Italian Time Series Back Recalculation

With the introduction of the new Regulation, starting from January 2021, differences in the definition of the employment status have been determined, thus leading to the necessity of a reconstruction of the previously released time series, in order to make the oldest data directly comparable with the newly produced ones. The process of back-recalculation was developed upon *macro-aggregates*, defined not only to allow the supplies to Eurostat, but also to provide suitable information to the various users (internal and external to ISTAT) who may need to use it.

In order to apply the established back-recalculation technique, it is essential to have a period of overlap between the old and the new survey, so that a model for linking the two segments of data can be defined. The overlapping period length between the two surveys must be at least one year in order to calculate a reconstruction of the series that also takes into account seasonal effects.

A macro-level, model-based and component-based approach was adopted to reconcile the series:

- macro level, because it is not possible to reconstruct the new estimates by aggregating individual data due to the absence in the microdata of the information needed to apply the new definitions;
- model based because it uses econometric and statistical techniques to analyse time series;
- component based because it separately reconstructs the long-term component, the annual period component and the short-term component, traditionally the so-called trend-cycle, seasonality and erratic components.

In view of the new Regulation, starting from January 2018 a number of additional questions were introduced into the questionnaire in order to simulate, even in the 'old' questionnaire, the new definition of employment status, although not yet operational. This made it possible to have as many as 36 overlap points (one for each month of the three-year period 2018-2020) for which estimates with respect to both the old and the new definition were available.

We used a multiple linear regression model to estimate the series levels according to the new definition at the unobserved points (the months from January 2004 to December 2017), while the series levels for the months between January 2018 and December 2020 were obtained by recalculation upon microdata.

Initially, the back-recalculation of the data from 2004 to 2017 was produced by status, gender and 5 age groups (15-24, 25-34, 35-49, 50-64, 65+), for a total of 30 series. The dependent variable for the multiple linear regression model (formula 1) is the amount of individuals classified according to the employment status of the new Regulation (CONDNR, where NR stands for New Regulation). As regards the independent variables, in addition to the level of the same series according to the old Regulation (CONDVR), seasonality and individual 'entries' or 'exits' from different employment status were kept separate. As well known, the substantial differences between the old and the new definition of employment status concern the different way of considering those who are absent from work for more than three months. Unlike the past, for example, individuals with lay-off allowances for more than three months represent an outgoing group, while workers on parental leave for more than three months represent an incoming one.

To summarise, the used independent variables are:

the same seasonally adjusted series according to the old definition (net of the amount of the outgoing groups, if used as explanatory variables) as X_1 ;
their seasonality (obtained as the difference between the unadjusted and the seasonally adjusted data) as X_2 ;
the number of individuals switching from one condition to another due to the change of definition, because lay-off allowances, as X_3, and reduced activity or leave, as X_4.

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \qquad (1)$$

Table 1 shows the details of the model for each of the 30 series initially back-recalculated, highlighting the explanatory variables considered and those left out being not significant.

Table 1. Series, explanatory variables used in the models and R^2 values.

Y	X_1	X_2	X_3	X_4	R^2
111	yes	yes	yes		0.99999
112	yes	yes	yes		1.00000
113	yes	yes	yes		1.00000
114	yes	yes	yes		1.00000
115	yes	yes			0.99995
121	yes	yes	yes	yes	0.99999
122	yes	yes	yes	yes	1.00000
123	yes	yes	yes	yes	1.00000
124	yes	yes	yes		1.00000
125	yes	yes			0.99991
211	yes	yes	yes		1.00000
212	yes	yes	yes		0.99998
213	yes	yes	yes		0.99998
214	yes	yes	yes		0.99999
215	yes	yes			1.00000
221	yes	yes	yes		0.99998
222	yes	yes	yes		0.99999
223	yes	yes	yes		0.99999
224	yes	yes	yes		0.99998
225	yes	yes			0.99999
311	yes	yes	yes		1.00000
312	yes	yes	yes		0.99998
313	yes	yes	yes		0.99996
314	yes	yes	yes		0.99998
315	yes	yes			1.00000
321	yes	yes	yes		1.00000
322	yes	yes	yes		1.00000
323	yes	yes	yes		0.99999
324	yes	yes	yes		1.00000
325	yes	yes			1.00000

SERIES NAME KEY *ijk*: individuals in employment status *i* (1=employed, 2=unemployed, 3=inactives) according to the new Regulation, sex *j* (1=males, 2=females), age group *k* (1=15-24, 2= 25-34, 3=35-49, 4=50-64, 5=65+)

In particular, the intercept is never significant, the parental leave is an explanatory variable only for employed women under 50 years of age and the lay-off allowances for the over-65s is never considered. The goodness of the models chosen is reflected in the R^2 values, which are always close to 1.

Having chosen a component approach it was necessary a careful evaluation of the choices made in the seasonal adjustment of the series. The presence of seasonality was tested also among the outgoing and incoming groups aside (in order to establish whether it should be considered as an additional independent variable), but it was absent; moreover, in compliance with international standards for assessing the quality of seasonal adjustment, the opportunity to adopt an indirect approach to back-recalculation was considered. To this end, models were estimated for people over 15 years old as a whole: they gave very similar results (without being better) than those obtained with the indirect estimation, finally used. This choice stems from the ability of indirect estimation to give more detail to the final result. Those series represent the benchmark for the subsequent breakdowns, which mainly concerned the employed people.

The strategy adopted for the disaggregation of the first 30 back-recalculated series is that historically adopted in the reconstructions of data before 2004; in particular, each time series is gradually disaggregated in several series, distinguished by the chosen detail of the newly added variable of interest, while maintaining consistency with the higher levels of aggregation.

For each employment status, the detail by age group and geographical breakdown is expanded, while employed people are distinguished in self-employed, permanent and temporary employees.

In particular, 15-24 age class was divided into two classes (15-19 and 20-24) to meet the requirements of the Regulation, while the last age group (65 or more) was split into three classes (65-74, 75-89 and 90 or more) to meet the requirements of the new definition. While the unemployed people remain limited to age 74 as in the old definition (and therefore the estimate of the unemployed aged 65 or more years in practice already represents only 65-74 years old), in the new definition, the employed people are limited to age 89 and people aged 90 year or more must be considered as inactive.

The technique used for these breakdowns is a generalisation of the statistical methodology for indirect estimation of small areas by analysing larger areas containing them, and more precisely of the SPREE (Structure Preserving Estimation, Rao, 2000) method. This approach uses information related to the variable to be estimated and splits the population into so-called crossclasses. The choice of these variables obviously plays a key role in the estimation procedure. In this work, the series disaggregation is obtained by associating all the age classes with structures related to the new variables determined on larger age classes containing them[1].

This allows the production of estimates for the whole population by employment status, gender, geographical breakdown and 8 age groups (15-19, 20-24, 25-34, 35-49, 50-64, 65-74, 75-89, 90+); for the employed also by position in the occupation (employees or self-employed) and nature of employment (permanent and temporary). This detailed information makes it possible to meet the requirement of the Regulation not only with respect to the 14 mandatory series (employed and unemployed by gender and age groups 15-24, 25-64, 65-74 plus 20-64 for employed persons only)

[1] The 8 age groups of interest, are disaggregated using the 3-mode structure 15-34, 35-49 e 50+

but also providing 20 other series among the optional ones, concerning employees, self-employed and employed on temporary contracts.

In order to ensure consistency between the total of the series and the known population totals by sex and age, we applied a reconciliation procedure by re-proportioning.

What illustrated so far refers only to the period between 2004 and 2017. As a matter of fact, the levels of the series for the period 2018-2020 have been obtained by recalculation of microdata, given the availability of the CONDNR, directly with the maximum detail. By construction, this branch of the series did not need the reconciliation step to the population totals, as it was entirely derived from weighted microdata and therefore already made consistent with the population totals during the process of estimation of the survey's weights.

The new intercensual population estimates are available for Italy, which took into account the latest available census data and implicated the revision of the entire population data series from 2011 to the present. In order to make the back-recalculated labour market data also consistent with the new populations, the reconciliation procedure by re-proportioning was reapplied for the period from 2004 to 2017, while for the 2018-2020 period the micro-data were used again, for which in the meantime the survey's weights have been recalculated using the new populations.