Methodological note

GUIDANCE ON ESTIMATION AND IMPUTATION OF MISSING DATA FOR SHORT-TERM BUSINESS STATISTICS IN THE CONTEXT OF THE COVID-19 CRISIS

EUROSTAT, DIRECTORATE B
UNIT B1 – METHODOLOGY; INNOVATION IN OFFICIAL STATISTICS

EUROSTAT, DIRECTORATE G
UNIT G3 – BUSINESS CYCLE; SHORT-TERM STATISTICS

20 APRIL 2020

Background

As a consequence of the COVID-19 outbreak, many enterprises (temporarily) closed down, work at lower intensity or partially changed their economic activity since February or March 2020. This concerns most countries of the world, where a variety of lock-down restriction measures have been adopted.

For a number of countries, data collections have been disrupted or is severely hampered, because the statistical office partially stopped working, many companies are closed down, or the collection of administrative data has been delayed. These factors will lead to under-coverage, breaks in series and comparability issues due to mode effects.

Eurostat issues this note, focusing in particular on short term business statistics, to provide guidance to the NSIs to guarantee - as far as possible - a harmonised approach on how to deal with the high number of missing data or delays, as well as to ensure sound imputation techniques.

Impact of the COVID-19 crisis on the business population

Depending on the extent of the confinement measures taken in the respective EU countries, all or part of the kind of activity units (KAUs) and enterprises (later called “statistical units”) with certain activities have ceased to function. These statistical units can be partially identified based on their NACE activity code, by administrative information, i.e. enterprises registering for temporary unemployment of their staff.

To increase the identification of the active business population that can still be surveyed, additional indicators on the activity status of enterprises or legal units could be added to the national business registers. The inactivity of statistical units should be taken into account when estimating or imputing the economic development as well as the date when they became inactive.

Impact of the COVID-19 crisis on surveys / administrative data

Despite the difficulties, the best possible efforts should be made to gather survey responses from businesses. In this respect, the communication side toward the observation units (enterprises) is very important, for example alerts on the websites of the statistical offices, alerts in the reporting tools for businesses:

- NL has posted a notice on their website
- FR took a similar initiative

In the current circumstances, businesses may not be able to provide the (administrative) data they are legally required to provide. In some countries, administrations have allowed additional time for providing administrative data.

In case a full "normal" survey is not possible, simplified questionnaires can be proposed, suitable to the current situation.

Methods for estimating and imputing missing data for short term business statistics

OPTIONS TO REPLACE MISSING SURVEY DATA

Whenever possible, Member States are encouraged to replace missing survey data with ***data from administrative sources***:

- VAT declarations
- e-invoicing data
- other monthly or more frequent tax data
- social security declarations.

Member States are also encouraged to use ***non-traditional sources*** of information where applicable, such as:

- data from the press
- data from the internet (social network sites, webscraping)
- payment statistics
- credit card transactions
- business and consumer surveys
- energy consumption data
- information from professional federations.

ESTIMATION AND IMPUTATION METHODS: OPTIONS WITH AND WITHOUT ADMINISTRATIVE DATA

It is important to differentiate between missing data coming from *unit non-response* or *item non-response*. Although unit non-response could, in principle, also be treated with imputation, weighting methods are more commonly used such as:

- adjusting the weights
- resampling and using secondary sampling unit

In the current specific situation, the non-response of a statistical unit might more often than in the regular situation be an indication of (temporary) ceased activity. Consequently, the imputation of an earlier value (not adjusted by the number of confinement days) may lead to an over-estimation of the target variable in the case of STS flow variables (turnover, production, hours worked and wages and salaries).

For adjusting the weights it is important to have information on the activity status, because this allows better estimation. The activity status might be obtainable using alternative sources (e.g. phone call, internet search). More detailed methods for adjustment for non-response is available in the Handbook of Theory and Research.

In short-term business statistics (STS), the indices are either nominal or fixed base year Laspeyres or Paasche indices (current base year 2015). Some countries also compile chain-linked indices with annual change of weights, in particular for the prices. Given that the purpose of the STS is to reflect the changes with comparison to the base year, it is not advised to change the overall structure of the weights caused by the COVID-19 crisis.

If for some reason an index is not available for one of the lower level activities, the weight of that activity should be distributed proportionately amongst the other activities that also contribute to the same activity one level higher in the activity classification.

For example, if there is no index for Class 10.73, the weight of Class 10.73 should be distributed between Classes 10.71 and 10.72, not simply by assigning half of the weight to each of these two Classes, but by dividing the weight of Class 10.73 according to the relative weights of Classes 10.71 and 10.72. The index for Group 10.7 is then compiled from the adjusted weights of Classes 10.71 and 10.72([1]). An analogue approach could also be used for re-distributing the sample weights of the missing statistical units in sample surveys.

Imputation is used more for item non-response, where there is existing partial information about the units, which can be extended by imputation to improve the quality of the statistics.

The **Memobust handbook** lists the following cases and methods for imputation:

- deductive imputation
- model-based imputation
- parametric
- non-parametric
- donor imputation
- cold deck

[1] Methodology term business statistics - Interpretation and guidelines, p. 23

- random hot deck
- sequential hot deck
- nearest-neighbour
- predictive mean matching
- longitudinal imputation (panel data)
- last information carry forward
- interpolation
- mean/ratio imputation
- nearest-neighbour/Little-Su method

The **Guidelines on the use of estimation methods for the integration of administrative sources** provides recommendations on how to use the above listed methods, if administrative data sources are available.

POSSIBLE ALTERNATIVES

Based on the above mentioned references, **deductive imputation is recommended** as the preferred method over all other imputation methods, as the imputed values will automatically satisfy certain restrictions or edit rules. If there is no unique value provided and multiple values need to be imputed for a given unit, then a donor imputation technique (nearest neighbour, random hot deck, sequential hot deck, or predictive mean matching) can be applied.

As in many cases the effect of the COVID-19 crisis can be different on enterprises depending on the industry characteristics, for modelling the non-parametric approach should be considered as it does not require the explicit usage of a model. Non-parametric models are more flexible in handling complex situations (a lot of variables of mixed type, categorical and continuous), but their draw back is that they are computational intensive. The imputed values from models and donor imputation do not always account for edit rules. Therefore, it is not guaranteed that the imputations made by those methods will satisfy all the restrictions.

There are two ways to solve this problem. The first option is a two-step approach, which is more commonly used in practice and easier to apply. In the first step, the missing values are imputed using an appropriate imputation method, which does not take (all) edit rules into account. Then, in a second step, the imputed values are minimally adjusted to satisfy the edit rules, according to some minimisation criterion. The second option tries to take the edit rules into account in the imputation method itself, for instance by choosing an appropriate model which generates imputations that automatically satisfy the edit rules. A drawback of this approach in practice is, that it can easily lead to very complex imputation models. However, there are some exceptions where direct modelling of the edit rules is feasible.

The cold deck donor imputation, **model based imputation** using parametric models and **imputation using longitudinal data** can have lower performance under the current circumstances, as the COVID-19 shock can change the parameters of the estimated models significantly and the previous values can differ largely to the values of the current period. In cases where detailed information are not available for a specific sector, imputation of the values on basis of the over-arching categories is possible. These imputations are considered to be an **acceptable** solution.

Carry-forward, **ARIMA** forecasts and **naive modelling** are going to be useful to estimate benchmarking values to be used for the estimation of the impact of COVID-19. **For sectors where significant impact is expected due to the COVID-19 measures, they are not recommended.**

However the **carry-forward** model **can be recommended** for data where the impact of COVID-19 cannot be measured or estimated, e.g. for prices([2]) of sectors which fully stopped their activities in the reporting period.

R PACKAGES SUPPORTING IMPUTATION

The statistical production can be assisted by existing **R packages**:

- VIM: Visualization and Imputation of Missing Values, package created in the ESS, colleagues from Statistics Austria ensure the maintenance
- Amelia: A Program for Missing Data, tool developed for imputing missing data based on bootstrapping based algorithm
- mice: Multivariate Imputation by Chained Equations, package for more advanced users

OTHER CONSIDERATIONS

The usual **procedures for non-response** correction or imputation might not be valid under the current conditions. These procedures usually assume continuity and will generate results that will be biased toward the “normal”, which might not describe the current situation. The composition of non-response might be very different in this period. Companies might be (temporarily) out of business and the questionnaires do not seem to apply to them. The best solution would be to design a non-response survey or conduct desk research on non-respondents. Also knowledge on what activities are forbidden in what periods can be used to make reasonable guesses.

To **minimise non-response**, efforts in CATI/CAWI should be increased, and in case respondents can be contacted, a simplified questionnaire shall be used to lower the burden. Best practices for the fieldwork are welcome, as well as for alternative sourcing such as for webscraping and machine learning.

The data compilers should critically review their regular **automatic editing procedures** because they might generate a bias. First of all, because due to the special situation, records can be identified as errors, where they are actually not. And secondly, because the automatic correction is likely to be biased towards continuity.

Some statistical units temporarily have taken up **alternative activities**, e.g. producing disinfectants, protective masks, offering home delivery services. It is not recommended to reclassify them at this point since these will be only temporary activities and might not affect the business structure in the long term. Although information on temporary conversions of businesses will be interesting, STS is not the right statistical domain for this purpose.

In the period of the crisis **wages and salaries([3]) might be partly subsidized by governments**. According to the definitions applied in STS, all wages and salaries paid by employers should be included and **such subsidies should not be netted out**. The countries

([2]) As regards prices Eurostat published on its webpage COVID-19 Support to statisticians a note: Guidance on the compilation of the HICP in the context of the Covid-19 crisis

([3]) As regards wage subsidies Eurostat published on its webpage COVID-19 Support to statisticians a Draft note on statistical implications of some policy measures in the context of the COVID-19 should report as “wages and salaries” only those payments that the statistical units make to employees, and not measures, addressed by government directly or indirectly to the employees.

Subsidies to production **should not be reported as turnover** of the statistical units.

Preference shall be given to the **sectoral approach** whenever possible as some branches are obviously impacted in some positive or negative way. NSIs could be looking at the respective national legislation for lockdown, where the list of activities subject to lockdown are explicitly indicated. This could be used for flagging sectors with potentially inactivity status in the business register.

Dissemination of statistics using estimation and imputation of missing data

Information on the adopted solution should be provided as complementary information. Such **metadata** could take into account the specific measures for each Member State, for example was there a general lockdown in the country, or only partial for some of the regions or sectors. From which date until when were enterprises forced to lock-down? Provide metadata on reliability of the values following the current **guidelines for quality reporting**.

The measures related to the virus will have an impact on the accuracy, on the timeliness or on the completeness of data. This is somehow a trade-off that requires strategic decisions. In any case, there has to be **some communication to the users, preferably coordinated at ESS level,** and at least there should be an exchange of information of the National Statistical Authorities on the communication approach and on the priorities.

When disseminating the data, **proper flags** for imputed or estimated values should be used (see **guidelines how to use flags** and guidelines on **STS data transmissions**). The currently applied flags of the STS domain seem to be sufficient for flagging the data:

- A - normal value, produced with usual procedures
- E - estimated value, based on limited amount of data, applying additional calculations
- P - provisional value, the source considers that the data are expected to be revised

For data transmissions marked with E and P flags as well as in case of missing data NSIs should inform the Eurostat STS domain managers on the reasons. Eurostat recalls that missing data should not be transmitted as “NaN” (not-a-number) in the STS data transmission files – instead, the time series should end with the last available observation. Eurostat will develop a nomenclature to standardise the future communication. Flags for confidential data can only be used for protecting data which would allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information.

When determining the strategy for STS, the **overall coherence with other domains** should be taken into account (i.e. IPI vs. GDP, etc). Relevant in this regard is the consideration that certain sectors, for example the production sector is locally more affected in a given country.

Further information and support

- ESTAT-STS-DATA@ec.europa.eu
- ESTAT-methodology@ec.europa.eu
- ESS Vision 2020 ADMIN online remote help desk