POSSIBLE IMPROVEMENTS IN THE ESTIMATION OF IT-LFS THROUGH INTEGRATION WITH THE REGISTERS SYSTEM

Antonella Iorio, Alessandro Martini

Abstract. In recent years, due to the increasing difficulties in maintaining high response rates and ensuring good representativeness of the collected samples, quality problems in surveys are getting more frequent. The impact on the final estimates is in terms of bias and lack of consistency between the different estimates. In recent decades, many NSIs have started to use administrative data, usually processed and organized in statistical registers, integrating them into statistical processes, with the aim of improving the quality of the output.
Availability of data coming from registers is increasing in the last years in Italy and it will improve even more in the next future.
The Registers should be exploited in statistical processes and in different phases: as frames for sample selections, to draw auxiliary variables needed in the estimation and to enrich the set of available information, through record linkage.
In this paper we focus on the evaluation of possible improvements on the accuracy in the IT-LFS exploiting the integration of auxiliary information coming from the ISTAT system of statistical registers. The Registers System (SIM) will be able to provide information about sex, age, citizenship, education level, employment signals and household structure for the resident population.
This more reliable information could be used through calibration in both the phases of estimation: treatment of total no response and final calibration.
For the no response treatment it would be possible to define constraints in a more flexible framework while, in the final calibration step, additional information about education level and employment signals coming from the register could be useful to improve different aspects of quality of the LFS estimates.
This experimental study is relevant to evaluate the possible improvements of the estimates and to collect evidences for introducing these changes in the next future.
The results show that with the proposed innovations the survey problems (MRT, bias) are addressed by improving the quality of the estimates.

1. Background

The Italian LFS is one of the most important source of statistical data referred to the Italian labour market. It provides monthly, quarterly and yearly figures of the main aggregates and, in a longitudinal perspective, flow estimates as well. The IT-

LFS is designed as a quarterly survey, the sample design is a two stage stratified sample with rotational pattern 2-2-2. The sample is uniformly spread across all weeks such that all geographical domains are represented in each month and in each of the four waves (rotational groups).

The estimation methodology in IT-LFS is design based, demographic administrative data are exploited through the calibration estimator.

Consistency of estimates produced by different sources is a key quality dimension for NSIs, so methodological approach and estimation procedures in IT-LFS take into account the need to provide consistent estimates for the different indicators in the statistical domain of labour market. More specifically, ISTAT is making an effort to improve coherency between LFS and different other surveys (Multipurpose Annual Survey, Time Use Survey, IT-Silc) and also with Continuous Population Census. "Coherence by design" is the idea of ensuring consistency trough the harmonization of sampling frames, designs, estimation methods and data collection protocols.

In this paper we focus on the evaluation of possible improvements on the accuracy of the estimates produced by IT-LFS exploiting the integration of auxiliary information coming from the ISTAT system of administrative and statistical registers (SIM).

The procedure is based on three step, and final weights are computed as follows:

- Step 1: Initial weights are obtained as the inverse of the inclusion probabilities of any selected household.
- Step 2: Intermediate weights are computed multiplying initial weights by correction factors for unit non-response worked out as the inverse of the response ratios; correction factors for non-response treatment are calculated using the information from the LFS theoretical samples to define 13 household typologies, considering the age, gender, citizenship of the head of the household and the number of members. When we started this kind of correction for no response, this information was not available in the population frame, but only for the municipalities selected in the sample.
- Step 3: Starting from intermediate weights, final weights are obtained solving a minimization problem under constraints through a calibration model. The estimates of some auxiliary variables have to be equal to the totals in the reference population. Final weights ensure that all members of a given household have the same weight.

Grossing weights are computed on quarterly basis, whereas annual estimates are calculated as averages of quarterly estimates.

The IT-LFS weighting procedure is more complicated than required by EU regulations, to satisfy specific national needs, in particular the weighting scheme includes, at NUTSII level:

- distribution of population by sex and seventeen 5-year age groups;

- distribution of non-national population divided as follow: male, females, other EU citizens, Non-EU citizens;
- number of households for each rotation group (1/4 of the total);
- distribution of population by sex, for each of the three months of the quarter.

The weighting scheme includes too:

- distribution of population by sex and five age groups for the thirteen large municipalities with more than 250.000 inhabitants;
- distribution of population by sex and five age groups at NUTSIII level.

According to European Regulations these control totals applied in calibration derive from demographic administrative sources. This auxiliary information is incorporated in the estimation procedure with the calibration estimator (Deville & Särndal, 1992; Särndal, 2007) to improve the quality of the estimates. As matter of fact that quarterly weights already include the benchmark to the NUTS 3 population, the annual datasets are obtained using all the quarterly interviews and annual weights are computed simply dividing the quarterly weights by four.

2. Calibration on control totals

Deville and Särndal (1992) formalized calibration in survey sampling according to the basic idea that if auxiliary variables strongly correlated with the target variables are available it "means that the weights that perform well for the auxiliary variables also should perform well for the study variables".

Consider a finite population $U = \{1, . , k, , N\}$. A sample s of fixed size n is drawn from population U according to a sampling design $(S, p(.))$, where S is the sample space and $p(\cdot)$ is a probability distribution on S. The first order inclusion probability, $\pi_k = Pr(k \in s)$, and the second order inclusion probability, $\pi_{kl} = Pr(k, l \in s)$, are assumed to be known and strictly positive. Throughout, let $\pi_{kk} = \pi_k$.

Let us assume to be interested in estimating the total of Y variable $t_y = \sum_{k \in U} y_k$ and to have collected on each element $k \in s$, besides the value of the interest variables y_k, an auxiliary vector value, $x_k = (x_{k1} \dots x_{kp} \dots x_{kP})$. Furthermore, population totals of Xs, $t_{x_p} = \sum_{k \in U} x_{k_p}$ with $p = 1, \dots, P$, are accurately known, without sampling error. Solving the optimization problem:

$$\begin{cases} min_{w_k} \{\sum_{k \in s} G(w_k, d_k) \,/\, q_k\} \\ \sum_{k \in s} w_k x_k = t_x \end{cases}$$

Deville and Särndal demonstrate that we can find a system of weights, w_k, calibrated (coherent) with the known population totals and the more the auxiliary variables are correlated with the target variables, the more efficiency of the estimates improves.

Furthermore, the calibrated weights are as close as possible to the design weights, $d_k = 1/\pi_k$, with respect to a given metric $G(\cdot)$.

When $G(\cdot)$ is the chi-squared distance, the resulting calibration estimator is equal to the generalized regression (GREG) estimator (see Cassel et al., 1979; Särndal, 1980; Isaki and Fuller, 1982; Wright, 1983; Bethlehem and Keller, 1987; Särndal et al., 1989; Fuller, 2002).

Even using other metrics with respect to the chi-squared distance (see Deville and Särndal, 1992 pp. 378-379) the calibrated estimator (CAL) is asymptotically equivalent to the GREG estimator. Therefore, also the asymptotic variance of CAL estimator can be derived referring to the GREG estimator.

3. Micro-linkage of administrative information and LFS data

Availability of data coming from administrative and statistical registers has increased significantly in recent years in Italy, and its quality and timeliness is expected to improve further over the next few years. These new sources open up the possibility of integrating such kind of data into some of the LFS production processes, usually as auxiliary information – used at the micro or macro level - for sample design, sample selection, correction for non-response, weighting, validation, imputation and modelled estimates. When administrative data are integrated into the processes at the macro level, usually there are no particular issues, apart from the consistency of definitions and the timeliness, i.e. the lag between their reference period and the moment they are available to LFS specialists for processing.

When administrative data are integrated into the processes at the micro level, a further problem arises and is related to the possibility for LFS specialists (or other dedicated specialists within statistical offices) to have an anonymized unique identifier (SIM code) that allow to link individual records (persons and households) of the administrative/statistical databases to those of the LFS databases (persons and households selected, whether interviewed or not). The process of assigning the anonymized unique identifier is called in Istat "pseudonymization". It is a crucial task performed by a specialized unit within Istat, following stringent criteria and procedures set by The Italian Data Protection Authority.

This "pseudonymization" process has to be carried out regularly, both on the new administrative databases made available to Istat and on the actual LFS samples. This process will assign the unique identifiers to persons and households that were already

"pseudonymized" and will create a new unique identifiers to persons and households that enter for the first time into the population or sample (e.g. new born, immigrants). Clearly, the integration at the micro level of up-to-date administrative data and LFS microdata during LFS processing would help to implement much more effective methods for the treatment of total non-response and under-coverage, thus improve the quality of monthly and quarterly LFS estimates that need to be disseminated at t + 30 days (for monthly data) and t + 56 days (for quarterly estimates and for transmitting validated data to Eurostat).

Until a few months ago, the main obstacle to such integration was represented by the timing and methods of "pseudonymization" of the actual sample of the LFS survey. While the unique identifier was available for the theoretical sample already during sample selection, given the complexity of the procedure it was used to be attached to all respondents of the actual LFS sample only on an annual basis, at the end of the year, for the purpose of estimating the informal employment for National Accounts.

In order to be able to regularly integrate administrative data for the treatment of total non-response in the LFS, a new "pseudonymization" process was developed - jointly by LFS specialists and the dedicated Istat unit - specifically for such purpose. The new process is semi-automated, based on a probabilistic record linkage methodology, and runs on daily bases on the new interviews transmitted by the enumerators. For individuals who are not linked through the probabilistic approach, a deterministic linkage procedure is used. It's important to note that this semi-automated daily process excludes visual checks, but these further quality checks are performed manually only at the end of the quarter to finalize the process.

This new process successfully assigns a unique code to over 98% of cases, thereby enabling the exploitation of register information for the estimation of both monthly and quarterly LFS data.

4. Use of calibration and administrative information for Non-response treatment

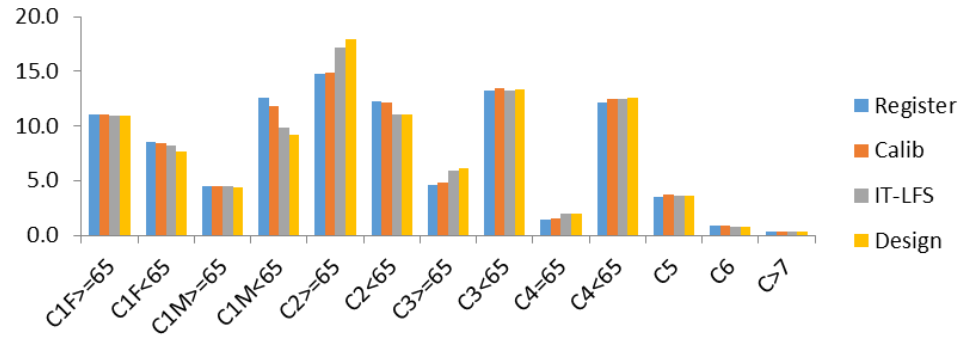
The availability to link the actual sample to the register data made it possible to review the non-response treatment. As seen above correction factors for non-response treatment are calculated with reference to the estimates obtained from LFS theoretical sample, through an Iterative proportional fitting procedure.

The availability of the population register allows to derive directly the totals for the distribution of the households by the specific typologies considered and geographical domains. Moreover, the use of calibration allows to define a more flexible framework to avoid very small adjustment cells, organizing constraints in a

more flexible framework (for instance a more detailed classification of households typologies at NUTS2 domain, a less detailed classifications at NUTS3 domain and strata level).

The results in Figure 1 show clearly how treatment of non-response trough calibration and administrative information is more effective in the context of the IT-LFS (blue and orange bars). The procedure we are currently applying is not able to entirely correct the overestimation, or underestimation, of certain categories of household, in particular the household of 2 members with the head of the household aged 65+ (C2>=65), we have an estimate of 17.9% with the base design weight (yellow bar) the intermediate weight of LFS provides an estimate of 17.2% (grey bar) while in the frame we observe 14.8% (blue bar). On the other side, households consisting of a single male member aged 18-64 (C1M<65) are 9.9% (grey bar), according to non-response adjusted weight, but they are 12.6% in the population register (blue bar). This is mainly due to the fact that the in the procedure we are currently using the correction factors, at stratum level, cannot exceed the correction factor calculated at province (Nuts3 level). This was a reasonable criterion and ensured a suitable correction when the procedure was defined but nowadays the profile of non-response by region and type of municipality has deeply changed and these bounds at province level do not allow an effective treatment of such different non-response patterns. In addition, the totals directly derived from the population register are more up-to-date, likely more reliable, and definitely more precise than the estimates we can derive from the theoretical sample.

Figure 1 – *Households by number of members and age of the head of the household, IT-LFS 2021Q4.*



Source: ISTAT, Labour Force Survey

5. Additional benchmarks derived from administrative and statistical registers

The Covid emergency had a big impact on the data collection process of IT-LFS. Since during lock-down all the interviews were conducted by CATI and the group of households that was entering the sample for the first time, was replaced selecting households that had already participated to the LFS in previous quarters and had provided a phone number for telephone interviews. So possible sources of bias arose in the data due to:

- selection of all households having phone number available;
- higher substitution rate (replacement of non-respondent households);
- higher final non-response rate.

The bias was studied comparing LFS data collected in 2020Q2 with previous quarters and with administrative sources. IT-LFS figures showed higher frequencies for elderly people, Italians, people having higher education level and employed ones. In those quarters were introduced in the calibration additional constraints regarding the distribution by education level at NUTS2 level in order to reduce the bias, maintaining the same structure of the population by education level. Updating these constraints was an issue since we had no data available for the definition of the totals.

On January 2021 new population figures were made available for the period 2011-2021, according to the results of the 2018 Population continuous Census. Consequently, LFS weights were recalculated to be coherent with IT Census population, starting from 2021Q1 and the constraints about education level have not been used anymore.

However, given the important relation noticed between the sampling distribution of the educational levels and the key labour market indicators, a further study has been conducted to evaluate the effect of the integration of auxiliary information coming from the registers on the IT-LFS estimates. This experimental study is relevant to evaluate the possible improvements of the estimates and to collect evidences for introducing these changes in the next future.

For this purpose, an additional set of constraints has been added to the final step of calibration, deriving them from the education level in the population register (RBI) and the signals of regular employment in the Labour Register (RTL/BOP). Both the registers are available with a variable timing compared to the production calendar of the LFS, a new population register (RBI) for year Y-1, referring to 31th december, is available for the estimation of quarter YQ4 while RTL/BOP has almost the same timing but with a provisional version. The same information is available for the actual sample, at microdata level, and the total population.

The first set of 24 constraints include education level in 4 groups (ISCED 0-1, 2, 3-4, 5-Higher) by sex and by 3 age groups (15-29, 30-49, 50+).

The second set of constraints include two different option to consider this auxiliary information:

- total quarterly signals of regular employment for 15-64 individuals by sex;
- annual mean of signals of regular employment for 15-64 individuals by sex.

Thus, starting from the intermediate quarterly weights, we used a calibration approach, with several different sets of constraints to get the new final weights. In this paper we present results for some of them, pointing out that all the constraints have been defined at NUTS 2 level, the level of partitioning of the calibration model.

6. Application and results

We compared the results of these different calibration models for quarter 2021Q4 LFS with the results of the census 2021. In the census, employment status at time t for unit k is modelled as a binary latent variable (t, k) (employed or not). Census survey, LFS and administrative sources are treated as imperfect measures of the target process. Coherence of employment estimates are open issue in our country, in particular for the census, that releases its figures 12 months after the reference period, with much less timeliness than then LFS.

Table 1 – *Calibration models considered.*

Calibration model	No-Response Adjustment	Final Calibration
IT-LFS	Current Iterative Proportional Fitting	Current Calibration applied in the survey
ED_FR	Current Iterative Proportional Fitting	IT-LFS + Education level in the Population register
ES_FR	Current Iterative Proportional Fitting	IT-LFS + Education level in the Population register + Regular Employment signals in Labour Register
PBCAL_ES	First step Calibration	IT-LFS + Education level in the Population register + Regular Employment signals in Labour Register

Comparing IT-LFS and the whole census framework (List-sample, Area sample, Register) LFS shows a higher level of individuals with university degrees (and a lower percentage of individuals with education up to a high school degree),

considering the education level collected on the register, in addition LFS has a higher share of respondents with no employment signal in the reference week.

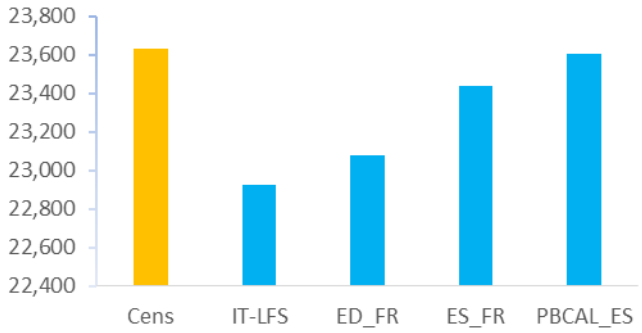
Given an estimate of employed people in the LFS that is lower of 3.0% than the census, at national level, the greatest differences are concentrated in the southern regions, in particular in Campania, Puglia, Calabria and Sicily. Overall, in these 4 regions, the Census estimates a higher number of employed people around 600 thousand (the differences exceed +10%).

Looking at Figure 2 we can summarize the results as follow. The current employment estimate from the LFS is 22.923 thousands (IT-LFS), about 3% lower than the Census figures (Cens). Introducing constraints in the calibration derived from RBI employment estimates, LFS employment estimate increases to 23 thousands including the education level (ED_FR).

Introducing the constraints on regular employment signals in the final weighting step, keeping the current adjustment for the total non-response increases the employment estimate to 23.437k (ES_FR).

Recalculating both the intermediate weight with a calibration step, as described in paragraph 4, and including in addition to the education level the constraints on regular employment signals in the final weighting step (PBCAL_ES) we get 23.603 thousand employed.

Figure 2 – *Total Employment Census 2021, LFS 2021Q4 with different weights.*



Source: ISTAT, Census, Labour Force Survey.

For employment rate 15+ at national level we get similar results, coherence with census improves, the last estimate with 2-step calibration almost closes the gap with the census figure. However, Nuts2 level analysis points out clearly a regional pattern (Figure 2). Including the auxiliary information of education level and, afterwards, of regular employment signals, the employment rate, based on the ILO definition, increases, in particular in the southern regions. In Campania employment rate raises

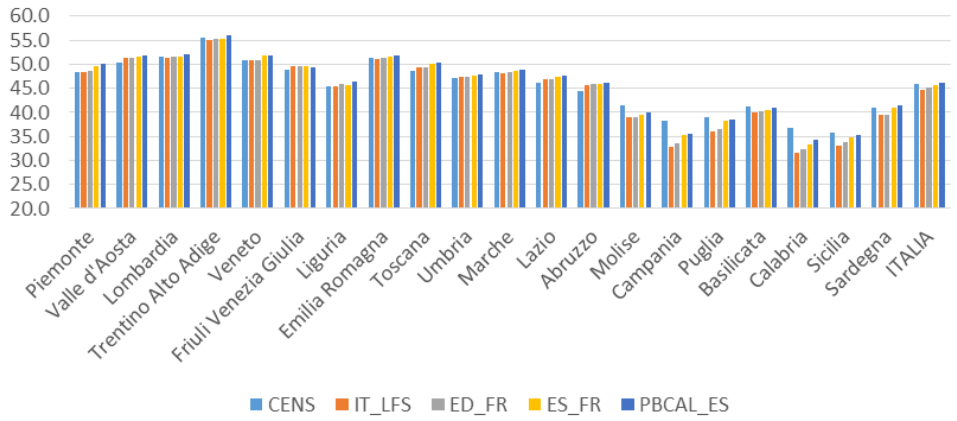
from 32.8 (the IT-LFS figure) to 33.5% including education level, 35.3% with regular employment signals and to 35.6% with 2-step calibration. Almost the same happens in Calabria, Sicilia and Puglia.

These first results represent further insights to study the differences between LFS and Census in Italy, an open issue that requires to be studied more in depth. For sure the response propensity of respondents is different for the two surveys and the under/over coverage as well.

Historically, the census is able to reach individuals who are very mobile in the country, perhaps a large part of them is residing in the south of the country but lives and works in the northern regions. Participation in the census is a strong motivation, linked to the administrative aspects of maintaining residence in a particular municipality, and the availability of CAWI interviewing mode for the Census may have been relevant for this particular subpopulation in order to improve the response rate.

On the other hand, participation of these individuals to the LFS could be more difficult, given that the survey technique is CAPI-CATI and they spend most of their time away from their residence, hence giving rise to a certain under representation in the sample. However, an advantage of the LFS, compared to census, is that data collection lasts less, so telescoping effect and measurement errors are generally reduced. The integration of the different sources and the use as auxiliary information in the LFS calibration seems to be able to reduce the bias due to the unbalanced coverage of certain sub-groups of population.

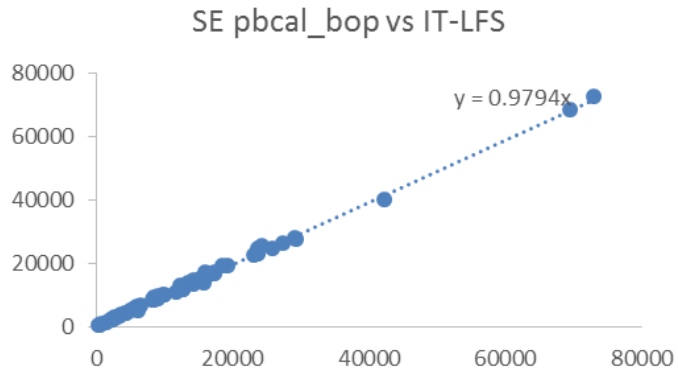
Figure 3 – *Employment rate 15+ at Nuts2 level, Census 2021, LFS 2021Q4 with different weights.*



Source: ISTAT, Census, Labour Force Survey.

Comparison of standard errors for a set of estimates obtained with 2-step calibration (PBCAL_ES) and the current calibration model (Figure 3) shows a slight increase for the precision of the estimates with the first option ($\hat{\beta}$=0.979).

Figure 4 – *Standard errors for the main aggregates at NUTS2 level, 2021Q4 LFS estimates and 2-step weights.*



Source: ISTAT, Census, Labour Force Survey.

7. Conclusions

The results show that the proposed innovations can be effective to address the survey problems (MRT, bias) and for improving the quality of the estimates. The crucial element is the integration with the information available from the registers. It is reasonable to think that introducing the same auxiliary information, education level and signals of regular employment, already included in the census latent class model, in the calibration model of LFS, leads to a better coherence of the estimates. For this purpose, calibration, beside the efficiency gain, is very effective to build a coherent system of surveys. Considering the different timeliness of the two surveys even a benchmarking of census employment to the LFS figure could be a suitable strategy, once the differences have been clearly identified.

The joint use of these kinds of auxiliary information in calibration estimators raises some questions.

Integrating administrative data into the LFS improves the accuracy of monthly and quarterly estimates? This should also be assessed by considering more detailed territorial domains, estimates of variations and particular subpopulations such as irregular workers.

Can we consider the integration of this information a good practice?

Which kind of signal for regular employment should we consider? The quarterly or the annual one?

Do they have a significant effect on the seasonal pattern of ILO Employment?

Further studies will be useful to finely tune the model and focus on specific aspects, but this work has already shown that the integration of administrative sources and its use in the calibration model is a very promising way to improve quality of IT-LFS.

References

DEVER J.A., VALLIANT R. 2010. A comparison of variance estimators for poststratification to estimated control totals, *Survey Methodology*, Statistics Canada, Catalogue No. 12-001-X, Vol. 36, No. 1, pp. 45-56.

DEVILLE J.C., SÄRNDAL C.E. 1992. Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, Vol.87, No. 418, pp.376-382.

DI ZIO M. FILIPPONI D. 2023., Multi-source statistics in the Italian permanent census . In *Proceedings of the Workshop on Methodologies for official statistics*, Rome, pp.31-44.

FULLER W. A. 2002. Regression estimation for survey samples, *Survey Methodology*, Statistics Canada, Catalogue No. 12001XIE, Vol. 28,No. 1, pp.523.

FULLER W.A., RAO J.N.K. 2001. A Regression Composite Estimator with Application to the Labour Force Survey, *Survey Methodology*, Statistics Canada, Catalogue No. 12001, Vol. 27, No. 1, pp. 45-51.

LUNDSTROM S., SÄRNDAL C.E. 1999. Calibration as a Standard Method for Treatment of Nonresponse, *Journal of Official Statistics*, Vol. 15, No. 2, pp. 305-327.

RENSSEN R.H., NIEUWENBROEK N.J. 1997. Aligning estimates for common variables in two or more sample surveys, *Journal of the American Statistical Association*, Vol 92, Issue 437, pp. 368-374.

Antonella IORIO, ISTAT, iorio@istat.it
Alessandro MARTINI, ISTAT, alemartini@istat.it