Check, edit and imputation of the variable turnover of the Italian Business Register

Viviana De Giorgi

1Istat – Italian National Institute for Statistics, e-mail: viviana.degiorgi@istat.it

Abstract

In 2005, Istat decided to evaluate the dissemination of a new variable of the Italian Business Register: the turnover. Since no check has been done until then, there was first an experimental phase on the data of the taxation year 2003. The Italian Tax Authority provided both the administrative sources used in the process. They were the VAT returns source, whose variable used is the Volume of Business, and the Sector Studies survey data, which provided with the same variable. The administrative data were integrated with the data coming from the Annual Survey on Italian Enterprises Accounts System carried out by Istat. The process has been running so far, and it has been included in the process of the Italian Business Register. It is indeed in line with the disposition of the Institute about the use of administrative sources for retrieving, integrating and comparing information about enterprises.

Since the reference year 2004, the use of the VAT time series returns and Sector Studies has made it possible to have the annual up-to-date turnover available for all the Italian Business Register units, thus improving the quality of the register and making it possible to disseminate data by turnover classes as well. In future, the turnover could possibly be used also as a stratification and/or domain variable for statistical surveys on Italian enterprises.

The process of checking, imputation and assessment of the Italian Business Register turnover consists of five steps, strictly in sequence: a) identification of the reference set to be checked, imputed and validated; b) identification of the turnover existence set of values for setting the turnover boundaries; c) data editing through time series consistency; d) data imputation of non-correct values (missing, zero and outliers) found on the basis of editing checks; 5) data validation using external indicators.

Keywords: administrative sources, outliers detection, data editing

1. Background and motivations

This work aims at reintroducing the topical issue of exploiting administrative sources in order to add information to the Italian Business Register (from now on ASIA1). The ASIA time series have been collected since 1996 and the register provides every year identifiers (name, location, etc.) and structural aspects (economic activity, size, etc.). It covers all the active enterprises (without any size limit), obtained from a process of integration of administrative files through statistical methodologies developed to estimate and impute all characteristics, yearly updated. The enterprises belonging to all the economic activity sectors, with the exception of the public and agricultural sectors are the reference population of the Italian Business Register. In addition, the

1 Italian Statistical Archive of Active Enterprises
data filed into ASIA are an integrated information system, suitable for the management of current surveys as it gives information on sectors, business size and structural classifications.

Information about the turnover of the Italian enterprises has been available in different administrative and statistical sources. The sources where we can find information about the enterprise turnover are:

1. **Sector Studies**, whose survey was instituted in 1996 by the Italian Tax Authority for evaluating the capacity of enterprises to produce income and knowing whether they pay taxes. The Sector Studies are undoubtedly one of the administrative sources of greatest interest for analysis because they gather a lot of information on more than 3 million enterprises (Bernardi et al., 2008). The subjects liable to the Sector Studies survey are Italian industrial, commercial and manufacturing income owners and VAT number proprietors for specific sectors of activity, though there are some exclusion and non-enforceability principles. Therefore, this survey does not cover the entire universe of ASIA. In the accounting table of the questionnaire the following information (variables) are available:
   a. **Volume of Business**, that is the amount of goods and services in the year, registered or liable to registration, taking into account changes in deduction (art.20 of the Italian D.P.R. 633/72)
   b. **Proceeds**, as they appear in letters a) and b) of art. 85, par. 1, of the TUIR, i.e. the total revenue for supplies of goods and services whose production or exchange is part of the business activity and the sale of raw materials and consumables, semi-finished goods and other personal property, excluding capital goods, purchased or produced for use in production, without taking into account other positive components of income
   c. **Agios and Income from the Sale of Goods Subject to Fixed Revenue**
   d. **Other Proceeds**, that is the amount of revenues other than those mentioned in letters a), b) and c) of art. 53, par. 1, of the TUIR
   e. **Other Positive Income Components**, that is all other revenues
   f. **Other Operations that Produce Income**, for example off-field operations (art. 2, art. 3, art. 7 and art. 74, par. 1, of the Italian D.P.R. 633/72); operations not subject to declaration (art. 36-bis and art. 74, par. 6 of the Italian D.P.R. 633/72)

2. **Annual VAT returns**, which are the declarations of the enterprises liable to the value added tax, gathered by the Italian Tax Authority. Whereas the Tax Authority provides Istat with all the Sector Studies data, for the VAT statements Istat gets only the variable named **Volume of Business**, whose definition is the same as it appears in the Sector Studies.

In the questionnaire of the Sector Studies there are more then one variable related to the business turnover, collected for different purposes to the VAT returns. The variable that is favoured as a proxy of the enterprise turnover is the **Volume of Business** from the annual VAT returns, mainly because the Sector Studies, as already mentioned, are a partial survey of the population of Italian enterprises, whereas the VAT statements are declared by (almost) all the Italian enterprises.

The ASIA turnover, except in some cases, was never disseminated before the taxation year 2004, despite Istat have been disseminating the results of surveys on the accounts of enterprises with: 1) an annual total survey on large enterprises (more than 99 employed persons) called **Survey on Italian Enterprises Accounts System**: the variable

---

2 Decree of the President of the Italian Republic
3 Italian consolidated act on income taxes
is the \textit{Sales and Services Proceeds}; 2) an annual sample survey on small and medium-sized enterprises (less than 100 employed persons) called \textit{Small and Medium-Sized Enterprises Survey}; also in this case the variable is the \textit{Sales and Services Proceeds}.

A new step of statistical quality for the Italian Business Register was represented by the usability of the new variable \textit{turnover}, available for all the ASIA enterprises after a procedure of check, edit and imputation that uses administrative and statistical sources.

Data from official business surveys is to be considered correct and, in some cases, it is preferable to impute the ASIA \textit{turnover} with such data instead of using the administrative ones. In particular, this is always true for large enterprises.

Before going on, it is worth clarifying some key points. In order to check, edit and impute the ASIA \textit{turnover}, we could use different solutions, such as automatic procedures for the identification of invalid data and/or software for editing and imputation of numerical data. In this context, and for the reasons we are going to describe now, we used instead an \textit{ad-hoc} procedure fitting the characteristics of \textit{turnover}.

One of the automatic procedures for the outliers detection retrievable on the Internet is the algorithm of Hidiroglou and Berthelot (1986) used also in the software BANFF described afterwards. Actually, the use of such a procedure, even if in the literature has been tested on several occasions, did not give good results in our case: in fact, the outcomes obtained by different checkouts did not meet a unique (or almost unique) set of outliers. Generalized software for checking, editing and imputing numerical variables called BANFF and developed by Statistics Canada (2005) is used in Istat. Such software requires the definition of control edits that have to be known in advance in order to be fitted to the data. Actually, in ASIA there are a few numerical variables, so we can hardly define effective edit rules to be applied.

For these reasons, we preferred using an \textit{ad-hoc} methodology that fits each time the available data rather than using an automatic procedure. With such a procedure the \textit{Volume of Business} declared in the annual VAT returns and Sector Studies and the \textit{Sales and Services Proceeds} declared in the Istat surveys are combined together after the checking, editing and imputation procedure. The \textit{turnover} of an enterprise in the year \(k\) obtained at the end of the process will be defined as the amount that enterprise generates in the year \(k\) supplying goods and services.

Now let us make explicit some mathematical notations, which are used in this paper. Let us define: 1) \(t_k\) as the \textit{turnover} an enterprise obtains in the year \(k\); 2) \(e_k\) as the number of persons an enterprise employs in the year \(k\); 3) \(tpc_k = \frac{t_k}{e_k}\) as the \textit{turnover per capita} an enterprise obtains in the year \(k\).

2. The \textit{ad hoc} process for the outliers detection and imputation

The methodology developed for the ASIA \textit{turnover} consists of single steps, namely: 1) identification of the set of enterprises to be processed; 2) identification of the \textit{turnover} range of permitted values; 3) outliers detection; 4) outliers imputation; 5) outcomes assessment.
2.1 STEP 1: Identification of the set of values to be processed

The enterprises whose value has to be subjected to the check, edit and imputation process are:

- the enterprises belonging to the reference population of the Istat surveys, that is all the economic activity divisions with the exception of the divisions 65 (Financial intermediation, except insurance and pension funding) and 66 (Insurance and pension funding, except compulsory social security). These enterprises, at the end of the process, will have a value of turnover (including non-zero values) equal to the Volume of Business as it appears in the annual VAT returns.

- the enterprises active during the reference period; the reason for that is due to the difficulty of imputing missing data for a great number of enterprises and assessing the detection process. In fact, the presence of the turnover value is one of the business activity signals, thus the percentage of turnover zero values among the non-active enterprises is about 85%.

- the enterprises active not less than six months during the reference period, since this is the reference population for the Istat surveys as well.

- the enterprises which by definition are liable to VAT declaration. The tax free ones, for which a zero value of the Volume of Business is valid, have been identified with those belonging to the economic divisions 67 (Activities auxiliary to financial intermediation), and 85 (Health and social work). These enterprises, at the end of the process, will have a value of turnover (including non-zero values) equal to the Volume of Business as it appears in the annual VAT returns.

- the enterprises that, in the last three years, have not been undergone an event of mergers or acquisitions or any other event that could have changed the capacity to produce income or, despite having suffered an event, have declared a zero value in the reference year. As a matter of fact, for the enterprises with events in the last three years and a non-null turnover value in the annual VAT return, sudden growths or declines can be justified from year to year due to the events they have been involved in. The enterprises excluded, at the end of the process, will have a value of turnover equal to the Volume of Business reported in the annual VAT returns.

- the enterprises belonging to the economic activity divisions with a significant number of enterprises, which means it can be possible to calculate the characteristic values. Regarding this, we excluded only the enterprises belonging to the division of “Mining and quarrying of energy producing materials” and “Mining of metal ores”, which are in general very few. These enterprises, at the end of the process, will have a value of turnover (including non-zero values) equal to the Volume of Business as it appears in the annual VAT returns.

2.2 STEP 2: Identification of the turnover range of permitted values

Using the annual VAT returns, the set of acceptable values of turnover declared in the current year is determined from the ASIA (checked and imputed) turnover value of the previous year.

In order to know the reasons for this choice, it needs to take into consideration that, first, since the taxation year 2003 the Asia turnover have been checked and corrected

---

4 Statistical Classification of Economic Activities in the European Community, Rev. 1.1
and, second, as the previous year data come from the same source of the current year, they are capable of managing the variability of the variable.

The Sector Studies source, which is a partial survey in respect of the VAT returns, will be used, in case, in the next step to impute (replace) the non-correct, missing or non valid VAT returns data. The point-to-point comparison of the turnover values between the two sources cannot be used for checking and validating the VAT returns data because the Sector Studies do not cover all the reference population, thus requiring a complementary method for businesses outside the observation range of the Sector Studies.

The previous year turnover values are clustered by economic activity groups (3 digits) and classes of persons employed. For each of these clusters the minimum and maximum values of turnover per capita are calculated. When in a group there is not a significant number of enterprises in order to determine these values, we consider the economic activity classes (2 digits) rather than the groups. Minimum and maximum values are used to define two sets of turnover per capita in the current year: 1) set P, that contains all values within the band-pass filter bounds or set of permitted values; 2) set NP, that consists of all values outside the band-pass filter, i.e. the non-permitted values. The values of the set P define the turnover range of permitted values. Missing and zero values are, by definition, assigned to the set NP.

In chart 1, we can see the algorithm of the step 2:

---

**STEP 2**

- Year k-1 values are grouped by economic activity and class of employed persons
- Calculation of \( \text{min}(tpc)_{k-1} \) and \( \text{max}(tpc)_{k-1} \)
- Pass-band filter for the year k
- \( \text{min}(tpc)_{k-1} < tpc < \text{max}(tpc)_{k-1} \)

** Provided that:**
- \( tpc \) = turnover per capita (from annual VAT returns in the year k)
- \( \text{min}(tpc)_{k-1} \) and \( \text{max}(tpc)_{k-1} \) = group i min and max values of turnover per capita in the year k-1
- (I) the min and max values of the group each enterprise belongs

---

**STEP 3**

**Chart 1: Identification of the turnover range of permitted values**
2.3 STEP 3: Outliers detection

The only use of the band-pass filter is not sufficient to identify the correct values: although it is quite effective against the extreme values, which are very high or very low data, it is not very effective with correct values, since it could identify non-correct values as permitted or permitted values as being outside the filter bounds. Therefore, another check phase is carried out with all the values: they have to pass a combined check through rules coming from: 1) the Sector Studies current year data; 2) the last three years annual VAT returns time series data. The latter one consists of a new filter defining the upper and lower bounds of the change in time series data of turnover and turnover per capita.

The check in time series consistency is carried out through the following ratios:

1) \[ R'_{k-j,k-i} = \frac{(t^i_{k-j} - t^i_{k-i}) \cdot (t^s_{k-i} - t^s_{k-j})}{t^i_{k-i} \cdot t^i_{k-j}} \times 100 \]

2) \[ R_{pc}^*_{k-j,k-i} = \frac{(tpc^i_{k-j} - tpc^i_{k-i}) \cdot (tpc^s_{k-i} - tpc^s_{k-j})}{tpc^i_{k-i} \cdot tpc^i_{k-j}} \times 100 \]

provided that: 1) \( s \) is the source taken into consideration: either the annual VAT returns or the Sector Studies (if their record values are available); 2) \( i \) and \( j \) (with \( i \neq j \)) can be either 0, 1 or 2, so that \( k \) is the current year, \( k-1 \) is the previous year and \( k-2 \) is the two years earlier and we can calculate the changes between years \( k \) and \( k-1 \), years \( k \) and \( k-2 \), years \( k-1 \) and \( k-2 \); 3) \( R'_{k-j,k-i} \) is the ratios for evaluating the time series consistency of the turnover passing from the year \( k-i \) to \( k-j \) in the source \( s \); 4) \( t^i_{k-i}, t^i_{k-j} \) is the turnover of the years \( k-i \) and \( k-j \); d) \( tpc^i_{k-i}, tpc^i_{k-j} \) is the turnover per capita of the years \( k-i \) and \( k-j \).

Therefore, provided that \( k-j \) and \( k-i \) are the years we want to carry out the changes, \( R'_{k-j,k-i} \) represents the cross product between the dual ratios for the source \( s \):

1) \( \left| \frac{(t^i_{k-j} - t^i_{k-i})}{t^i_{k-i}} \right| \), that is the absolute difference between the source turnover in the two years divided by the turnover value of the year \( k-i \); 2) \( \left| \frac{(t^s_{k-i} - t^s_{k-j})}{t^s_{k-j}} \right| \), that is the same difference divided by the turnover of the year \( k-j \).

Mutatis mutandis, we can also compute the same ratios for turnover per capita.

Since such ratios are symmetric, they consider in the same way positive changes (increasing in value from the year \( k-i \) to \( k-j \)) as well as negative ones (decreasing in value).

As we mentioned above, if the Sector Studies record information is available, it is used in some cases to confirm the annual VAT returns data and in some other cases to correct the VAT returns value when it does not respect the time series consistency (Sector Studies imputation).

In order to consider at the same time the values reported in all three years \( (k, k-1, k-2) \) we construct a binary vector of three elements (3-tuple) as it follows:

\[ I' = [I'_{k,k-1}, I'_{k,k-2}, I'_{k-2}] \]

where \( I'_{k-j,k-i} \) is the binary value (0 = good, 1 = not good) corresponding to \( R'_{k-j,k-i} \).

Each binary value allocation is subject to the source ratio thresholds. These thresholds are computed based on the minimum and maximum values of the source absolute and
relative turnover and turnover per capita changes in the two years taken into consideration. In this way, we obtain:

- consistent indexes, for both the Sector Studies and the Vat returns source
- non consistent indexes, for both the Sector Studies and the Vat returns source
- a (possible) signal about the non-correctness value of the previous years, should they have been imputed

We check first the consistency in time series of the VAT returns current year indexes, and in case of non-consistency the consistency of the Sector Studies binary 3-tuple is assessed. This two-step time series check operates in the same way with both sets P and NP, with the only difference on the threshold values: they are stricter or more relaxed according to the changes taken into consideration (years and/or data source). The following chart shows the algorithm of the step 3.

At the same time of the check of the current year data the accuracy of the two previous year data is assessed, thus identifying any mistakes in the procedure applied the previous years. Should the previous year data item turns out as non-correct, an appropriate flag of incorrectness is assigned to it, so that this information can be used in the subsequent step.

In order to understand how time series checks operate, we can read the following chart.
1.4 STEP 4: Outliers imputation

The planning of this step was the most difficult phase of the process, since we had to go backward to the previous steps in order to identify the most reliable method of imputation.

For large enterprises, the values declared in the Istat survey is replaced as they are in this survey, whereas for small and medium-sized enterprises we use the VAT returns and the Sector Studies sources to impute a new data item. The enterprises with a non-correct value of turnover (missing, zero or invalid) are divided into: 1) enterprises with a missing, null or non-correct previous year turnover value; 2) enterprises with a correct previous year turnover value. As we mentioned earlier the correctness of the previous year value is already assessed in the step 3. The values of the first set enterprises are imputed using the following formula:

\[ k_{xy} \cdot tpc_{xy} \cdot e_k = k_{xy} \cdot tpc_{xy} \cdot e_k \]

provided that: 1) \( k_{xy} \) is the current year imputed turnover value of the enterprise; 2) \( tpc_{xy} \) is the current year turnover per capita of the activity groups (3-digit) or division (2-digits) \( x \) and the class of employed persons \( y \) (group \( xy \)). The value \( tpc_{xy} \) is calculated dividing the current year amount of turnover correct values by the number of persons employed in the group \( xy \). We call it mean value imputation.

The first time we tested the procedure we used a set of different values, concluding at the end that using data coming from the same source was a more accurate method. As a matter of fact, the value \( tpc_{xy} \) is the result of calculations on the correct values of the current year in the class \( xy \), thus not taking into consideration any zero, missing or non-correct values.

For the enterprises with a correct previous year turnover value, we use the turnover per capita average changes in value between the previous and the current years that
occur in the group \( xy \), i.e. the following formula: 
\[
\dot{it}_k = (tpc_{k-1} + tpc_{k} \cdot \Delta tpc_{xy}) \cdot e_k .
\]
The quantity \( \Delta tpc_{xy} \) is the percentage turnover change in value between the previous and the current years that occurs in the group \( xy \), that is the difference in turnover per capita between the current year and the previous year in the group \( xy \) divided by the previous year value multiplied by 100.

In this case, the average change is calculated only with the correct data of the group \( xy \) as well. We call this algorithm previous year imputation.

**Chart 4: Outliers imputation**

2.5 STEP 5: Outcomes assessment

The last step, in the experimental phase, had helped us to redefine some characteristics of the previous steps, in particular the identification of the set of enterprises to be processed.

The assessment of the imputation outcomes is done by aggregated data, which means comparing the values between ASIA and the Istat surveys by economic activity divisions and classes of persons employed. Each one of the following indexes calculated for all groups is able to recognize each comparison aspects:

i. by group absolute differences and/or ratios of the number of enterprises between ASIA and the Istat surveys

ii. by group absolute differences and/or ratios of the number of employed persons between ASIA and the Istat surveys

iii. by group differences of the amount of turnover between ASIA and the Istat surveys, before and after the imputation process

iv. by group amount of imputed value of turnover and percentages of enterprises whose turnover has been imputed, for both all the values and only missing and zero values.

v. by group average values, before and after data imputation

Some explanations of the indexes have to be given. As the universe of the Istat small and medium-sized enterprises surveys is an expansion through sample coefficients, the number of enterprises and employed persons by groups could be quite different from ASIA, thus justifying part of the difference of the turnover amount calculated by groups. Therefore, we have to calculate the first two indexes in order to explain
correctly the turnover indexes. The remaining indexes are useful to understand whether and how imputation represents an improvement of data quality, if all the indexes are analyzed jointly.

Mutatis mutandis, from the second year of turnover availability for dissemination, we have also calculated similar indexes between the current and the previous years, thus having another mean for assessing the results.

3. Conclusions
The dissemination of the ASIA turnover is done by classes of turnover, which for value greater than 1 million euros agrees with the Eurostat ones. If we consider the last 10 years of data available on the enterprises subjected to the process, we can see that the number of missing values falls drastically in 1999 and decreases in the following years, stabilizing in recent years on about 3%, with a significant improvement of data quality. Moreover, prior to that year missing values were indistinguishable from zeros. After the process of check, detection and imputation of the turnover outliers, about a 5% of values have been imputed.

Getting on in years, we have ascertained that for particular economic activities the variable whose value is the nearest to those declared and disseminated by the Istat surveys is the variable called Proceeds. So, there is an ongoing study for improving the procedure in order to make the ASIA turnover more comparable with that coming from Istat surveys and, second, to build up an effective imputation for economic activities non liable to VAT. This procedure should use both the Proceeds coming from the Sector Studies and the Volume of Business coming from the VAT returns, taking into consideration that the Sector Studies are not a total survey. The development of such a procedure is not trivial, as Sector Studies do not cover the entire population, so that the two variables would need to be worked out together keeping unchanged both time series and internal consistency.

References
Italian DPR n. 633/72 on VAT returns (Italian).
TUIR Italian consolidated act on income taxes (Italian).