Fiscal data supporting Research and Development statistics

Orietta Luzi, Giulio Perani, Giovanni Seri
Italian National Statistical Institute (Istat)
Direction for Structural business statistics, Agriculture, Foreign trade and Consumer prices
{luzi,perani,seri}@istat.it

Introduction

The Italian annual Business Research and Development (R&D) survey produces statistics on the R&D expenditure by business enterprises, as well as on personnel involved in R&D activities. It is a census survey where the target population includes all Italian enterprises potentially performing R&D in the reference year.

In order to improve the process of identification of potential R&D performers, Istat is implementing a process of integration between statistical and administrative (fiscal) data. Data on tax incentives for R&D (both tax credits and allowances) at enterprise level are provided to Istat by the Italian Tax Agency (Agenzia delle Entrate). The main problems related to the use of tax credits data to complement the statistical production in the R&D field will be discussed, including the discrepancies between Istat and Tax Agency in the definition of R&D or the negative effects of frequent changes in the tax legislation.

The use of fiscal data to improve the statistical production on R&D will be discussed with to reference to:

1. editing/validation of the business R&D survey's results;
2. to deal with the business R&D survey's total and partial non-response;
3. to keep updated the business R&D survey's target population list;
4. to support the development of a sample R&D survey on small and medium enterprises by providing the information needed to identify a target population of marginal (unknown) R&D performers.

1. The Italian business R&D survey

An obvious matching can be proposed between the fiscal data collected by the Italian Tax Agency on the expenditure for R&D activities by the Italian enterprises and statistical data annually produced by Istat on the R&D expenditure and personnel of Italian enterprises.

Istat is collecting business R&D data, and producing official statistics on the Italian business R&D activities, since 1963. The Istat business R&D survey, which is carried out every year, is based on the methodological recommendations provided by the "Frascati Manual", the main source of theoretical and practical guidelines to undertake statistical R&D surveying at international level[1].

The Istat R&D survey is an annual census of the population of Italian business enterprises which can be identified as "potential R&D performers" in the reference year. In order to identify the target population, Istat is constantly collecting information from a range of statistical and administrative sources to keep updated a "register of R&D performers". Such sources of information include: previous Istat R&D surveys, other Istat business surveys with R&D-related questions, the Istat official statistical business register, the Italian Register of R&D performing institutions, managed by the Ministry of University and Research, data on national and EU funding to research projects, patent databases, business reports.

Following the Frascati Manual recommendation, on the basis of such register of the R&D-performing enterprises, all enterprises known, or assumed, as being R&D performers are included in R&D survey. Since the year 2004, also fiscal data from the Italian Tax Agency have been used to identify R&D performers.

---

[1] The "Frascati Manual. Proposed Standard Practice for Surveys on Research and Experimental Development" is an OECD technical manual, firstly published in 1963, aimed at providing the OECD countries with common methodologies to estimate their own investments on R&D. The Manual, now available in its sixth edition, is currently used all over the world and is considered – according to the EU legislation – as the main methodological source for the production of official R&D statistics in the European Union as well.

As to the size of the surveyed enterprises, it has to be pointed out that there is no a cut-off point: all enterprises which are potential R&D performers are included in the target population irrespective of their number of employees. No special treatment is implemented for data collected from micro-enterprises (0 to 9 employees) if they are employing at least one researcher[2]. However, for selected industries and technological areas, also micro-enterprises with less than one researcher are included in the realised sample.

Up to 27,000 enterprises are currently under monitoring as "potential R&D performers" in Italy. Most of these enterprises are constantly surveyed to identify the "actual R&D performers". Besides this census of potential performers – and in order to provide a highly consistent information on the time series of R&D expenditure and R&D personnel - data from a panel of around 200 largest performers (covering about 85 per cent of total R&D spending) are regularly collected. The rate of response for this panel is assumed to be very close to 100 per cent. In the event that an enterprise included in the panel should not fill in the questionnaire, data from the previous year's questionnaire is imputed after careful checking and up-dating.

Istat is not performing any grossing up of the R&D survey results. The treatment of total non-response is only based on "cold-deck imputation" from previous R&D or other business surveys.

So far, administrative data have never been used for imputation of missing data but only for the identification of potential R&D performers. The increasing availability of fiscal data (including tax credits data) will give Istat the opportunity of using them also in the process of data imputation and estimation.

2. R&D tax credits in Italy

Since the fiscal year 2007, Italian businesses have been allowed to apply for a corporate tax relief proportional to their own total R&D expenditure. The R&D incentive scheme developed by the Italian Government was based on tax credits with two different reduction rates: 10 per cent for investments in own R&D projects and 40 per cent[3] for investments in R&D projects developed in co-operation with universities or public R&D institutes. A maximum of 50 million Euros of R&D expenditure can be declared by a single enterprise for each fiscal year until 2013[4].

The Italian R&D tax credits scheme is based on the automatic subtraction from the income corporate tax (or, as an alternative, from other corporate fees including the "regional tax on business activities", IRAP) of a percentage of the enterprises' expenditure on R&D. Tax credits can be used both in the same fiscal year when the R&D expenditures have been undertaken, and in the next fiscal years. Because of the concerns about the costs for the Italian Government budget of such scheme, the enterprises have now to inform the Tax Agency, some months before filing the yearly tax statement, about the expected volume of their R&D expenditure in the current year in order to make the Tax Agency able to monitor the total future costs – in terms of loss of revenues – for the public budget[5]. As a result, two datasets will be made available every year by the Tax Agency to Istat for statistical use: a dataset containing data on the expected R&D expenditures and a dataset on actual R&D expenditures as reported in the enterprises' tax statements.

The definitions used by the Italian legislators in establishing such R&D tax reduction scheme follow more the European Commission (EC) standards[6] than those recommended by the OECD's Frascati Manual (FM). Because of the discrepancies between these two institutional frameworks in defining the R&D, it was even questioned whether the basic R&D (according to the FM definition) should have been included in the activities supported by this legislation[7]. In order to provide a clarification, the Tax Agency accepted to link as follows – although not very properly – the EC definitions on R&D with the FM concepts:

<p align="center">

Fundamental research   <==>   Basic research

Industrial research   <==>   Applied research

Experimental development (EC)   <==>   Experimental development (FM)

</p>

---

[2] Usually expressed in "full time equivalent" (FTE) as recommended by the Frascati Manual.
[3] The initial rate was 15 per cent, later (2008) increased to 40 per cent.
[4] This incentive scheme has been recenty extended to 2013 but it was initially limited to the period 2007 - 2009.
[5] A maximum amount of revenue losses have been set by the Government: for instance, in the fiscal year 2010 will be granted R&D tax credits for no more than 654 million Euros.
[6] European Commission, "Community Framework for State Aid for Research And Development and Innovation", (2006/C 323/01), Brussels 2006.
[7] The guidelines by the Tax Agency explicitly included "basic research" among the activites to be reported by enterprises to get tax credits but the legislation (Law No.296/2006) was not so clear on this point.

Discussing the differences between these two sets of definitions goes beyond the scope of this paper. What is relevant here it is to assess whether the use of the EC definitions in the fiscal legislation could make the resulting administrative data not reliable for statistical use. In this case, we will accept that the Tax Agency's interpretation had provided a basis for comparing tax credits and statistical data.

About the information to be provided by the applicants, it should be pointed out that, for the very first time, a detailed breakdown of the businesses' R&D expenditure (mainly by "type of cost") is requested by the Tax Agency. The tax statement has to be filled in with information on the following items (as well as on the share of them related to R&D co-operative projects).

- **Total R&D expenditure** (arguably, both intra and extra-mural) – Tax Agency code: RS15
- R&D personnel expenditure (only researchers and technicians are explicitly mentioned) – code: RS01
- Expenditure for laboratories' equipment and tools – code: RS03
- Acquisition (renting) of buildings to be used as research centres – code: RS05
- Research contracted-out and patent-related costs (to be possibly excluded from the intra-mural R&D expenditure) – code: RS07
- Expenditure for R&D consultancy – code: RS09
- Administrative and general costs – code: RS11
- Materials and consumables costs– code: RS13

In principle, by excluding the item coded RS07 which could largely include extra-mural R&D expenditures and innovation costs, the expenditure data collected by the Tax Agency could allow for a reliable estimation of the total R&D intra-mural expenditure by the Italian businesses. Indeed, a few caveats have to be stressed:

- each enterprise can apply for tax reduction up to a maximum of 500,000 Euros, so it can not report yearly more than 50 million Euros of R&D expenditure;

- as the burden of a detailed book-keeping of R&D costs may prevent small enterprises by applying for tax credits, so the uncertainty about the availability of sufficient public resources to have their applications accepted may prevent large enterprises (mainly those listed in the stock exchange) to apply, as well.

As a conclusion, tax credits data can be very helpful to complement the Istat statistical activity in the field of R&D but both the coverage of this administrative source with reference to the overall business R&D activities in Italy, and the consistency with the R&D statistics regularly produced by Istat on the basis of its R&D surveys, have to be carefully checked.


3. Matching statistical and tax credits data

In this paragraph some preliminary statistical analyses aiming at evaluate the consistency of the 2007 tax credits data with R&D survey data are illustrated. These analyses have been performed on a set of enterprises with both tax credits and survey data available for the reference year of 2007. Since the 2007 Istat R&D survey is still in progress, to perform the analysis we used data from questionnaires completed and checked by the end of July 2009. Therefore, the results presented here have to be considered as provisional.

A common identification enterprise code allowed for matching the fiscal and the statistical datasets. Around 40 per cent of the R&D survey respondents were also found in the tax credits file. Records presenting inconsistencies have been removed. As a first step, the relationships between the corresponding variables in the two datasets have been investigated. In Table 1 (columns 1 and 2) the correspondence between fiscal items (as described above) and R&D survey variables is shown. Since we can not assume that a one-to-one relationship would exists between the two sets of variables for all the enterprises, in some cases different relationships have been tested.

In Figure 1 the values of the survey's variable on total intra-mural R&D expenditure (code: C312) is plotted against the corresponding variable in the tax credits dataset (code: RS15) (logarithmic scale). The hypothesis of a (log)linear relationship between them seems to be graphically supported. The correlation (considering all records) between the variables C312 and RS15 is 0.7 (see Table 1, column 3), however several isolated points in the graph suggest the presence of a few measurement errors in the data (e.g. unity measure errors). These cases have been clerically revised, and some further records have been removed from the matched dataset, leading the size of this dataset from 1,799 to 1,655 records. After the clerical review of the data, the correlation between the variables described in columns 1 and 2 of Table 1 increased remarkably (Table 1, column 4).

Figure 1. Plot of variables C312 and RS15 (logarithmic scale)
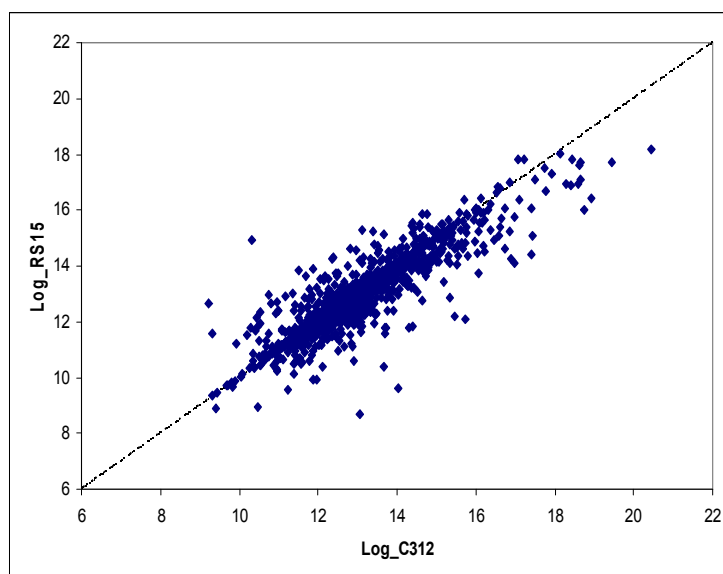
In Table 1, further results are shown. In order to perform a preliminary analysis of the capability of tax credits data to predict the R&D survey non responses, a 10 per cent of total non responses on the R&D survey data has been assumed (corresponding to 165 records), based on a Missing Completely At Random (MCAR) mechanism (Little et al., 1987). Then, the auxiliary information provided by the tax credits data have been used in a simple scheme of robust regression model (without residual error) for predicting the simulated missing values. In particular, the Least Trimmed Squares (LTS) algorithm (Rousseew at al., 1987) has been used for obtaining robust estimates of the parameters of the regression models with the R&D survey variables (Table 1, column 1) as dependent variables and the tax credits items (Table 1, column 2) as explanatory variables[8]. In the column 5 of the Table 1 the R-Square of the models are reported.

Table 1. Correspondence between R&D survey variables and tax credits items

1	2	3	4	5	6	7	8
R&D survey variables (descriptions and codes)	Tax credits items (codes, see above)	Corr-1 1,799 rks	Corr-2 1,655 rks	R-Square	Diff% (Sum)	Outliers	N_Obs
Total R&D intra-mural expenditure (code:C312)	RS15	0.70	0.89	0.966	11.5	0	1490
Total R&D intra-mural expenditure (code:C312)	RS15-RS07	0.73	0.94	0.965	12.6	9	1490
R&D personnel costs (code: C302)	RS01	0.75	0.90	0.978	17.1	48	1487
R&D personnel costs excluding Other R&D personnel, i.e. only Researchers and Technicians (code: C303+C304)	RS01	0.69	0.88	0.908	4.9	25	1433
Expenditure for R&D equipment (code: C310)	RS03	0.30	0.65	0.005	-204.7	0	441
Expenditure for land and buildings to be used for R&D (code: C309)	RS05	0.12	0.20	0.004	-8701.4	0	39
Total capital R&D expenditure (code: C308)	RS03+RS05	0.35	0.77	0.01	-233.4	0	482
Total R&D intra-mural expenditure (code:C409)	RS07	0.19	0.65	FAILURE			304
Costs of R&D consultants (code: C307)	RS09	0.04	0.18	0.834	10.3	103	734
Total R&D current costs, excluding R&D personnel (code: C306)	RS11	0.65	0.56	0.155	-11.2	74	1111
Total R&D current costs, excluding R&D personnel (code: C306)	RS13	0.24	0.56	0.051	-379.3	0	1111
Total R&D current costs, excluding R&D personnel and consultants (code: C306-C307)	RS11+RS13	0.40	0.63	0.751	-20.0	39	836
Total R&D current costs, excluding R&D personnel (code: C306)	RS09+RS11+RS13	0.43	0.91	0.804	-36.0	37	1111

The model-based predicted values were used to replace the simulated missing information (only non-zero values for the dependent variables which have been preliminarily selected). The differences (as percentages) between the sum of the concerned variables (column 1) for the 165 observed records and the sum of the same

---

[8] The LTS algorithm has been applied by using the Proc ROBREG of the SAS package.

variables computed on the corresponding predicted values, are reported in column 6. The number of outliers detected by the procedure for each variable can be found in column 7 and the number of observations used in the model (as above, only non-zero values for the dependent variables) is shown in column 8.

The above results seem to be promising for an extensive use of tax credits data for statistical purposes. However, since not all the results are as expected, further investigations are required. Additionally, as analyses have been conducted so far only on provisional data, the results of the above exercises will have to be confirmed with reference to the final 2007 R&D survey data..

4. Conclusions: future use of tax credits data in the production of R&D statistics

After having checked for the consistency between data on R&D tax credits and survey data, a potential use of fiscal data in the production process of R&D statistics can be envisaged. On the other hand, it has to be stressed that the availability of administrative data is strongly affected by policy choices and that a certain level of "instability" of this source of information (with reference to data availability, coverage and, even, financial sustainability) has to be taken into consideration.

The most obvious use of R&D tax credits data in connection with the R&D surveys will be that of keeping updated the R&D survey's target population list. In Italy, as in other countries, a range of different administrative and statistical sources is used to identify the enterprises belonging to the population of "potential" R&D performers to be regularly surveyed in order to get detailed information on their R&D activities. In this respect, tax credits data offer a highly valuable input under the assumption that all the R&D performers would find an advantage in applying for tax reduction. Istat has included the fiscal source among those used to update its "register" of R&D performers since 2004, when some information on R&D performers were made available for the first time by the Tax Agency.

Administrative data, notably R&D tax credits data, can be extensively used to edit/validate the results of R&D surveys. Enterprises are reporting that the process of quantification of the resources spent on R&D in a given period of time is much more accurate when carried out for fiscal rather than for statistical purposes.

Finally, R&D tax credits data can provide additional information to deal with the R&D surveys' total and partial non-response. Imputation of missing data can be largely based on administrative sources but under some conditions. First, a full knowledge of the administrative source is needed, including detailed information on possible bias in terms of coverage or data reporting. This is needed in order to avoid the use of partial data (for instance, an enterprise could not be either willing or allowed to report all its R&D expenditure). Second, a suitable model has to be develop in producing (robust) estimates for those survey variables which are not available from the administrative source. Estimated data should not bring any distortion, for instance at regional or sectoral level, to the aggregated estimates as a result of possible under or over-estimation of the variables of interest. Third, it can be argued that – since only a few variables are usually available - not all the total non-responses are worth of being imputed by using administrative data. The use of administrative data has to be chosen when the potential R&D performer would have an expected level of R&D expenditure (or number of persons employed in R&D activities) remarkably high and relevant to produce accurate results for some statistical domains.

As to the Istat future plans on the exploitation of R&D tax credits data, it can be mentioned the project (currently under development) of using such administrative data to design a sample R&D survey on small and medium enterprises (SMEs). Tax credits data should provide a quite detailed information to be used to identify a target population of marginal R&D performers not included in the current survey. In addition to this, the availability of data on the R&D investments of a large number of SMEs could allow for calculating the probability an enterprises can have undertaken R&D in a given reference year, as well as, to predict the size of its R&D expenditure. Such analyses will be developed further by Istat as longer time series will be available (only the year 2007 has been processed, so far) and also the longitudinal dimension of the R&D activity could be taken into consideration.

Selected bibliography

European Commission (2006), *Community Framework for State Aid for Research and Development and Innovation*, (2006/C 323/01), OJ 30/12/2006, Brussels.

Istat (2008), *La Ricerca e Sviluppo in Italia nel 2006*, Statistiche in breve, Rome.

Little, R. and D. Rubin (1987), *Statistical Analysis with Missing Data*. Wiley & Sons, New York.

OECD (2002), *Frascati Manual 2002: Proposed Standard Practice for Surveys on Research and Experimental Development*, Paris 2002.

Rousseew P.J., Leroy A.M. (1987). *Robust Regression and Outlier Detection*. Wiley & Sons, New York.