istat
working
papers

N. 15
2011

Towards the streamlining of the editing process of Italian data on the external merchandise trade

Alessandra Nuccitelli

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N. 15
2011

Towards the streamlining of the editing process of Italian data on the external merchandise trade

Alessandra Nuccitelli

Istat

Istat Working Papers

Towards the streamlining of the editing process of Italian data on the external merchandise trade

N. 15/2011

ISBN 88-458-1701-6

Istituto nazionale di statistica
Servizio Editoria
Via Cesare Balbo, 16 – Roma

Towards the streamlining of the editing process of Italian data on the external merchandise trade

Alessandra Nuccitelli

Sommario

Lo scopo principale di questo lavoro è quello di evidenziare i vantaggi dei nuovi criteri automatici introdotti recentemente nel processo di controllo dei dati italiani sul commercio estero di beni. Considerando la gran quantità di osservazioni da trattare in un tempo relativamente breve e il livello molto dettagliato di diffusione, tali procedure svolgono un ruolo importante nel garantire la qualità dei dati finali. La suddivisione - per operatore commerciale o per paese partner - dei domini utilizzati per l'individuazione di dati anomali conduce a intervalli di accettazione più affidabili rispetto al passato. Inoltre, l'introduzione di una misura dell'errore potenziale associato ad ogni osservazione sospetta consente di concentrare la revisione interattiva solo sui dati anomali con maggior impatto sulle statistiche finali; ciò contribuisce a ridurre il numero dei controlli senza compromettere la qualità complessiva dei dati finali. La validità delle stime preliminari viene controllata tramite una tecnica di macro-editing basata sulla previsione automatica di serie storiche di dati opportunamente aggregati.

Parole chiave: analisi della varianza, editing selettivo, intervallo di accettazione, macro-editing, previsione di serie storiche.

Abstract

The main aim of this work is to highlight the advantages of the new automated criteria recently introduced in the editing process of the Italian data on the external merchandise trade. Considering the huge amount of records to be processed in a relatively short time and the very detailed level of dissemination, the editing procedures play a major role in ensuring the quality of the final data. The breakdown - by trader or partner country - of the domains used for the outlier detection turns out to give more reliable acceptance intervals than in the past. Besides, the introduction of a measure of potential error related to each suspicious observation enables to concentrate the interactive review only on the most influential outliers; this helps to reduce the number of checks without compromising the overall quality of the final figures. The validity of the preliminary estimates is checked through a macro-editing technique based on the automatic forecasting of time series of properly aggregated data.

Keywords: acceptance interval, analysis of variance, macro-editing, selective editing, time series forecasting.

1. Introduction*

In Italy, the statistics on the merchandise trade with countries outside the European Union (EU) are compiled from documents provided by importers or exporters (or trade operators) to the Customs authorities.

The relevant document is the Single Administrative Document (SAD). The information reported on the SAD is collected for both fiscal and statistical purposes; data items include product description, value,[1] net mass (in kilograms), country of origin and/or destination and trade operator. At present, electronic declarations account for about 95 per cent of the total trade value (they are obligatory for exports since July 2009). Collected data are transferred monthly by the Italian Customs Agency to Istat, the National Statistical Institute, for the subsequent processing and dissemination. Istat receives approximately 2.5 million records per month, which correspond to about 800,000 transactions.

Italian statistics on the external merchandise trade are disseminated every month by flow (import or export), eight-digit commodity code (according to the CN)[2] and country (from which Italy imports or to which Italy exports), or at higher aggregation levels.

Considering the huge amount of records to be processed in a relatively short time and the very detailed level of dissemination, editing procedures play a major role in ensuring the quality of the final data.

The selection of observations (*i.e.* transactions) with potential errors is based primarily on the use of acceptance intervals for:

- the AP (Average Price), *i.e.* the ratio of value to net mass;
- the UW (Unit Weight), *i.e.* the ratio of net mass to number of SUs (Supplementary Units, *e.g.*: litres, number of parts, square metres), if any.

If the AP or the UW for a particular transaction falls outside the respective interval, then that observation is considered as suspicious and should be further analysed by the clerical staff. Clerks may also call back customs offices or traders in an attempt to verify or correct suspicious data. Henceforth, lower and upper bounds of acceptance intervals will be called edit parameters as well.

In the past, the AP acceptance intervals were updated and maintained manually by subject matter experts. Those intervals were defined by flow and eight-digit commodity code. Due to time and resource constraints, maintaining effective AP edit parameters became increasingly challenging; besides, their updating proved to be not so timely in following price dynamics in international markets, causing too many false suspicious observations (from now on, these observations will be also called rejects or outliers). Moreover, since not all outliers could be reviewed, specific thresholds for revision were used. Those thresholds were defined on a judgmental basis by flow and chapter (two-digit commodity code according to the Harmonized System) for value, net mass and number of SUs, in order to concentrate the review only on the rejects with high value and/or high net mass and/or high number of SUs.

Recently, automated methods have been investigated to maintain and refresh the AP edit parameters, while increasing the effectiveness and the efficiency of the editing procedures.

As a first step towards the streamlining of the editing process, automated statistical criteria have been adopted to detect outliers from January 2007. According to such criteria, the AP edit parameters have been updated every month by using the distribution of the raw micro-data in the latest two years (Nuccitelli, 2007; Narilli and Nuccitelli, 2008). Afterwards - more precisely, from January 2009 - the AP parameters have been estimated for more specific strata (or domains) than in

---

* An early version of this paper was presented at the European Conference on Quality in Official Statistics, Helsinki, Finland, May 4-6, 2010.

1 In this paper, 'value' means the value of the goods (in euros) when crossing the national frontier, *i.e.* including the transport and insurance costs up to the national frontier. Further details on definitions, classifications and basic methodology used for the compilation of statistics on the trading of goods can be found in Eurostat (2006).

2 The CN, or Combined Nomenclature, is a European Community's classification of goods (based on the Harmonized System) that includes nearly 10,000 eight-digit commodity codes. The Harmonized System is used worldwide as a reference for classifications of foreign trade statistics and for customs tariffs. The CN is updated yearly by introduction, deletion or alteration of commodity codes and/or their descriptions.

the past; a finer stratification by trade operator or country has yielded more reliable acceptance intervals. At the same time, the pre-defined thresholds for revision have been replaced by a criterion that selects the outliers to be reviewed by considering their suspicion and potential error. This criterion has enabled to review rejects with low value (and/or low net mass and/or low number of SUs) as well, containing the introduction of bias in the final statistics.

Finally, the above selective micro-editing approach has been combined with a macro-editing procedure based on the fully automatic forecasting of monthly value or monthly net mass.

The main aim of this work is to highlight the advantages of the procedures and criteria lately introduced in the editing process of the Italian data on the external merchandise trade.

This paper is organized as follows. The methodology used to produce the AP edit parameters is discussed in section 2, with particular emphasis on the importance of using domains specific to trade operator or country. The criterion adopted for selecting the rejects to be reviewed is described in section 3. Some results about the introduction of these recent developments in the editing process are reported in section 4. The procedure for macro-editing is examined in more detail in section 5. Finally, some conclusions are drawn.

2. Automated production of edit parameters

The definition of ranges in which 'good' data are expected to fall within is based on the assumption that the majority of data are reported correctly. Under this assumption, using the bulk of import or export data for a commodity, it is possible to produce lower and upper bounds, such that any value lying outside these limits is unusual enough to be considered as corresponding to an outlier. Acceptance intervals can be defined by using quantiles derived from the distribution of properly stratified raw micro-data (Fescina *et al.*, 2004; Garcia *et al.* 2006).

On this point, because of the low quality of the raw micro-data concerning both net mass and number of SUs, estimating UW edit parameters through the above statistical criteria may be a tricky task. So, in the editing process under consideration, for some products the UW acceptance intervals are made to correspond to the ranges of acceptable values specified by the CN in the commodity description; in the other cases, the acceptance intervals are settled on a judgmental basis by subject matter experts.

With reference to APs, there are some important issues to keep in mind when discussing how edit parameters can be produced.

Firstly, within a particular domain on which quantiles can be computed, the more homogeneous the APs are, the more effective the outlier detection method is. Mostly, the lack of homogeneity within a domain identified by flow and commodity can be accounted for by factors such as trader or country. In these cases, the frequency distribution of the transactions by AP in such a domain can be seen as a mixture of sub-distributions with different centres. Besides, the more distant the centres are, the less accurate the acceptance interval is.

Secondly, a detailed grouping is in conflict with the demand for a minimum amount of observations needed in each domain to compute quantiles. Each commodity is involved in a different number of transactions every month; so, in order to have enough data to produce reasonable limits, a large file of historical data is needed.[3] Thus, in the current editing process, the observations in the latest 24 months - including the month under examination (from now on, reference month) - are exploited.

Strata identified by flow and eight-digit commodity code were considered until the end of 2008. Since January 2009, more specific domains have been used according to the priority order defined by the following levels of data stratification:

[3] On the other hand, using data even more distant in time may offer some drawbacks; for example, in case of price rise in the month under examination (or reference month), edit parameters might prove to be more easily unsuitable.

a) by flow, eight-digit commodity code, trade operator;
b) by flow, eight-digit commodity code, country; (1)
c) by flow, eight-digit commodity code.

More precisely, for a particular transaction to be checked, the trader-specific range takes priority over the others; if this interval is not available - as there are not enough data[4] in the corresponding *a*-type domain to estimate reasonable bounds - the country-specific range is used. The generic interval (on the *c*-type domain) is employed only in the last resort, when the limits for the *b*-type domain are not available either. In case no range is usable, that transaction is considered as suspicious and should be further analysed by the clerical staff.

The suitability of using AP acceptance intervals specific to trader or country is discussed in more detail in next subsection. The formulas for computing AP parameters on a domain are provided in subsection 2.2.

2.1 Edit parameters specific to trade operator or country

Some drawbacks rising from strong heterogeneity between APs within a *c*-type domain are explained through the following two cases.

The former concerns the imports of "*Textile hosepiping and similar textile tubing of synthetic fibres*" (Figure 1). In 2008 and 2009, this good came mainly from China, Croatia and the United States. The generic acceptance interval (2.79-17.71 euros/kg) is too narrow and unsuitable, especially for the imports from the United States (all related observations are flagged as outliers). In such a situation, the country-specific parameters appear more appropriate (1.66-2.98, 4.03-6.98 and 38.09-58.65 euros/kg, respectively for China, Croatia and the United States).

In the latter case, related to the exports of "*Vitamin B_6 and its derivatives*" (Figure 2), the generic range (5.34-853.61 euros/kg) is too wide because of the very big difference between the APs for two - *A* and *B*[5] - among the main traders (no transaction is flagged as suspicious). A great gain in the accuracy of the outlier detection can be achieved by using trader-specific parameters (11.79-18.33 and 177.97-277.00 euros/kg, respectively for *A* and *B*).

The breakdown by trade operator or country turns out to be particularly useful within domains corresponding to residual headings or sub-headings of the CN, which often cover goods with very different features.[6]

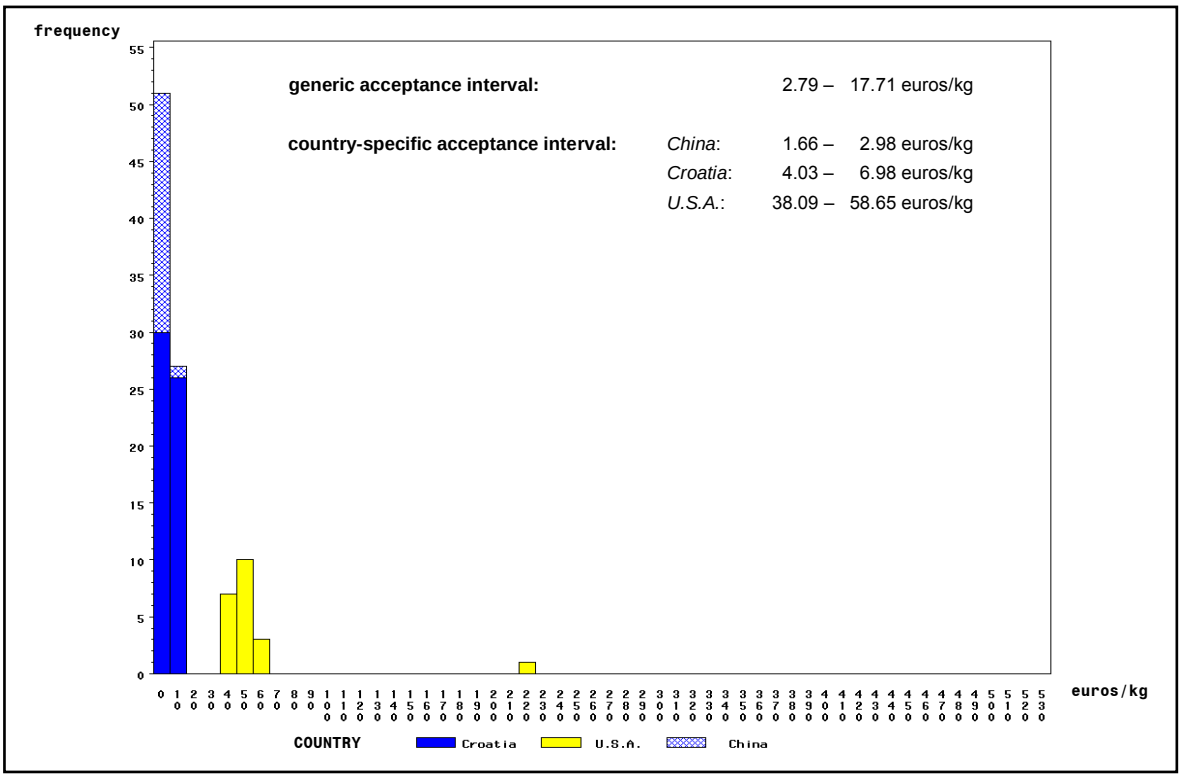
[4] Acceptance intervals are only computed for domains with at least 20 observations in the latest 24 months.

[5] For reasons of confidentiality, the VAT (Value-Added Tax) identification number of the involved traders can not be disclosed.

[6] Residual headings or sub-headings generally occur last in numerical order among those that equally merit consideration in classifying a particular good. For instance, with reference to chapter 14, the residual sub-headings corresponding to eight-digit commodity codes are 14019000 and 14049000:

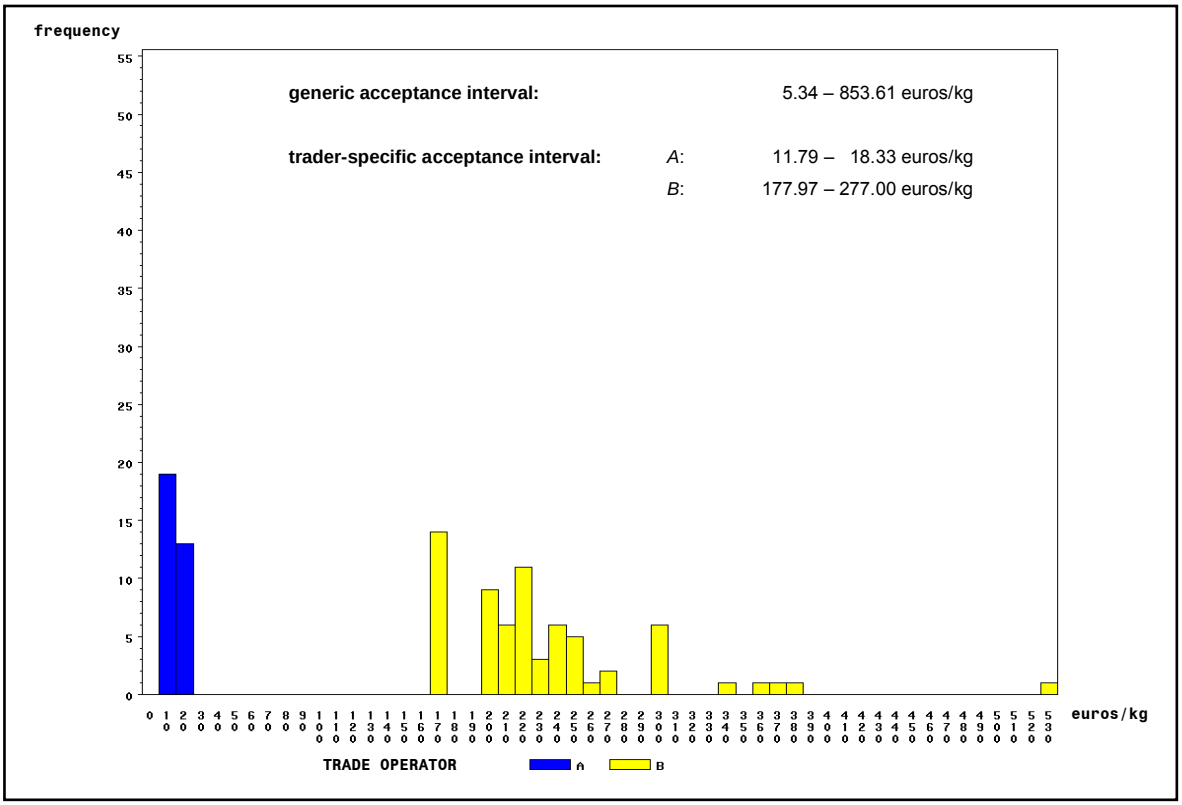
1401	*Vegetable materials of a kind used primarily for plaiting:*
14011000	*- Bamboos;*
14012000	*- Rattans;*
14019000	*- Other.*
1404	*Vegetable products not elsewhere specified or included:*
14042000	*- Cotton linters;*
14049000	*- Other.*

Figure 1 - Imports of Textile hosepiping and similar textile tubing of synthetic fibres by average price - Extrastat data *(January 2008-December 2009)*



Source: Elaboration based on Istat data

Figure 2 - Exports of Vitamine B6 and its derivatives by average price - Extrastat data *(January 2008-December 2009)*



Source: Elaboration based on Istat data

In order to evaluate the suitability of adopting specific domains, the technique known as analysis of variance (or ANOVA) can be used. In this framework, the aim is to search for more homogeneous strata within every domain identified by flow and eight-digit commodity code.

The natural logarithm of the AP is considered as response variable, and trade operator and country as explanatory variables (or factors).

As the number of observations in a domain represents a critical aspect, the analysis of variance is performed by considering one factor at a time (one-way ANOVA); besides, for each factor, only the *c*-type domains with a sufficient number of levels and observations by level are selected for the subsequent analysis.

The results presented here are obtained by selecting, for each factor, the *c*-type domains for which the number of levels with at least 20 observations is greater than 1.

Table 1 gives, by factor and flow, the number of commodities for which:

- the number of levels with at least 20 transactions is greater than 1 (column 2);
- the factor has a significant effect on the response variable, and at least 45 per cent of the variability in the response variable is due to the factor (column 3).

The commodities eligible for the analysis of variance turn out to be less than half of the total. In reading table 1, it is necessary to keep in mind that, when both country and trade operator have a significant effect on the response variable, only the factor accounting for the higher percentage of variability is considered.

As displayed in table 1, at least one factor has a significant effect on the response variable for over half of the commodities eligible for the analysis of variance. Besides, even though the number of these commodities is higher for the country factor, the trader is more often decisive in accounting for the variability in the response variable, especially for exports.

Table 1 - Commodities eligible for the ANOVA and commodities for which at least 45 per cent of the variability in the response variable is accounted for by one of the factors, by factor and flow - Extrastat data *(January 2008-December 2009)*

FACTOR	Commodities eligible for ANOVA	Commodities for which at least 45% of the variability in the response variable is accounted for by one of the factors
	IMPORT	
Country	4,184	532
Trade operator	3,686	2,101
	EXPORT	
Country	4,620	238
Trade operator	4,278	2,411

Source: Elaboration based on Istat data

2.2 Computing edit parameters on a domain

For a particular domain, AP edit parameters are computed as follows.

Let AP_i be the AP for the observation *i* in the reference month, *i.e.*:

$$AP_i = \frac{value_i}{net\_mass_i} \qquad net\_mass_i \neq 0;$$

besides, let $AP_{Q1}(i)$ and $AP_{Q3}(i)$ be, respectively, the first and third quartiles of the distribution of the raw APs related to the transactions occurred in the latest 24 months (including the reference month) in the same domain of *i*.

Since ratios like APs have by nature skewed distributions, data are symmetrised using the natural logarithm transformation.[7]

When $net\_mass_i \neq 0$, the suspicion for *i* can be defined as:

$$suspicion_i = \begin{cases} \dfrac{log(AP_{Q1}(i)) - log(AP_i)}{log(AP_{Q3}(i)) - log(AP_{Q1}(i))} & \text{if } AP_i < AP_{Q1}(i) \\ 0 & \text{if } AP_{Q1}(i) \leq AP_i \leq AP_{Q3}(i) \\ \dfrac{log(AP_i) - log(AP_{Q3}(i))}{log(AP_{Q3}(i)) - log(AP_{Q1}(i))} & \text{if } AP_i > AP_{Q3}(i). \end{cases} \quad (2)$$

Any observation whose suspicion is greater than a non-negative real number *C* is considered as suspicious. In other words, a transaction is considered as suspicious when the distance (in logarithmic terms) between the observed AP and the closer of the two quartiles is greater than *C* times the interquartile range.

The bounds $AP_{min}(i)$ and $AP_{max}(i)$ ($AP_{min}(i) \leq AP_{max}(i)$) can be derived easily as functions of *C*:[8]

$$\begin{aligned} AP_{min}(i) &= exp\big(-C \times log(AP_{Q3}(i)) + (1+C) \times log(AP_{Q1}(i))\big) \\ AP_{max}(i) &= exp\big(-C \times log(AP_{Q1}(i)) + (1+C) \times log(AP_{Q3}(i))\big). \end{aligned} \quad (3)$$

3. Selecting micro-data to be reviewed

The transactions a priori subjected to review fall within at most one of the following three categories:

1. observations related to commodities that require special attention (*e.g.*: petroleum, natural gas, electrical energy, diamonds, pearls, aircrafts, spacecrafts, ships, boats and works of art);
2. observations for which net mass is 0 or missing;
3. observations for which any acceptance interval can not be estimated.

Every observation that does not fall into one of the above categories can be given a score, in order to concentrate the review only on the suspicious micro-data most affecting the final statistics.

The proposed score function is adapted from the one developed by Jäder and Norberg (2005) for the Swedish foreign trade data and includes measures of suspicion and potential error (in terms of value).

[7] This transformation has also the advantage of eliminating the problem of negative lower bounds, which are not meaningful in a foreign trade context.

[8] In choosing a value for *C*, the primary concern is to reduce the incidence of too narrow acceptance intervals - such that there is an unusually large number of rejects - and too wide intervals, where all observations pass through. The trade-off, however, is that, as *C* increases, the number of domains without rejects increases.
On the basis of experimentations, the value 0.5 has been considered as 'optimal' for *C* and it has been used up to now to calculate the AP parameters (further details can be found in Nuccitelli, 2007).

The potential error for *i* can be defined as:

$$error_i = \left| value_i - net\_mass_i \times AP_{Q2}(i) \right|, \tag{4}$$

where $AP_{Q2}(i)$ is the median of the distribution of the raw APs related to the transactions occurred in the latest 24 months in the same domain - as specific as possible, according to the priority order proposed in (1)[9] - to which *i* belongs.

In other words, the median AP, multiplied by the observed net mass, is used as the best expected value; the potential error is defined as the absolute difference between the observed value and the estimated one. It is worth noting that an error either in the observed value or in the observed net mass results in a potential error measured in value.

The observation *i* is selected for review, only if the following conditions are satisfied simultaneously:

$$\begin{aligned} &suspicion_i > C; \\ &error_i > D; \\ &(suspicion_i - C) \times (error_i - D) > E, \end{aligned} \tag{5}$$

where *D* and *E* are constants (measured in euros).

Representing each observation by its suspicion and potential error in a two-dimensional Cartesian coordinate system, the boundary of the acceptance region for data revision, according to conditions (5), is the branch of the hyperbola described by the equation $(suspicion - C) \times (error - D) = E$ in the first quadrant.[10]

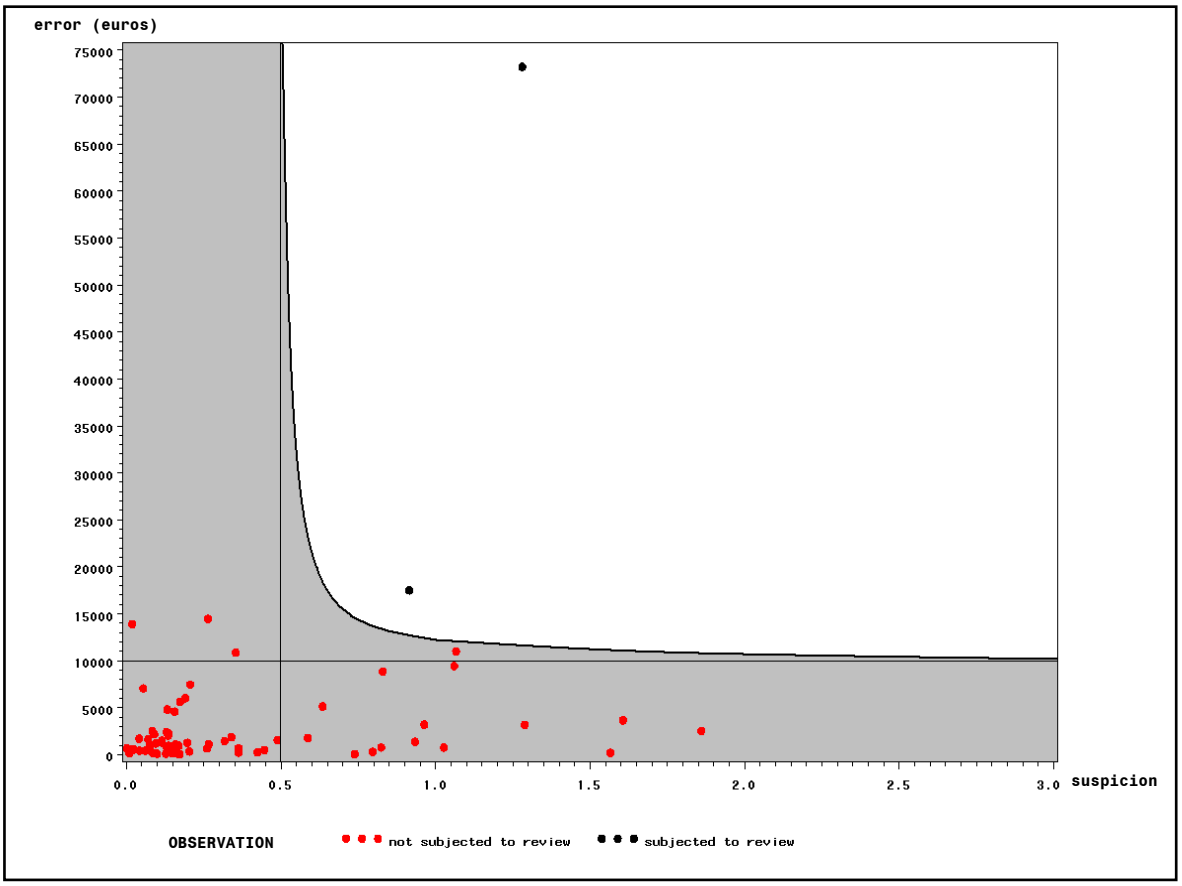
The number of observations to be reviewed is an important measure of the clerical workload. As available resources and time represent unavoidable constraints, this number can be used to calibrate *D* and *E*.

By way of an example, with reference to the imports of goods related to chapter 1 ("*Live animals*") in December 2009, the observations selected for review are represented by the points located above the branch of the hyperbola identified by $C = 0.5$, $D = 10{,}000$ euros and $E = 1{,}000$ euros (Figure 3).

[9] This precaution makes $AP_{Q2}(i)$ a good predictor of the AP for the observation *i*.

[10] The asymptotes of the hyperbola - $suspicion = C$ and $error = D$ - are parallel to the coordinate axes.

Figure 3 - Imports related to *Live animals* in December 2009 by potential error (in euros) and suspicion - Extrastat data *(January 2008-December 2009)*



Source: Elaboration based on Istat data

4. Some results

Measuring the effectiveness of the editing criteria lately introduced is a difficult task. Since reporting mistakes in some final transactions may still look reasonable, even if those data truly are incorrect, there is no way of accurately identifying a truly 'good' or truly 'bad' observation.

The effectiveness of the new micro-editing criteria can not even be assessed by using validated data from the past, as conclusions would depend heavily on the old editing choices.

Besides, the huge amount of monthly transactions - about 800,000 - together with the small number of clerks in charge of the review have not allowed to test in parallel the new editing criteria [edit parameters computed according to (3) for more specific strata and outliers selected for review according to (5) with $C = 0.5$, $D = 10{,}000$ euros and $E = 1{,}000$ euros] versus the former ones [edit parameters computed according to (3) on generic domains and outliers selected for review by using judgmental thresholds].[11]

Some results in terms of number of observations selected for review and potential error are reported respectively in tables 2 and 3, for the month of December 2009.

[11] With reference to domains identified by flow and eight-digit commodity code, a comparative assessment of the edit parameters produced according to (3) versus the old ones, settled on a judgmental basis, is reported in Narilli and Nuccitelli (2008). In that paper, a comparison between the two types of acceptance interval by their length and mutual position shows the old bounds to be often underestimated or not updated, especially for exports.

As displayed in table 2, acceptance intervals can not be produced only for a very small percentage of transactions (less than 0.3 per cent of the total). More than half of the observations selected for review according to the new editing criteria are flagged as suspicious by trader-specific intervals and about one third by country-specific intervals. Only a small percentage (8 and 14 per cent for imports and exports, respectively) is flagged as suspicious by using generic ranges.

Table 2 - Observations selected for review in December 2009 by type of acceptance interval and flow - Extrastat data *(January 2008-December 2009)*

OBSERVATIONS TO BE REVIEWED (*)	Type of acceptance interval			
	Trader-specific	Country-specific	Generic	**Total**
	IMPORT			
Obs. for which any acceptance interval can not be estimated	-	-	-	**571**
Obs. selected for review according to the old criteria	-	-	11,250	**11,250**
Obs. selected for review according to the new criteria	6,238	4,734	977	**11,949**
	EXPORT			
Obs. for which any acceptance interval can not be estimated	-	-	-	**608**
Obs. selected for review according to the old criteria	-	-	26,954	**26,954**
Obs. selected for review according to the new criteria	13,588	8,579	3,477	**25,644**

Source: Elaboration based on Istat data

(*) For the sake of brevity, the observations falling into categories 1 and 2 (see section 3) are not considered.

With almost the same number of observations selected for review than in the past, the new editing choices enable to check in-depth data more affected by potential error (Table 3).

Table 3 - Frequency distributions of total observations and observations selected for review in December 2009 according to the old and new editing criteria, by potential error category and flow - Extrastat data *(January 2008-December 2009)*

POTENTIAL ERROR (IN EUROS)	Total observations (*)		Obs. selected for review according to old and new editing criteria			
			Old editing criteria		New editing criteria	
	absolute frequencies	%	Absolute frequencies	%	Absolute frequencies	%
			IMPORT			
≤ 10,000	168,303	81.14	6,681	59.39	0	0.00
10,000 –| 100,000	35,071	16.91	3,658	32.52	10,015	83.81
100,000 –| 1,000,000	3,683	1.78	798	7.09	1,743	14.59
1,000,000 –| 10,000,000	367	0.18	108	0.96	185	1.55
> 10,000,000	9	0.00	5	0.04	6	0.05
Total	**207,433**	**100.00**	**11,250**	**100.00**	**11,949**	**100.00**
			EXPORT			
≤ 10,000	347,269	81.99	17,656	65.50	0	0.00
10,000 –| 100,000	66,004	15.58	6,691	24.82	20,266	79.03
100,000 –| 1,000,000	9,294	2.19	2,152	7.98	4,696	18.31
1,000,000 –| 10,000,000	936	0.22	415	1.54	636	2.48
> 10,000,000	62	0.01	40	0.15	46	0.18
Total	**423,565**	**100.00**	**26,954**	**100.00**	**25,644**	**100.00**

Source: Elaboration based on Istat data

(*) Total observations are net of the transactions related to categories 1, 2 and 3 (see section 3).

5. Macro-editing

A macro-editing procedure is combined with the above micro-editing approach. Whereas micro-editing basically treats each record to be edited separately, macro-editing treats the data set to be edited as a whole and can lead to the detection of errors that could go unnoticed with the other approach. Besides, human operators themselves might introduce recording errors (*e.g.* the inclusion or omission of one or more digits) in correcting the micro-data selected for review.

The macro-editing is performed at the end of the checking and correction process, in order to assess the validity of the preliminary estimates and identify possible residual influential errors in the data. Initial estimates of monthly value (stratified by flow and trader, or by flow, eight-digit commodity code and country) or monthly net mass (stratified by flow, eight-digit commodity code and country) are compared with the corresponding 'predictions' or forecasts for the reference month.

The forecasts are based on models applied to validated data from the past. It is not necessary that these predictions are completely accurate; they are only used to select the strata for which the preliminary estimates of value or net mass are atypical, in order to prioritize the editing efforts.

In particular, the one-month ahead forecasts and confidence limits for value or net mass are generated by the stepwise autoregressive method implemented in the FORECAST procedure available in SAS software. That procedure provides a quick and automatic way to produce forecasts, parameter estimates and goodness-of-fit statistics for many time series in one step. For each series, the forecasts are functions of time and past values of the series, not of other variables.[12]

In this macro-editing context, the forecasting procedure is applied to monthly data validated in the latest five years[13] and stratified by flow and trader or by flow, commodity code and country.[14]

It is necessary to keep in mind that the procedure is just applied to those strata for which the aggregates can be properly compared over the period under examination (the latest five years). Therefore, in order to have time series long enough, only the strata identified by lively traders or by commodities and countries not affected by any change over the whole period are actually considered. Overall, the forecasts are produced for a very large number of series at a time (nearly one million).

Then, suspicious strata are selected to find in them transactions with potential residual errors.

A stratum is considered as suspicious when the following conditions are satisfied simultaneously:

- the absolute difference between the preliminary estimate of monthly value (or monthly net mass) and the forecast exceeds a pre-selected threshold in euros (or in kilograms);
- the absolute difference between the preliminary estimate of monthly value (or monthly net mass) and the closer of the two confidence limits is greater than F times the confidence range, where F is a pre-selected non-negative real number.

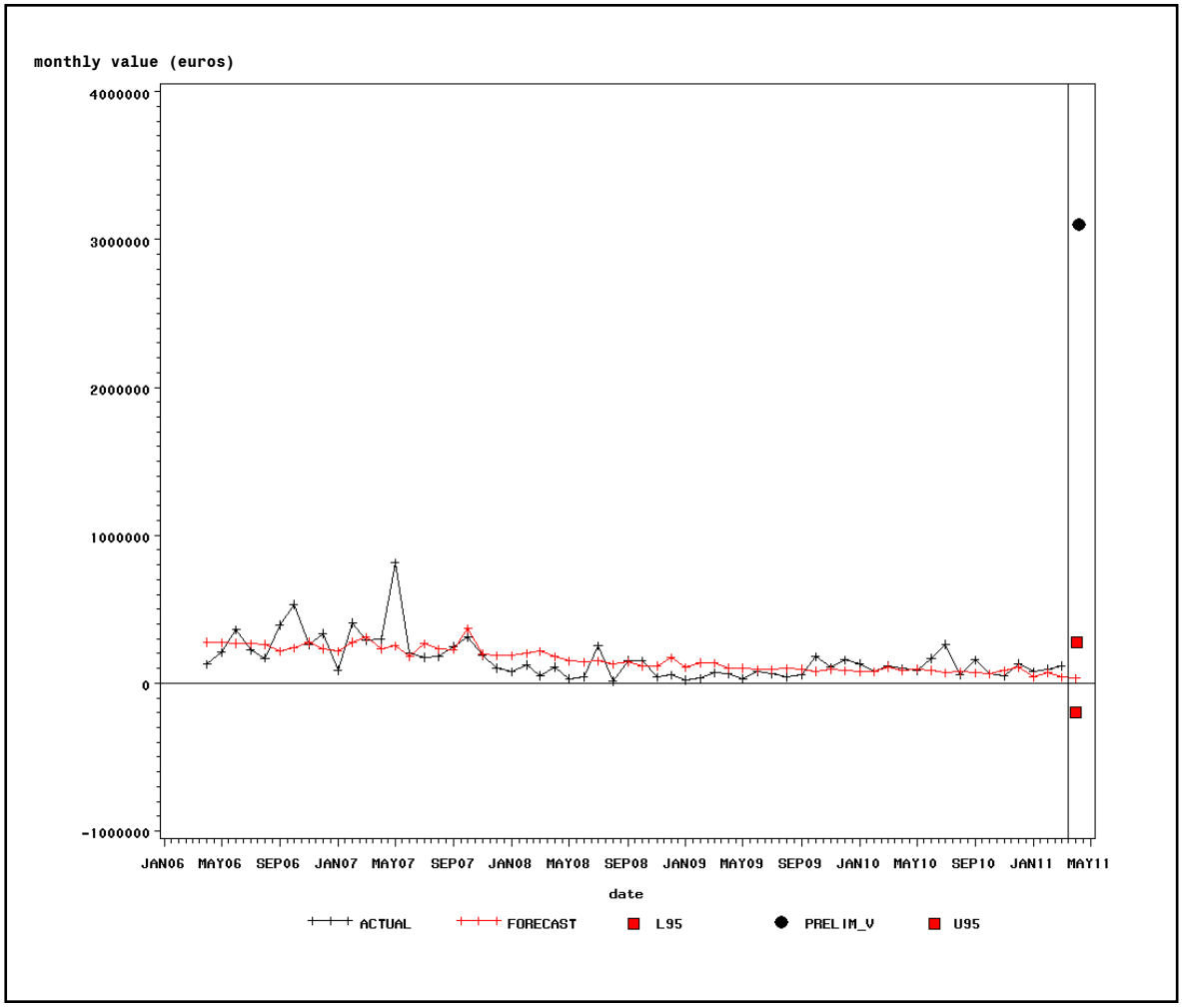
[12] The stepwise autoregressive method combines a time-trend regression with an autoregressive model for departures from trend. The procedure first fits a time trend model (constant, linear or quadratic) to the series, using ordinary least-squares regression. Then, the residuals from the estimated trend are taken and an autoregressive model is fit to them.
The autoregressive parameters included in the model for each series are selected by a stepwise regression technique, so that autoregressive parameters are only included at those lags at which they are statistically significant (in this specific context, the maximum order considered for the autoregressive model is twelve, to account for possible seasonal fluctuations of each time series about the trend). Since the trend and autoregressive parameters are fit in sequence rather than simultaneously, the parameter estimates are not optimal in a statistical sense; however, the estimates are usually close to optimal, and the method is computationally inexpensive. Further details can be found in SAS Institute Inc. (1993).

[13] Longer series do not necessarily yield better forecasts. If a time series is too long, there is a danger that the early part of the data is not relevant to making forecasts from the end of the series, perhaps because of changes in the underlying model or in the data themselves (*e.g.* changes in the classification of goods or in the register of trade operators).

[14] A less detailed stratification can be inappropriate, since possible outliers could go undetected.

By way of an example, figure 4 shows a time plot of the monthly values of the goods exported by a trader[15] from April 2006 to April 2011. The time series exhibits a slight downward trend up to March 2011; this is followed by an upward spike in April 2011, which represented a preliminary estimate for the reference month. The stratum at issue was considered as suspicious, since the distance between the preliminary estimate and the upper confidence limit was about six times the whole confidence range for the predicted value. A close examination of the involved micro-data revealed that the unexpected increase in value for the reference month was due to an error in the currency declared for a single transaction (Japanese yen were used instead of euros).

Figure 4 - Validated monthly values (in euros) of goods exported by a trader from April 2006 to March 2011; preliminary (wrong) estimate of monthly value and one-month ahead forecast with confidence limits (in euros) for April 2011 - Extrastat data *(April 2006-April 2011)*



Source: Elaboration based on Istat data

[15] For reasons of confidentiality, the VAT (Value-Added Tax) identification number of the involved trader can not be disclosed.

6. Concluding remarks

With reference to APs, the methodology for editing data on the merchandise trade with countries outside the EU has been also adopted - with suitable changes - in the context of the Intrastat[16] data processing from January 2010.

In comparison with the past, the new editing criteria ensure a higher degree of accuracy and objectiveness in detecting suspicious observations and a reduction of time and costs related to the parameter updating. In particular, acceptance intervals specific to trade operator or country turn out to be more suitable than the generic ones, defined by flow and commodity code. The selective micro-editing criterion, which combines both suspicion and potential error, enables to concentrate time and energy on the most influential outliers, limiting the re-contacts with customs offices or traders as well. Besides, it allows to calibrate easily the number of transactions to be reviewed as a function of the available resources and to check in-depth rejects even with low value (and/or low net mass and/or low number of SUs), containing the introduction of bias in the final statistics.

As regards the drawbacks, the method to determine acceptance intervals may be useless for detecting outliers, when the quality of the input data is poor. Since systematic mistakes may affect the net mass or number of SUs, the UW edit parameters can not be produced automatically by using the proposed statistical approach.

The micro-editing phase is followed by macro-editing, in order to assess the validity of the preliminary estimates and identify possible residual influential errors. The macro-editing procedure is based on the fully automatic forecasting of monthly value or monthly net mass stratified at a very detailed level (by flow and trader, or by flow, commodity code and country). Therefore, the forecasts are produced every month for large numbers of time series. As it is quite impractical to develop a model tailored to each series, a simple, automatic forecasting method is used. Moreover, the selective macro-editing criterion allows to calibrate the number of strata (and observations within them) to be examined as a function of the available resources.

Other kinds of checks can be introduced in the editing process under consideration. In such a framework, a further significant increase in the efficiency and effectiveness of the editing procedures requires a broader approach, involving organizational aspects as well. Systematic mistakes could be prevented by anticipating editing at the data collection phase, *e.g.* through validity checks on the values to be entered when filling in electronic forms.

[16] The Intrastat survey covers the trading of goods among countries within the EU.

Appendix - Determining edit parameters allowing for annual changes in the Combined Nomenclature

Changes are incorporated into the CN at the beginning of each year, either at the request of trade federations or national and Community authorities, or for legal reasons.

The CN is updated by introduction, deletion or alteration of commodity codes and/or their descriptions.

In this section some directions on how to produce AP edit parameters according to the proposed approach, allowing for annual changes in the CN, are provided.

In the year $t+1$ the AP edit parameters for new commodity codes can be estimated as follows:

- if one or more new codes correspond to one old code, the edit parameters for each new code can be estimated by using both the observations related to this new code (in the year $t+1$) and the observations related to the old code (in the year *t*);
- if a new code corresponds to several old codes, the edit parameters for the new code can be estimated by using both the observations related to it (in the year $t+1$) and the observations related to all the corresponding old codes (in the year *t*).

In other words, the parameters for the year $t+1$ can be produced by using new and old raw micro-data according to the correspondence between old and new commodity codes.

Riferimenti bibliografici

Eurostat (2006). *Statistics on the trading of goods - User guide*. Methods and Nomenclatures, Theme External Trade, Luxembourg: Office for Official Publications of the European Communities.

Fescina R., Jennings A., Wroblewski M. (2004). Automated production of foreign trade data edit parameters using resistant fences. *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association*, Toronto, Canada, August 8-12, 2004, pp. 3492-3497.

Garcia M. M., Gajcowski A., Jennings A. (2006). Selective editing strategies for the U. S. Census Bureau Foreign Trade Statistics Programs. *UN/ECE Work Session on Statistical Data Editing*, Bonn, Germany, September 25-27, 2006.

Jäder A., Norberg A. (2005). A selective editing method considering both suspicion and potential impact, developed and applied to the Swedish foreign trade statistics. *UN/ECE Work Session on Statistical Data Editing*, Ottawa, Canada, May 16-18, 2005.

Nuccitelli A. (2007). Determining edit parameters to revise data on external trade with non-European Union countries: the Italian experience. Paper presented at the *Seminar on External Trade Data Production*, Eurostat, Luxembourg, December 12-13, 2007.

Narilli M., Nuccitelli A. (2008). Re-designing the editing process for the Italian data on foreign trade with countries outside the European Union: some quality issues and results. *Proceedings of the European Conference on Quality in Official Statistics*, Rome, Italy, July 8-11, 2008.

SAS Institute Inc. (1993). *SAS/ETS User's guide*, Version 6, Second edition, Cary, NG: SAS Institute Inc..

Informazioni per gli autori

La collana è aperta ad autori dell'Istat e del Sistema statistico nazionale, e ad altri studiosi che abbiano partecipato ad attività promosse dal Sistan (convegni, seminari, gruppi di lavoro, ecc.). Da gennaio 2011 essa sostituirà Documenti Istat e Contributi Istat.

Coloro che desiderano pubblicare sulla nuova collana dovranno sottoporre il proprio contributo alla redazione degli Istat Working Papers inviandolo per posta elettronica all'indirizzo iwp@istat.it. Il saggio deve essere redatto seguendo gli standard editoriali previsti, corredato di un sommario in italiano e in inglese; deve, altresì, essere accompagnato da una dichiarazione di paternità dell'opera. Per la stesura del testo occorre seguire le indicazioni presenti nel foglio di stile, con le citazioni e i riferimenti bibliografici redatti secondo il protocollo internazionale 'Autore-Data' del *Chicago Manual of Style*.

Per gli autori Istat, la sottomissione dei lavori deve essere accompagnata da una mail del proprio dirigente di Servizio/Struttura, che ne assicura la presa visione. Per gli autori degli altri enti del Sistan la trasmissione avviene attraverso il responsabile dell'ufficio di statistica, che ne prende visione. Per tutti gli altri autori, esterni all'Istat e al Sistan, non è necessaria alcuna presa visione. Tutti i lavori saranno sottoposti al Comitato di redazione, che valuterà la significatività del lavoro per il progresso dell'attività statistica istituzionale. La pubblicazione sarà disponibile su formato digitale e sarà consultabile on line.

Gli articoli pubblicati impegnano esclusivamente gli autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat. Si autorizza la riproduzione a fini non commerciali e con citazione della fonte.