ANALISYS ON ECONOMIC FISCAL DATA FOR STATISTICAL USES

Seminar: "Using Administrative Data in the Production of Business Statistics - Member States Experiences".
18th-19th March 2010

Mr. A. Bernardi, Ms F. Cerroni, Ms V. De Giorgi

Abstract

Thanks to the availability of a huge amount of information coming from administrative sources, the division of Istat in charge of administrative sources acquisition has led some analyses and completed important experiences on the Tax Authority survey source (Sector Studies survey), in order to make it available to Istat departments requiring such data to analyse their use for statistical purposes. The Sector Studies survey can be considered a very powerful source of information for around 4 million enterprises about accounting data, occupation, structural and specific aspects of the business activity. It has been acquired by Istat since the taxation year 2004, and all the analyses conducted so far reveals that fiscal information gathered by it are comparable to SBS regulation variable and thus can be used for statistical purposes. This document shows how data coming from administrative sources can be processed in order to be validated for statistical purposes. It illustrates some formal schemes and their applications to check the meaning and to evaluate numeric variables derived from Sector Studies. Although the procedures have been developed and applied to Sector Studies source, they suggest a standardized statistical approach that allows their use on other similar sources: they can change according to the existence of benchmark variables and to the degree of similarity to the administrative ones under investigation.

Preface

For several years, the Italian Statistical Institute (Istat) has been gathering sources from administrative bodies. The activity to ensure their overall inter-connectibility and compatibility with existing surveys data falls within the context of the more general program for public expenditure containment and promoting the more efficient use of public data by determining whether such data are able to integrate/substitute existing surveys.

Since other national statistical institutes have already been opting for the same method, this approach shows some promises in spite of several known disadvantages. The most significant one is that the collection and handling of administrative data occur outside of the control of the national statistical institutes. As a result, critical elements such as the definition of survey units, classification variables and analytical variables are guided by administrative decisions, which means they always exhibit some differences from those used by national statistical institutes (Wallgren and Wallgren, 2007). The two-fold consequence is, first, the considerable effort to determine the statistical utility of administrative data (P.J.H. Daas et. al., 2009) and, second, the dependence that the Institute will tend to acquire on the availability of externally-gathered data, and thus on their quality. This situation makes it extremely important to identify a procedure for assessing the overall quality of administrative sources in a systematic, standardized and reliable way, including whatever operations are required to conclude the validation process.

This paper is based on the experiences reached gradually at the Department for administrative sources and business registers - Division for administrative sources (Dcar/Dam/B) in regards to the validation process for economic-accounting variables from section F of the Sector Studies form (Bernardi et al., 2008 – Cerroni and De Giorgi, 2008), which were used to produce the first validation process nucleus for turning administrative sources into statistical ones. It should be noted, in particular, that even though the method used with the Sector Studies source has involved training on such specific data, it is still generalizable, with appropriate modifications, to any administrative source.

1 General features of the validation process

The acquisition of an administrative source is a process that begins when requested by a division of the Institute and after the approval of the persons in charge. This process, summarized in the table 1, is divided into three phases, and each of them involves verification and/or elaboration phases with associated quality indicators as defined on the basis of the results.

Working phases
1. Analysis of the body detaining the administrative source and its relationship with the Institute
2. Analysis of metadata from the administrative source
3. Assessment of the administrative source

Table 1 - Analyses for converting administrative sources into statistical ones

The verifications and/or elaborations done in each phase generate quality indicators that can summarize the possibility to use the outcomes with a judgement of: 1) positive assessment (+), which indicates that the operation can be worked out and completed within the scheduled time period, with the available resources and with a good quality of the results; 2) partially-positive assessment (+/-), which indicates that additional time and/or resources were required to complete the operation in question or the outcomes was of lower quality; 3) negative assessment (-), which indicates that it was not possible to complete the operation; this includes cases in which the achievement of an adequate quality standard required excessive time and/or resources relative to the benefits obtained.
The scheme just described makes it possible to specify a purely qualitative summary of the entire validation process for a specific source in the form of a simple count of the signs acquired in each of the three phases and even compare their degree of integration relative to others. Except for cases requiring different data treatment methods, it can be assumed that the operations from each phase are equally significant, so that a negative outcome for any one of the three is sufficient to disqualify the administrative source for statistical purposes.

1.1 Analysis of the body detaining the administrative source

The first phase involves a general assessment of the administrative source, and begins with an assessment of how feasible it is for the Institute to establish a collaborative relationship with the administrative body. This is followed by an assessment of how significant this administrative source is for statistical purposes, e.g. for expanding the informational basis of surveys and reducing the statistical burden. It is equivalently important to verify precise compliance with data confidentiality and privacy laws, the manner in which the administrative source is delivered, the safe procedures for uploading it, including all related economic and/or information-based aspects. It is also important to observe whether the timing of the availability of the administrative source and the information-gathering needs of the Institute are acceptable, and whether the administrative body applies its own quality controls to the source. Lastly, the stability over time of the above characteristics needs to be evaluated.
Assuming a significance equivalent to the previous operations, each of these can be assigned one of the three qualitative scores indicated above (+, +/-, -), with the count of results being used to summarize the overall assessment for this first phase of operations.

1.2 Analysis of metadata from the administrative source

The second phase involves verifying the administrative source metadata. This consists of checking whether there is any identifier identical to those used by Istat, and whether there is a clear correspondence in the definitions of statistical units, classification variables and analytical variables. One can then check how many variables (relative to the total number of variables in the source) are subjectable to the validation process and whether the previous metadata exhibit temporal stability as determined by comparing the current validation process with previously applied methods. Also in this second phase, we can summarize the outcome by means of the count associated with each individual indicator, according to its correspondence to statistical purpose.

1.3 Assessment of the administrative source variable

In the third phase, the administrative source integration into a statistical one is assessed through a series of operations and controls of the variables being integrated, which are examined one by one. The first step is to distinguish whether or not each administrative variable has a fully corresponding statistical variable. The first possibility, that starts the statistical procedure 1, is that one of the existing variables from surveys measures the same phenomenon as the administrative variable under examination. The measurements of these two variables are then compared to determine whether the analogous definitions correspond to compatible, or even identical, values. In this case, we can assume the existence of a direct control variable. The second possibility is subdivided into two approaches: the first refers to links with outside variables, external to the administrative source, that exhibit reasonable functional links with the administrative variable under examination. The second refers to situations in which there are considerable doubts about the functional connection hypothesis. The first approach is a quality check for the administrative variable to be integrated. It represents an effort to estimate the functional relationship of the variable with another one that may not be fully correspondent, but definitely shows some connection, so undertaking the role of functional control, different from the direct control described above. The second approach assesses the quality of the administrative variable in question by looking for connections with other variables in the administrative source itself that are suitable for supporting its interpretation.
This process is showed in the figure 1.

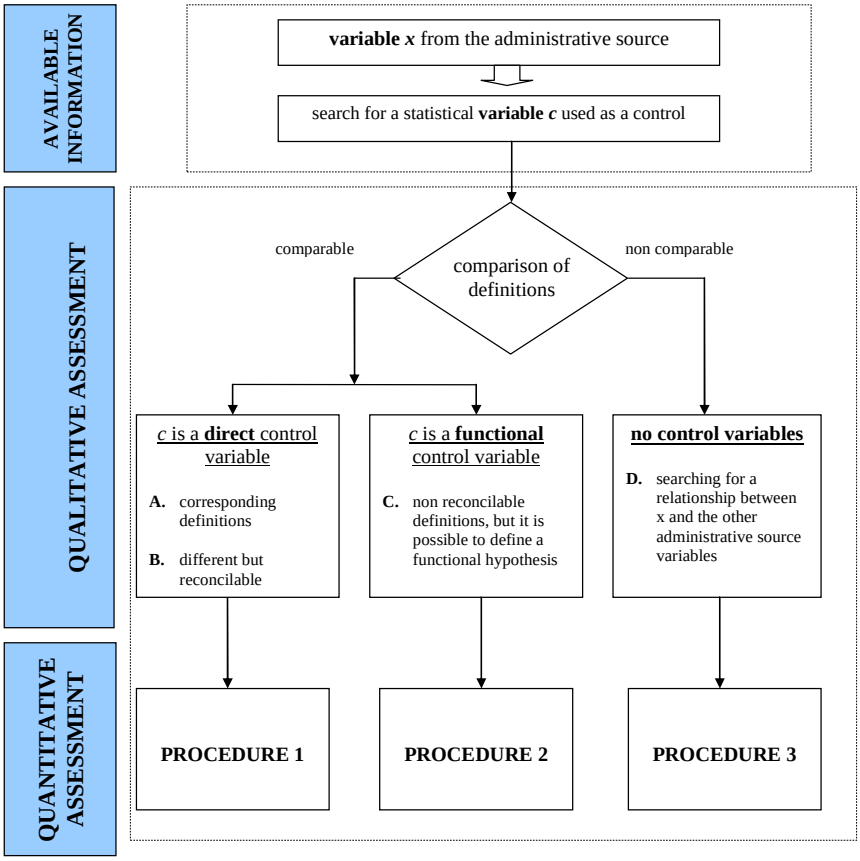

Figure 1. Assessment of an administrative variable

As seen in the diagram above, the administrative variable under examination (indicated with an x) is validated using methods that depend on the presence or absence of a control variable (indicated with a c). The diagram above illustrates the four distinct results, indicated with the letters A through D.

Case A is more rare because it requires a precise correspondence between the original definitions, whereas case B is more common because of how the definitions can be reconciled. Suppose x and c are aggregates that can be re-modeled using different combinations of elementary components: for example variable x represents the sum of n+1 variables and variable c represents the sum of the first n variables, so that excluding the extra component from x can make it equivalent to c.

Case C is for situations in which there are no control variables with definitions corresponding to x. If we assume, however, the existence of a variable c that can be related to x, a different type of validation could be attempted. While the previous validation were based on a precise comparison of observations of x and c, the c can now play the role of functional controller for x, which could be considered validated if its overall behaviour relative to c presents logical and coherent functional dynamics. Each case calls for great care and for checking whether the comparison between the new, modified definitions (with a component curtailed, for instance) preserve the meaning of the previous comparison based on all entries in order to confirm that the weight of the component excluded from x (for equating it to c) is insignificant relative to all of the others.

The last scenario, case D, accounts for the absence of any control variable for x (neither direct nor functional), and the quality analysis begins by relating it to one or more variables from the same administrative source. Here we shift from the search for external coherence (by comparing x with a variable c from outside of the administrative source) to an internal coherence, comparing x with variables from the same source. In the Sector Studies source analyses this situation can be encountered with variables from other sections, such as section B (structural elements) and section D (specific aspects of the economic activity in question), that contain extremely specialized information that could effectively be internally related.

The specific statistical procedure to apply is determined by the categorization into types A-D. Types A and B use statistical procedure 1, which is useful whenever it is reasonable to presume that x actually corresponds (with possible adjustments) to c. It consists of getting an overview of the possible differences between x and c through a series of indexes of location, scale and distributive form.

If a correspondence between x and c is confirmed, the values of x and c supply the range of admissibility for any values of x that could not be connected to c. This is what happened with the Sector Studies survey variables and the control variables from the small and medium-sized enterprises survey. The first ones result from a global survey while the second ones are sample-based, which in practice made it possible to match up most of the enterprises from the small and medium-sized enterprises survey (of the sample) with the same enterprises in the Sector Studies, with the rest of the Sector Studies enterprises obviously remaining unmatchable.

If a correspondence between x and c is not confirmed, the proposed control variable can no longer be treated as such, though it could still be used as a variable that is functionally connectible to the first, as described next for category C.
Category C calls for the application of statistical procedure 2, which consists of a series of exploratory analyses (data mining and regressive) designed to reveal the functional relationship between x and c, a variable that is presumed to exist. This leads to the construction of a confidence zone for selecting cases of x that are compatible with c, including both matched and unmatched cases. If the outcome is negative, i.e. when no satisfying functional connection is found between the presumed control variables and x, the process shifts to category D, which is addressed next. Category D begins with statistical procedure 3, which calls for targeted analyses that are designed to uncover linkages between x and other variables from the administrative source with the potential detection of outlier.
Three comments should be added. First, due to the fact that all the procedures described above are being applied to data sets that are difficult to handle due to their large size, a single analytical step may at times require two alternative techniques to be used in a complementary fashion in order to clarify otherwise ambiguous results. Secondly, the operations could also create intermediate situations in which the data analysis does not point clearly to only one of the four categories (A, B, C, D). Thirdly, the descriptions found in the pages that follow refer to the relevant literature for various clarifications.

2 Statistical procedure 1

This statistical procedure begins with the individuation of case A or B, and is subdivided into steps as showed in figure 2.

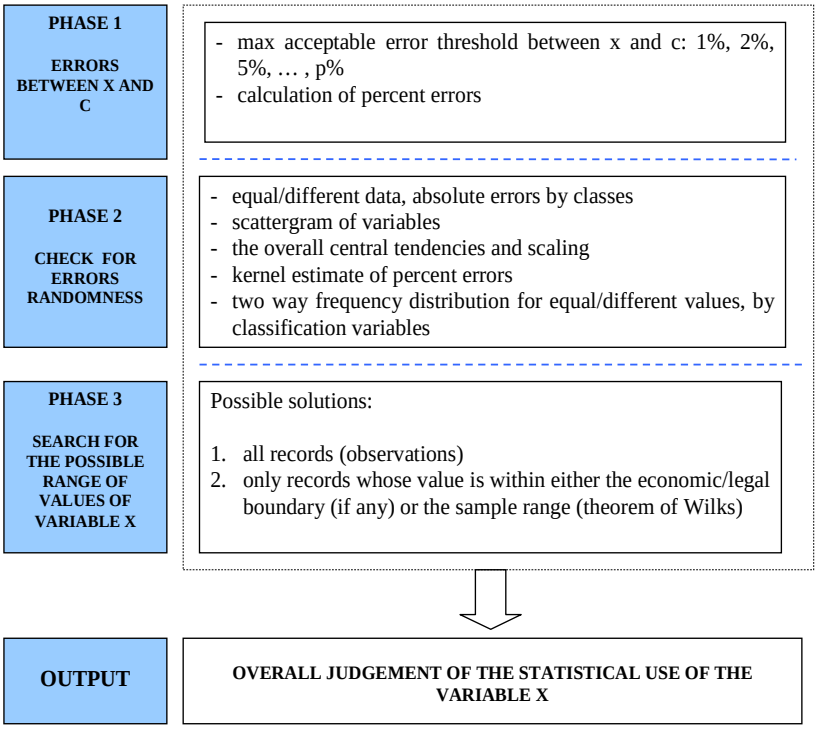

Figure 2. Scheme of the procedure 1

Given that different forms of analysis can overlap to some extent and that this can be justifiable, given that interpreting the results of a statistical analysis is not always easy and unambiguous when dealing with extremely large data sets, cases A and B presume that there is an adequate level of correspondence between the definitions of variables x and c. It starts out by counting the number of equal observations and the number of different observations, which already provides an important initial assessment of x. Based on the specific variable in question, an admissible deviation of x is then hypothesized relative to a c maximum of ±p%, for example, where p can vary depending on the specific situation (type of variables, type of source, purpose of the data integration, etc.). A grid is set up with deviations of different classes: ±1%, ±2%, ... , ±p%, over ±p%, and the relative frequencies are calculated. This is followed by the examination of the frequencies in each class and selection of cases that fail to respect the $|(x-c)/c|*100<p\%$ constraint in order to confirm that they are few numbers and have some feature that makes them appear to be random. In general, if the definitions correspondence hypothesis is true, it should be observed that:
1) the excluded observations should be the minority, and a clear prevalence of smaller deviations occur;
2) a scattergram on x and c should show a cloud of points around a 45° straight line, with the points distributed equally above and below the line;
3) the overall central tendencies and scaling for x and c should be very similar;
4) the distribution of (x-c)/c deviations should be characterized by a unimodal kernel estimation, be symmetrical around

zero and show a marked kurtosis[1].
5) the two way frequency distribution by classification variable (Nace code, geographical area, etc.) as a row variable, and with the modality given equal/different as a column variable, should exhibit similar row profiles (Agresti, 2002; Cameron and Trivedi, 2009) pointed out by a chi-squared test measuring the association between the variables and distinguishing statistically between differences and significant differences.
Subordinated to point 1, the definitions correspondence hypothesis can still be acceptable when the number of different cases observed between x and c is not small, but the deviations for most of these cases are close to zero. In this case, it could be useful to further qualify the deviations, such as by fitting it to a normal curve with 0 mean and variance to calculate, even if it should be remembered that in the presence of large numbers, goodness of fit tests receive too much power, which in practice induces them to reject normality, i.e. the h_0 hypothesis even for small discrepancies.
Finally, in case of an effective quantitative correspondence between x and c, the range estimated for cases of x matched with c can be used to define an admissible range for selecting the unmatched cases of x. This is due to the theorem of Wilks, which asserts that for large data samples, apart from the distribution shape, the range between the maximum and the minimum as results from the sample provides a consistent estimate of the corresponding population range. In other terms the probability that the sample range includes the entire population tends towards 1. It is well known that the SME sample units (variable c) represent a random, balanced sample of the world of SME enterprises. If we assume that the same holds true for observations of Sector Studies (variable x) matched up with c, this means that such observations of x also constitute a random, balanced sample of the population.
As an example of the application of the procedure 1, we can describe the case of turnover of Sector Studies survey compared with the SME variable. It can be treated as a type B case, since the variables do not correspond exactly in definitions but it needs a combination of elementary components. About 77% of units have the same value of the variable, so that different observations count for 23% in terms of frequencies and 28% for values.
As the overall central tendencies and scaling for Sector Studies and SME variable are very similar, the focus on different values units points out errors less than 5% for more than 90% of units and they graphically show a distribution which is unimodal, symmetrical around zero and with a marked kurtosis. The scatter plot describes a cloud of points around a 45° straight line, with points equally distributed above and below the line and only few units very far from them: the definitions correspondence hypothesis is well confirmed as smaller deviations are prevalent and different values are a minority and have random features.
Also the two way frequency distribution of equal/different value classified by economic activity exhibit similar row profiles.

3 Statistical procedure 2

This statistical procedure performs when the definitions correspondence hypothesis does not work and as an alternative it is possible to define a functional hypothesis where c is the functional control variable (case C of figure 1).
Figure 3 displays three phases included into statistical procedure 2 aimed to analyze the relationship between x and c.

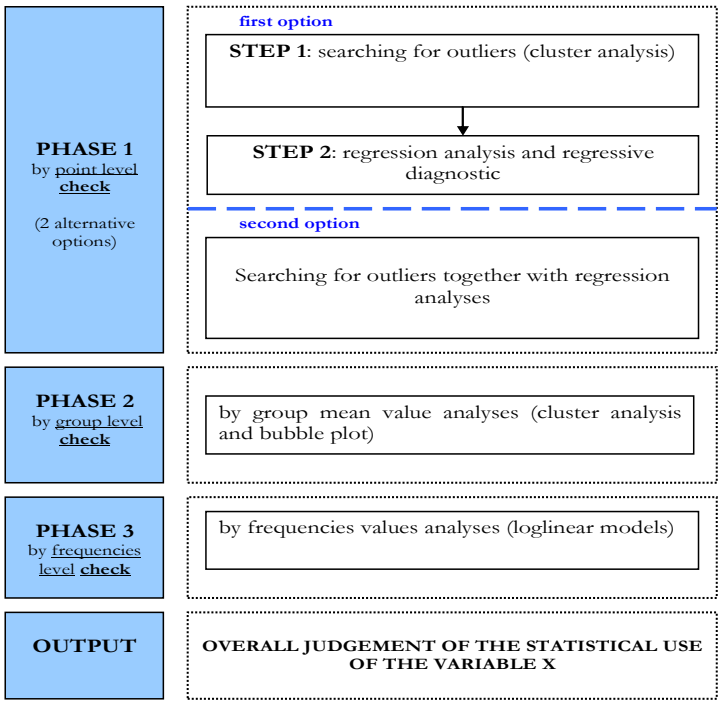

Figure 3. Scheme of the procedure 2

[1] The kernel estimation is a smoothed representation of the histogram that makes it easier to check its main characteristics (unimodality, symmetry, etc.).

The first phase has two alternative options: the former works in two steps by initially removing outliers and only afterward estimating the relationship between x and c. The latter performs the same task working the two steps simultaneously. The first option may result less satisfying, therefore it is used only if the second option can not be applied. If both methods fail or give ambiguous results, the subsequent phases may be used as an alternative especially for large datasets. In fact the third phase describes graphically the relationship between x and c through the analysis of group averages trends by using the cluster techniques. Finally the fourth phase, instead of analysing the relationship on a point level, works on a frequency level by applying log-linear models to verify a nearly-independent model hypothesis.

3.1 Statistical procedure 2, phase 1 – first option

This phase begins with the search for irregular values and outliers that have to be identified and excluded from the set before going on with any quantitative assessment. Irregular data and outliers are distinguished in the following way: a value or a specific attribute that should not be in the set of data is defined as irregular; data distant from all other observations, because they are extreme, are instead defined as outliers. For example, if a source aimed to hold sole proprietorship companies includes a private company, this one is an irregular data; if Sector Studies were to include a company with a turnover that exceeds the survey maximum value established by law in € 7,5 thousands euros, this is an outlier.
Irregular data concerns with the nature of the data and can be detected by using legal criteria. Outliers concern with the structure of the data as they are due to a high variability of the phenomenon under examination. They can be detected by a scatter gram of x and c, however when the graph is not easy to understand - as often happens in large data sets - a cluster analysis can be used to identify extreme data as they belong to isolated clusters (i.e. distant from the others), have a low number of elements (i.e. low frequency) and have relatively modest dimensions (i.e., a limited radius).
Clustering analysis is carried out using one of the following two techniques which provides unambiguous results:

A. two steps k-mean clustering procedure
B. single linkage clustering procedure (Everitt, 2001).

The k-mean clustering procedure is applied in two steps. In the first step the clustering procedure assigns extreme data to clusters which are cut off in terms of their *large* distance from the nearest cluster, their *low* frequency and their *small* size. To confirm the results, a correlation matrix is built up by using clusters as rows and the three variables mentioned above - distance, frequency and size - as columns. This matrix should exhibit a positive correlation between frequency and size, a negative correlation between distance and size and between distance and frequency. Once the cut off centroids are leaved out, in the second step the k-mean cluster procedure is only applied to centroids found as regular, then it is carried out under the constrain that none of the remaining clusters has to be larger in size than a threshold value derived from the previous step. The assignment of regular centroids at the beginning of the clustering procedure makes the process easier to converge in the two sub-groups of extreme and regularly clusters of data; the constraint on the maximum admissible cluster size keeps instead the cut off clusters from being included in regular clusters by chance.
The single linkage clustering procedure is based on a nearest neighbour clustering which tries to achieve the same scope as the iterated clustering A using a different method: data are grouped into clusters by using the distance between their two closest points (nearest neighbour) as a measure of the linkages between two groups, so that the procedure ends up leaving out points in clusters that are very far away from the others. The advantage of this procedure seems to be that it avoids the two steps of the k-mean method and the setting of the cluster radius, which is not always a simple operation. Sometimes the graphical evidence can be difficult to understand, especially when dealing with large data sets. All the observations identified by both procedures are classified as extreme and discarded and their identification is followed by the check for their randomness after which the first part of statistical procedure 2 ends up. The next step is to calculate the relationship between x and c by estimating the regression and checking the consistency of the coefficient in terms of sign, value and significance.
Lastly, a regressive diagnostic can be applied to the estimated relationship. Some of the following exams can be used (Cameron and Trivedi, 2009, p. 91-97):

1. a residual diagnostic plot. The graph summarizes the regressive model results by reporting the adapted values on the abscissas and the residual values on the ordinates. However it is well know that such a scatter plot is "less used in micro-econometrics" due to the large number of points involved
2. a check for influential observations of x, where a check of the leverage of the dependent variable at the point level is needed
3. a test of the distribution of the relationship between x and c: the Box Cox transformation of the relationship is calculated
4. a specification test for the omitted variables: models with powers of x, dummy variables etc. are tested by the Wald test
5. a test to verify the distribution of the conditional average: the Ramsey reset test is complementary to the Wald test for omitted variables
6. a test of heteroscedasticity: this is the Cook-Weisberg test
7. an omnibus test: this refers to the information matrix (IM) test, that is used to check for symmetry, kurtosis and heteroscedasticity of errors.

If relatively few cases are excluded, the model can be estimated again using only the regular points.

The regressive diagnostic above ends up the first phase of procedure 2.
Given the hypothesis that predictive variable c is a representative sample as it contains all values that are generally assumed, the result of the model estimation will be to provide a range of confidence or admissible values for x which can be used to distinguish, among the unmatched units, the values of x that are compatible from those that are not compatible as they lie outside of the aforementioned range.

3.2 Statistical procedure 2, phase 1 – second option

As an alternative to the previous one, this alternative option estimates the relationship between x and c and at the same time identifies the outliers. Firstly an m-band regression and/or a lowess regression are applied to explore the distributional form of the relationship between x and c, afterwards a quantile regression and/or a robust M-type regression can be applied to identify the relationship.
The m-band regression is a non-parametric regression model as it does not generate an explicit regressive equation by estimating the parameters of the relationship. Since it allows the researcher to explore data without any specification on the shape of the distribution, the method can be used as exploratory instrument to verify a non-linear distribution relationship.
Data have to be sorted by increasing value of the regressor (c in this case) and the number of desired sub-groups has to be fixed in advance. For example if you have 1,000 coordinate points and 10 is the number of sub-groups specified, this means to fill in the first sub-group with the first 100 coordinate points, then from the points 101 to 200, …, and so on up to the points from 901 to 1,000. The median which is the index of robust central tendency is calculated for variables c and x by sub-group. These points (x_1 and c_1, x_2 and c_2, … , x_{10} and c_{10}) joined by a straight line can be displayed by a scatter plot together with the 1,000 original points.
The lowess (LOcally WEigthed Scatterplot Smoothing) regression is similar to the common regression of vector x on vector c where in the first vector observations x_i are replaced by the mean of a fixed number of previous and subsequent observations of x_i values with weights increasing by growth of the error between observed and adapted value.
Quantile regression is a robust type of regression which has the advantage of been semi-parametric (in the sense that it generates estimates for coefficients and does not require hypotheses on the parametric distribution of the errors, except for their been i.i.d.). Quantile regression can also check the relation between x and c for specific quantiles of the dependent variable (Cameron and trivedi, 2009, p. 209), not just for the entire distribution as a whole, and allow to monitor sections where the relation between x and c get worse.
The robust M-type regression is also semi-parametric and may turn out to be even suitable to the purpose here, since it is highly resistant to extreme data. Being the control variable, in this context c is presumed to be free from errors while they can still be found in the dependent variable x. The OLS regression is applied to all observations that have a Cook D statistic greater than 1 as all the remaining observations are too influential. A Huber Window (smaller weights are assigned to less adaptable data, and vice versa) is used to calculate the weights for each observation of x, and the iterated weighted regression is calculated. This is followed by a new window of weights that can provide 95% efficient OLS coefficient estimates (Hamilton, 2009, p. 256-257).

3.3 Statistical procedure 2, phase 2

If both alternative options of phase 1 fail to show any connections between x and c by point level, another solution is to check whether a simple relationship can be detected by group averages level. This approach is largely suggested in the economic literature for large sample data (Agresti, 2002, p. 180 - Khattre and Naik , 2000, p. 300). Each group is synthesized by average group values applying the k-mean clustering procedure and then the concordance is confirmed by checking the linearity of the average group values through a bubble graph. If they do line up, this indicates that variables x and c are in fact compatible in terms of their average values, despite the fact that they exhibit no close linkages from the point-based perspective.
As an example of the application of the procedure 2, phase 2, we can describe the case of personnel cost of Sector Studies compared with the Sme variable, which can be treated as either a type B or type C case. Actually, since the Sector Studies variable does not correspond exactly with the sample variable because of some differences in definitions that leads to different values in 30% units, we can go ahead by considering only the units having a non-zero personnel cost, of which 54% have equal values. All the explorative regressions (band regression, lowess regression) have brought to similar conclusions: the Sector Studies and sample variables show a I-III bisector relationship, if we exclude both some high values (besides 3/4 million euros) and some units having non-zero Sector Studies variable value and zero survey variable value. Having explored the overall relationship, we can now consider only the units having different personnel cost values. In this case too, all the regression analyses (linear regression, median regression, robust regression) strengthen that a very high percentage of units (from 2-3% to 7%, depending on the model used) having comparable values in the two sources. In fact, the slope is around 1 and the intercept is very low (few thousand euros) which means that the variable can be validated.

3.4 Statistical procedure 2, phase 3

Whenever the previous approaches fail to give a clear idea of the connection between x and c, another test involves an analysis of the data by summarizing their frequencies through the discretization of variables. This method, which is recommended in the literature (Basilevsky, 1993, p. 508 and subs.), is based on the idea that data measurement errors will

have less effects if one analyzes the frequency distributions obtained from the elementary data instead of the elementary data themselves. To accomplish this, the optimal number of classes is identified using an automatic criterion (Piccolo, 2000, p. 66) and the equal frequency classes criterion with a fine grid, i.e. non-masking the details of the distribution - classes that contain each 5% of the cases, for example - and the concordance suggested by the two criteria is then verified. An exploratory analysis is then performed on the double entry table by calculating the weight of the cases located on the main diagonal and in the upper and lower triangular parts over the total number of cases, after which the Cohen k coefficient is calculated. Introduced in 1960, the kappa coefficient estimates the degree of concordance between the two variables of a square table. It considers the concordant elements located on the main diagonal and the concordant elements that would be obtained if the variables themselves were independent.

Lastly, the exploratory analysis is followed by the estimation of specific log-linear models, such as:

1. the IM model (Independence Model), which hypothesizes that the frequencies are randomly distributed in the contingency table cells and, in fact, it translates into a testing of the independence between the rows and the columns (Allison, 1999, p. 44). If the model passes the verification phase, we conclude that the data match up primarily on the main diagonal and this means that no other dynamic is present in the table;
2. the UL model, which tests whether the association between x and c weakens from the main diagonal.

4 Statistical procedure 3

This procedure has to be used when a control variable c is not available. In general, it can be observed that when a lack of a control variable is detected the only way to assess the quality of is to link x to other variables from the same source. A positive outcome means that x is reasonably connected to other variables of the same source and there is internal coherence. With respect to the selection of observations, high factorial point scores can be used. After the analysis, it will be possible to point out and understand links between variables inside the groups and their respective latent factors in terms of the complementary and substitute concepts of variables.

The scheme shows 4 steps:

1. the search for theoretically-connectible variables;
2. preliminary check and identification of the extreme data;
3. factor analysis;
4. if necessary, a confirmation analysis using multidimensional scaling.

The procedure starts searching for the potentially connected variables, by following the general criteria to carry on an higher number of possible variables rather than selecting only few from the beginning.

After a preliminary analysis with *sqm* and *mad* it is clear data include outliers. A method based on principal components (statistics Hisq and Disq) is applied to identify outliers and since they are not easy to understand due to the large number of data, these statistics are examined by a box plot graph. This first phase ends with an evaluation of the number of observations selected and whether they exhibit non-random patterns.

Then next step is the factor analysis which allows us to identify correlations at the level of groups of variables as it represent a multi-variate extension to the bi-variate correlation analysis. Factor analysis is applied for identifying latent variables which are variables able to attract clusters of variables providing an explanation for correlations of the variables within the cluster. The multidimensional scaling method is applied to confirm (or discharge) the results of the factor analysis. If the results are confirmed, variables from the administrative source can be assessed as they have passed the internal coherence test. If the results do not hold up, an alternative to multidimensional scaling is to partition the data randomly into two subsets of equal dimension and perform a factor analysis for exploratory scope on a subset and a factor analysis for confirmatory scope on the second subset.

As an example of the application of the procedure 3, we can describe the case of total costs paid by freelances of Sector Studies. Since no control variable neither direct nor functional has been found this can be treated as a type D case. Here we search for internal coherence by exploring the relationship with other theoretically-connectible variables from the same source, explicitly from the Sector Studies accounting box variables. Comparison between standard deviation and *mad* directly points out the presence of outliers. Hisq and Disq Statistics displayed in a box plot graph are used to identify outliers which show non-random patterns. As far as factor analysis is concerned, the seven variables loading on the four factors is very clear since there is no variable charging on 2 or more factors together: the first axis (factor) indicates a link between four cost variables which are correlated, while the other three axes do nothing but represent, each, a variable cost unrelated to the others. The multidimensional scaling method confirms the factor analysis results. We can conclude that the variable from the administrative source can be assessed as internally coherent from a statistical point of view.

Conclusions

As we know, administrative definitions are in some cases different from statistical ones, and this leads to the difficulty to reconciling them. By providing a standardized method of reconciliation, we can move forward integration in a statistical register. The twofold consequence is to either put the variable into the register as it is, or look for a method to understand and integrate a new variable from the administrative source. The first solution has the advantage of having immediate available data but could make it more difficult the subsequent interpretation of functional relations between this variable and the others already existent. The second is more burdensome but allow easier examination of the functional relationships among variables. In any case it should be checked if there is any similar variable that measures the same phenomenon as the variable in question in statistical sources.

The need to define a standardized way to compare administrative variables with statistical ones provides methods differentiated by type and degree of similarity. In short, every variable of the administrative source can be split into two parts: the matched units and the unmatched ones. If the matched part of the source would detect coherence with the statistical benchmark, we can say that the administrative variable is relevant. If the unmatched cases have features similar to the matched part, we can integrate the variable into a statistical register.

The Sector Studies source as a very powerful source of information was used to develop such a general method for assessing administrative source. The results obtained have been used to validate Sector Studies source to be used, not only in Administrative sources and registers Department, but also in other Divisions that have requested Sector Studies data. This is the example of National Account Division and the Structural Business Statistics Division. In the second case the data were used to write a description of SME enterprises into the 2008 Annual Report and now they have been used to experiment the integration of SME sample survey data together with other administrative sources.

References

A. Agresti, *Categorical Data Analysis*. Wiley 2002

P.A Allison. *Logistic Regression Using Sas*. Cary, NC: Sas Institute, 1999.

A. Basilevsky. *Statistical Factor Analysis and Related Methods*. Wiley, 1993.

A. Bernardi, F. Cerroni, V. De Giorgi *A Methodological Process for Assessing Variables coming from Administrative Sources: an Application to the Tax Authority Source (Sector Studies)* European Conference on Quality in Official Statistics Q2008. Roma, 9-11 July 2008.

A.C. Cameron, P.K. Trivedi. *Microeconomics using Stata*. Stata Press, 2009

F. Cerroni, V. De Giorgi *The Tax Authority Source as an example of the use of an administrative source as a statistical one*. IAOS2008. Shanghai, 14-16 October 2008.

P.J.H. Daas, S. J.L Ossen, J. Arends-Tóth. *Framework of Quality Assurance for Administrative Data.* Paper for the 57th Session of the ISI, Durban, South Africa, 16-22th August 2009.

ESC. Pros and cons for using administrati ve records in statistical bureaus. Paper presented at the seminar on increasing the efficiency and productivity of statistical offices, Economic and Social Council conference of European statisticians, Geneva, Switserland, 2007.

B. Everitt, S. Landau, M. Leese. *Cluster Analysis.* A Hodder Arnold Publication, 4th edition, 2001.

L. C. Hamilton. *Statistics with Stata.* Brooks/Cole, Cengage Learning, 2009

R. Khattre, D. N. Naik. *Multivariate Data Reduction using Sas*. Crystal Dreams Publishing, 2000, p. 300.

SAS. On-line Documentation. *Sas/Stat module*, *example 28.2*, 2009

D. Piccolo. *Statistica.* Il Mulino, 2000.

A. Wallgren, B. Wallgren. *Register-based Statistics : Administrative Data for Statistical Purposes.* Wiley. March 2007.

A. Zuliani, C. Scala. *Complementi di Statistica Metodologica*. Ed. Kappa, 1964. p. 264.