the total measurement component of the estimation error. In order to apply this methodology to composite indices, each index should be expressed as a linear combination of elementary data.

The computation of agricultural price indices was described in section II. The relationship among the national, regional and elementary indices was indicated. Grace to equation (3), it is possible to apply the selective editing procedure to the elementary indices. Indeed, the national monthly indices may be expressed as a linear combination of elementary monthly indices. Consequently, the national monthly indices may be naturally interpreted as the criteria measuring the contribution of each elementary monthly index.

IV. Selective editing implementation strategies and results

12. In this section we discuss the implementation strategies that were tested for the automatic application of the SeleMix methodology to national agricultural price indices. A step-by-step approach was considered. The implemented strategies all consists of the definition of the same stages. First, the definition and estimation of the contamination model parameters $\theta \equiv (\mathbf{B}, \mathbf{\Sigma}, w, \lambda)$ is necessary. Secondly, based on the obtained predictions, the score function is computed and consequently, the selection of the units with the highest impact is performed. Finally, some correction procedure is implemented. To evaluate the entire procedure, two criteria are considered; the first one is based on the convergence of the algorithm while the second one is an accuracy-based comparison with the manual procedure. Additionally, a comparison with the manual procedure was performed: the comparison concerns the number of units identified as erroneous by the automatic and manual procedures.

13. The three strategies compared in this project are mainly identified by the way the model parameters $\theta \equiv (\mathbf{B}, \mathbf{\Sigma}, w, \lambda)$ are estimated, i,e, the way the anticipated values are estimated. Indeed, three approaches were implemented, based on the choice of the $\mathbf{X}$ variable in the contamination model. The first approach, called M1, compares the elementary price indices of month *t* with the corresponding indices of month *t-12*,i.e.

$$_{prov,v}\mathbf{I}^{t}_{r,a} = \mathbf{B}\,_{prov,v}\mathbf{I}^{t-12}_{r,a} + \boldsymbol{\varepsilon} \qquad (4)$$

where the notation introduced in previous sections holds and the bold symbols represent the vectors of observations. The second strategy, called M2, consists in comparing the elementary price indices of month *t* with the corresponding indices of month *t-1*, i.e.

$$_{prov,v}\mathbf{I}^{t}_{r,a} = \mathbf{B}\,_{prov,v}\mathbf{I}^{p(t-1)}_{r,a} + \boldsymbol{\varepsilon} \qquad (5)$$

It should be noted that, in case of method M2, the month denoted by *p*(*t-1)* is not necessarily the previous month, but it represents the latest month in which the elementary prices were registered. This observation is particularly important in case of seasonal products which cannot be monitored in each month.

In both cases, M1 and M2, the anticipated values are represented by the fitted values $_{prov,v}\hat{\mathbf{I}}^{t}_{r,a}$ corresponding to models (4) and (5), respectively.

Finally, for the third strategy, called M3, no contamination model is used. The anticipated values are estimated by means of the relative changes with respect to the previous year. In this case, assuming that the monthly relative change *R* with respect to the previous month is constant over the consecutive years, the anticipated values may be written as $_{prov,v}\hat{\mathbf{I}}^{t}_{r,a} = (_{prov,v}\hat{\mathbf{R}}^{t}_{r,a} + 1) * _{prov,v}\mathbf{I}^{t-1}_{r,a}$.

The relative change *R* may be estimated using the monthly relative changes of the previous year, i.e. $_{prov,v}\hat{\mathbf{R}}^{t}_{r,a} = (_{prov,v}\mathbf{I}^{t-12}_{r,a} - _{prov,v}\mathbf{I}^{t-13}_{r,a}) \big/ _{prov,v}\mathbf{I}^{t-12}_{r,a}$.

14. In this work, based on some preliminary simulations, the procedure for the definition of the estimation domains together with the treatment of missing values were equal in each of the three implemented strategies. The choice of the estimation domains was guided by the trade-off between the number of available real-valued observations and the convergence of the maximum likelihood algorithm. In these preliminary simulations, the regional weights were not used, hence it was not really possible to take advantage of the linear relationship between national and

elementary monthly indices. The possible estimation domains are given groups of agricultural products. For method M1, the MLE algorithm converges in more than 90% of cases, independently on the chosen level of classification. When the estimation domain is the most detailed level of the classification of agricultural products, the method M2 converges in 34% and 75% for agricultural products and means of production, respectively. In order to reach convergences in more than 90% of cases, the method M2 has to be applied using synt_3 for agricultural products and synt_4 for the means of production. To enable the comparisons between methods, the three strategies were applied for the same estimation domains, i.e. synt_3 for agricultural products and synt_4 for means of production. Moreover, in the estimation step, the regional weights were used. Only the results obtained for these estimation domains are reported later in this section.

15. When applying the MLE algorithm implemented in the R package SeleMix, one problem to be faced is the treatment of missing values. As already described in section II, the elementary price indices are characterized by large percentages of missing values, non uniformly distributed among the products. In order to reduce the impact of the missing values, based on the real-valued observations, we firstly computed the probability p_{eq} of an observation being equal to the corresponding comparison value. In case of method M1, this comparison value is defined by the value observed at time *t-12*, while for method M2, it equals the value observed at time *t-1*, i.e. the previously registered observation. Using this probability p_{eq} ,we randomly imputed some missing values with the corresponding comparison values, maintaining constant the probability p_{eq} across the vector of observations.

16. In the next stage, using the anticipated values, the selective editing procedure was applied using different thresholds. Indeed, in this step, the observations whose corresponding errors exceed a given threshold are considered influential values. The thresholds *t.sel* used in this simulation are 0.01, 0.02 and 0.05.

17. In this stage it is possible to compare the SeleMix strategies with the manual editing procedure. In Table 3, the number of influential values for each implemented strategy and for the manual procedure is shown. The columns called "#infl" indicate the number of influential values identified by the SeleMix methodology. With respect to the number of influential units identified by the automatic procedure, the columns called "%Conc" indicate the percentage of observations simultaneously identified as influential by both the automatic and manual procedure. As expected, as the parameter *t.sel* increases, the number of influential units decreases. By comparing the three methods, it is possible to observe that only the usage of method M2 it is possible to approach percentages greater than 90%. Anyway, it should be noted that the agreement with the manual procedure does not represent one of the objectives of our project. The main goal of the Istat project is to reduce the number of observations to be edited, given some predefined accuracy results. Consequently, the number of influential values has to be compared with the overall number of manually edited observations, information shown in the last column of Table 3. By adopting the criteria of the minimum number of influential units, it is possible to consider method M1 slightly better than method M3. Anyway, method M2 is the one minimizing the number of influential values for means of production and non-seasonal agricultural products. On the contrary, method M1 performs better than method M2 for seasonal agricultural products. A possible explanation may be given by the problems encountered when estimating the models (5) in presence of high percentages of missing and duplicated values.

Means of production							
	***t.sel*=0.01**		***t.sel*=0.02**		***t.sel*=0.05**		
	#infl	**%Conc**	**#infl**	**#Conc**	**#infl**	**#Conc**	**Manual**
M1	1253	5.59	684	6.43	267	8.61	175
M2	176	46	86	60.5	26	80.8	175
M3	1635	6.61	696	9.48	142	19	175

Seasonal agricultural products							
	***t.sel*=0.01**		***t.sel*=0.02**		***t.sel*=0.05**		
	#infl	**%Conc**	**#infl**	**#Conc**	**#infl**	**#Conc**	**Manual**
M1	433	43.19	238	47.05	81	54.32	959
M2	486	58.4	300	68.3	153	81.7	959
M3	1399	40.61	911	45	383	50.91	959

Non-seasonal agricultural products							
	***t.sel*=0.01**		***t.sel*=0.02**		***t.sel*=0.05**		
	#infl	**%Conc**	**#infl**	**#Conc**	**#infl**	**#Conc**	**Manual**
M1	1123	16.18	643	20.06	259	30.5	771
M2	287	79.4	187	85	81	92.6	771
M3	2820	16.17	1545	22.14	525	34.1	771

Table2. Number of influential values identified by M1, M2 and M3.

17. In the final stage, in order to compare the manual procedure with the SeleMix methodology, two correction methods were applied, involving both missing and influential values. In the first one, called *Imp1*, the missing values were imputed by the anticipated values $\hat{}_{prov,v}\mathbf{I}_{r,a}^{t}$, while the influential values were corrected as if the manual procedure were in use. In the second one, called *Imp2*, both missing and influential values were imputed by the manually corrected values. In both cases, *Imp1* and *Imp2*, a cost reduction is anyway obtained if the number of influential units is reduced with respect to the manual procedure.

In Table 3, a comparison with the published indices is shown. The first three columns indicate the editing methodology, the used threshold and the simulated imputation/correction method. The other columns indicate the percentiles of the differences between the indices computed using the manual procedure and the automatic one. In general, it may be observed that there are no significant differences among the methods and the manual procedure. Thus, the assumptions of models (3) and (4) may be considered validated. Moreover, M2 and M3 seem to perform slightly better than M1. The extreme values in Table3 are due to the presence of products with high percentage of missing values.

Means of production

M	*t.sel*	Imp	p1	p5	p25	p50	p75	p90	p95	p99
M1	0.01	*Imp 1*	-4.82	-1.18	-0.14	0.00	0.12	0.59	1.15	2.72
		Imp 2	-0.93	-0.36	0.00	0.00	0.04	0.38	0.83	2.09
	0.02	*Imp 1*	-4.82	-1.23	-0.14	0.00	0.13	0.60	1.16	2.72
		Imp 2	-15.00	-0.38	0.00	0.00	0.04	0.40	0.84	2.09
	0.05	*Imp 1*	-4.89	-1.27	-0.14	0.00	0.13	0.60	1.16	2.81
		Imp 2	-1.48	-0.40	0.00	0.00	0.04	0.40	0.85	2.10
M2	0.01	*Imp 1*	-0.70	-0.21	0.00	0.00	0.00	0.14	0.41	1.42
		Imp 2	-0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.18
	0.02	*Imp 1*	-0.79	-0.22	0.00	0.00	0.00	0.15	0.42	1.48
		Imp 2	-0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.32
	0.05	*Imp 1*	-1.02	-0.23	0.00	0.00	0.01	0.15	0.48	1.77
		Imp 2	-0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.54
M3	0.01	*Imp 1*	-0.99	-0.21	0.00	0.00	0.02	0.23	0.55	2.38
		Imp 2	-0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.06
	0.02	*Imp 1*	-1.06	-0.22	0.00	0.00	0.02	0.24	0.60	2.55
		Imp 2	-0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.39
	0.05	*Imp 1*	-1.25	-0.23	0.00	0.00	0.02	0.25	0.62	2.99
		Imp 2	-0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.51

Non seasonal agricultural products

M	*t.sel*	Imp	p1	p5	p25	p50	p75	p90	p95	p99
M1	0.01	*Imp 1*	-34.02	-15.08	-0.70	0.00	1.03	4.56	8.89	18.11
		Imp 2	-5.06	-0.97	0.00	0.00	0.04	0.49	2.04	5.59
	0.02	*Imp 1*	-34.02	-15.08	-0.70	0.00	1.03	4.56	8.89	18.11
		Imp 2	-5.72	-0.96	0.00	0.00	0.03	0.48	2.04	5.59
	0.05	*Imp 1*	-34.02	-15.08	-0.79	0.00	1.08	4.90	8.94	18.11
		Imp 2	-5.75	-1.34	0.00	0.00	0.04	0.65	3.17	6.39
M2	0.01	*Imp 1*	-11.44	-2.83	-0.12	0.00	0.37	3.00	5.96	20.36
		Imp 2	-0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.13
	0.02	*Imp 1*	-11.44	-2.93	-0.12	0.00	0.39	3.00	5.96	20.36
		Imp 2	-0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.34
	0.05	*Imp 1*	-11.44	-3.93	-0.14	0.00	0.39	3.16	6.14	20.36
		Imp 2	-4.11	-0.07	0.00	0.00	0.00	0.00	0.00	1.68
M3	0.01	*Imp 1*	-11.84	-3.14	-0.12	0.00	0.41	2.88	8.12	17.84
		Imp 2	-0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.11
	0.02	*Imp 1*	-11.84	-3.14	-0.12	0.00	0.41	2.88	8.12	17.84
		Imp 2	-0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.12
	0.05	*Imp 1*	-11.84	-3.95	-0.18	0.00	0.42	2.98	8.12	17.84
		Imp 2	-4.04	0.00	0.00	0.00	0.00	0.00	0.00	0.32

Seasonal agricultural products

M	*t.sel*	Imp	p1	p5	p25	p50	p75	p90	p95	p99
M1	0.01	*Imp 1*	-15.28	-10.38	-2.73	0.00	3.05	6.36	8.41	14.09
		Imp 2	-5.26	-1.38	0.00	0.00	0.78	2.10	3.98	5.87
	0.02	*Imp 1*	-16.04	-11.44	-2.84	-0.04	3.05	6.37	8.41	14.09
		Imp 2	-9.25	-1.93	0.00	0.00	0.78	2.13	4.01	5.87
	0.05	*Imp 1*	-16.04	-11.32	-2.85	-0.07	3.31	6.36	8.41	14.09
		Imp 2	-9.25	-2.03	0.00	0.00	0.78	2.46	4.18	6.02
M2	0.01	*Imp 1*	-12.19	-5.85	-1.30	0.00	1.39	3.95	6.56	15.36
		Imp 2	-2.48	0.00	0.00	0.00	0.00	0.00	0.00	0.77
	0.02	*Imp 1*	-12.19	-5.85	-1.37	0.00	1.39	3.95	6.56	15.36
		Imp 2	-4.01	0.00	0.00	0.00	0.00	0.00	0.00	1.23
	0.05	*Imp 1*	-12.19	-6.28	-1.40	0.00	1.46	4.06	6.75	15.36
		Imp 2	-6.70	-0.70	0.00	0.00	0.00	0.00	0.00	4.32
M3	0.01	*Imp 1*	-15.24	-7.87	-1.31	0.61	4.58	9.63	15.25	28.22
		Imp 2	-0.58	0.00	0.00	0.00	0.00	0.00	0.00	0.45
	0.02	*Imp 1*	-15.24	-7.87	-1.31	0.61	4.64	9.63	15.25	28.22
		Imp 2	-0.99	0.00	0.00	0.00	0.00	0.00	0.00	1.62
	0.05	*Imp 1*	-15.24	-7.87	-1.54	0.63	4.64	9.63	15.63	28.22
		Imp 2	-2.94	-0.09	0.00	0.00	0.00	0.00	0.01	2.01

IV. Conclusions

This analysis is part of a group of methods that include models currently used by the EU for the analysis of agricultural markets.

In the context of price statistics, the agricultural products is a sector very peculiar especially for the nature of phenomena detected and the manner in which they are observed.

According the analysis of this paper can be seen as a contribution to evaluate by comparing products on evaluation of the imputation process related to the three different techniques.

In order to assess the method, for each threshold, the interactive editing has been simulated by replacing the selected values with the ones available from the official "cleaned data". Moreover, two different approaches have been adopted to impute missing values. The adopted methodology is a type of correction. In the table they are imputation1 and imputation2.

This report contains the distinction as it is possible to see in the table of purchased and sold goods and services the total number of each month that is less in correspondence of 0,05; 0,02 and 0,01 per cent.

In presence of missing value we have used two different options to impute these values, in the first case the price of missing values are equals to zero. In the second case the price of missing values are imputed to be equal to price used.

As we can see the biggest difference of the results between the imputation1 and the imputation2 about the method M3 is products purchased.

References

Di Zio, M., and Guarnera, U. (2013). A Contamination Model for Selective editing. *Journal of Official Statistics*, Vol. 29, No. 4, pp. 539-555.

Eurostat (2008). Handbook for EU Agricultural Price Statistics, Version 2.0, available at http://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/-/KS-BH-02-003.