EUROPEAN COMMISSION
EUROSTAT

Directorate F: Social statistics
Unit F-2: Population and migration

Luxembourg, 20 April 2020
ESTAT/F-2/GL*

TECHNICAL NOTE

Subject: correction of provisional nowcast data

1. INTRODUCTION

Eurostat regularly collects provisional data referring to the first six months of the reference year within the month of November of the same year. This data collection is known as NOWCAST, shorter version of Nowcasting Data Collection. Therefore, the latest available nowcast data are those referring to (at least) the first six months of the year 2019 and they were collected in November 2019. Vital events (i.e., live births and deaths) are regularly provided with a monthly disaggregation, while migration data (that are not mandatory) are less available and sometimes provided with different temporal disaggregation (e.g., 6-month aggregate).

In particular occasions, such as the projections exercises, these data may be further provided by the countries at a later date. Sometimes only the revised aggregated value is transmitted by then. Even though usually covering a larger period of the reference year, these data may be of varied nature: for instance, they may be still provisional or based on national definitions not complying with EU regulations. For any further use of these nowcast data it is thus advisable to take into account their provisional and/or national features. This note indicates basic methods to correct such potential biases.

2. METHODS OF CORRECTION OF NOWCAST DATA

2.1. Method based on previous NOWCAST

This correction is based on a measure of the data changes occurred in previous data collections. Considering the NOWCAST in the reporting year *t*-1 those monthly data – referring to the same year *t*-1 – are transmitted to Eurostat again the following reporting year *t* in the Unified Demographic Data Collection (UNIDEMO).

* Giampaolo Lanzieri (giampaolo.lanzieri@ec.europa.eu). This note has not been officially edited and it is released to inform interested parties about ongoing work. The information and views set out in this note are those of the author and do not necessarily reflect the official opinion of the European Commission / Eurostat.

The ratio of these quantities gives a correction factor that can be applied on the nowcast data reported in the year *t*:

$$k_p^1 = \frac{d_{U,t,p}}{d_{N,t-1,p}} \quad [1]$$

where $d_{N,t-1,p}$ are any nowcast data (births or deaths, possibly also immigration and emigration) in the reporting year *t* -1 covering the period *p* (e.g., 6 months), and $d_{U,t,p}$ are the corresponding monthly data from the UNIDEMO data collection in the following reporting year.

An example of application on the latest available data collections is reported in the Table 1, where the values of the factor k_p^1 are reported together with the number of months covered by the nowcast data, which varies from country to country. In no case the value of the factor is equal to one, which means that all nowcast data would require a correction. Among vital events, such a correction is on average slightly more important for deaths, as indicated by the Mean Absolute Error (MAE) at the bottom of the table. From the same Table 1 it can be noted that not all countries provide data on migration flows and even in that case, it would not be possible to compute the k_p^1 due to the lack of data on monthly migration in the UNIDEMO data collection. Therefore, for those countries that provide a partial coverage of the migration in the nowcast data, only the period *p* is reported.

The correction factor could also be an average over several years of the [1]:

$$\bar{k}_p^1 = \frac{1}{T}\sum_{t=t_0}^{t_n} k_{t,p}^1 = \frac{1}{(t_n - t_0 + 1)}\sum_{t=t_0}^{t_n} \frac{d_{U,t,p}}{d_{N,t-1,p}} \quad [2]$$

where t_0 and t_n are respectively the first and last reporting years of the UNIDEMO used for averaging the correction factor. For instance, considering the reporting years 2000-2002, it is:

$$\bar{k}_p^1 = \frac{1}{(2002 - 2000 + 1)}\sum_{t=t_0}^{t_n} \left(\frac{d_{U,2000,p}}{d_{N,1999,p}} + \frac{d_{U,2001,p}}{d_{N,2000,p}} + \frac{d_{U,2002,p}}{d_{N,2001,p}}\right) \quad [3]$$

Using a single year should be sufficient in regular situations; extending the average over too many years backward brings the risk of including in the correction factor data production peculiarities not anymore in place.

This method is best applied in times close to the data transmission of NOWCAST data. When these data are revised in later data transmissions, the correction factor in [1] or [2] might not be the best solution, as the latest data may be closer to the final values and the factor would then over-correct the remaining bias – if any. Additionally, migration data are usually less covered in the NOWCAST data collection. A better solution for these cases can then be provided by the next method.

2.2. Method based on DEMOBAL

When nowcast data are revised by the data provider, sometimes even in the following year, a better picture of the potential biases can be derived by the changes measured

between DEMOBAL and the UNIDEMO data collections. The Demographic Balance Data Collection (DEMOBAL) is carried out within six months of the end of the reference year and it contains the main aggregates for births, deaths and – more frequently, even though not mandatory – immigration and emigration. The UNIDEMO, carried out within 12 months of the end of the reference year, is meant to collect the final version of data. The DEMOBAL data are supposedly closer to the final version than the NOWCAST data and in some cases there are no further changes in between DEMOBAL and UNIDEMO data.

Likewise in the previous method, a correction factor can be so derived:

$$k^2 = \frac{d_{U,t}}{d_{D,t}} \qquad [4]$$

where $d_{D,t}$ are the data usually from the DEMOBAL of the same reporting year of the nowcast data under correction. The data from DEMOBAL are better taken from their first version transmitted to Eurostat, as they might as well be revised and getting closer to the final values. The factor k^2 does not have a subscript referring to the time period covered by the data because in both DEMOBAL and UNIDEMO data cover the entire (previous) year, thus it holds $p \equiv 12$. The k^2 can as well be computed in its averaged version:

$$\bar{k}^2 = \frac{1}{T}\sum_{t=t_0}^{t_n} k_t^2 = \frac{1}{(t_n - t_0 + 1)}\sum_{t=t_0}^{t_n} \frac{d_{U,t}}{d_{D,t}} \qquad [5]$$

The country-specific values of k^2 have been computed from the latest data collections and they are reported in the Table 2. Values in green cells are exactly equal to 1 and therefore indicate that no correction would be needed. The Mean Absolute Error (MAE) at the bottom of the table shows that emigration data would be those potentially more biased, but the value of the MAE is there nuanced by the presence of an outlier (which should be further investigated) and two missing values. Nowcast migration data result in general more prone to biases also due to differences in definition (e.g., migrants with stay of at least one year vs. those with stay of four months). Last but not least, the size of corrections is generally lower than in the previous method, which is coherent with the expectation that the accuracy of the data improves over time.

3. CONCLUSIVE REMARKS

The two methods presented above are two very basic tools for the correction of potential biases in nowcast date, either later revised or not. The choice between the two methods can be made looking at the time of the last version of nowcast data under examination: when closer to their regular deadline, then k_p^1 should be preferred; when closer instead to the first sending of the following DEMOBAL data collection, it is k^2 that could apply the most appropriate correction. This latter factor may be the only option when dealing with migration data, given their reduced availability in regular NOWCAST data collections and the lack of monthly migration data in UNIDEMO.

Let illustrate the approach above by an example. The nowcast data transmitted in November 2019 to Eurostat contain monthly vital events and migration data for the first months of 2019; they are soon followed in December 2019 by the UNIDEMO data that

contain the entire set of monthly vital events data and the whole migration data for the year 2018. For nowcast data transmitted in November 2019 only, the suggested method for correction makes use of the factor k_p^1, computed using monthly data from that latest UNIDEMO compared to the previous nowcast reporting, both these latter referring to the year 2018. If those nowcast data are revised at a later moment, possibly closer to the deadline of the DEMOBAL 2020 data collection, then it is the factor k^2 that should be preferred. For instance, nowcast data reported in November 2019 but revised in March-April 2020 have supposedly a quality closer to DEMOBAL data than to the original NOWCAST data, which is the rationale for the application of k^2.

Relatively high values of any of the two correction factors are warning signals and they should be followed by further investigation to understand the specific reason(s). Data declared by the country as 'final' despite of their early delivery[1] should not be subject to corrections.

[1] Countries may have the total values of the demographic balance components for a given year *t* relatively early in the following year and anyway well in advance of their transmission to Eurostat. Taking advantage of such early availability, nowcast data may be revised to improve their completeness and accuracy. This is for instance the case in the occasion of projections exercises, when it is beneficial to incorporate the latest empirical evidences.

Table 1: values of k^1 and corresponding number of months covered as from reporting years 2018 (NOWCAST) and 2019 (UNIDEMO)

country	births	*p*	deaths	*p*	immigration	*p*	emigration	*p*
BE	0.975	6	0.977	6	NA	NA	NA	NA
BG	0.990	12	1.006	12	NA	NA	NA	NA
CZ	1.014	6	1.006	6	NA	6	NA	6
DK	1.000	9	1.001	9	NA	9	NA	9
DE	1.008	7	1.003	7	NA	6	NA	6
EE	1.006	10	0.999	10	NA	NA	NA	NA
IE	0.958	6	0.980	6	NA	6	NA	6
EL	1.003	8	1.000	8	NA	NA	NA	NA
ES	1.009	6	1.006	6	NA	6	NA	6
FR	1.000	6	0.999	6	NA	NA	NA	NA
HR	1.002	9	0.984	9	NA	NA	NA	NA
IT	0.993	7	1.002	7	NA	7	NA	7
CY	1.012	6	1.144	6	NA	NA	NA	NA
LV	1.007	10	0.999	10	NA	10	NA	10
LT	0.973	12	0.997	12	0.781	12	0.768	12
LU	1.028	8	1.125	8	NA	NA	NA	NA
HU	1.002	12	1.010	12	NA	NA	NA	NA
MT	1.008	12	1.014	12	1.553	12	1.452	12
NL	1.003	9	1.001	9	NA	9	NA	9
AT	1.006	9	1.014	9	NA	6	NA	6
PL	1.002	9	1.001	9	NA	NA	NA	NA
PT	1.002	9	1.001	9	NA	NA	NA	NA
RO	1.084	12	1.018	12	NA	NA	NA	NA
SI	1.006	6	1.001	6	NA	6	NA	6
SK	1.007	9	1.032	9	NA	9	NA	9
FI	0.995	12	1.011	12	0.981	12	1.222	12
SE	1.105	9	0.998	9	NA	9	NA	9
IS	0.995	11	0.983	11	NA	11	NA	11
LI	1.052	11	1.085	11	NA	NA	NA	NA
NO	1.000	9	1.019	9	NA	9	NA	9
CH	1.049	8	1.023	8	NA	NA	NA	NA
MAE / average *p*	**0.017**	**9**	**0.020**	**9**	**0.263**	**9**	**0.302**	**9**

Table 2: values of k^2 as from reporting year 2019 (DEMOBAL and UNIDEMO)

country	births	deaths	immigration	emigration
BE	1.000	1.000	1.025	1.093
BG	1.000	1.000	1.000	1.000
CZ	1.000	1.000	1.133	1.370
DK	1.000	1.000	0.741	0.880
DE	1.000	1.000	0.566	0.453
EE	1.000	1.000	1.000	1.000
IE	1.001	0.997	1.000	1.000
EL	1.000	1.000	NA	NA
ES	1.009	1.004	1.001	1.001
FR	1.001	0.993	NA	NA
HR	1.000	1.000	1.000	1.000
IT	1.000	1.000	1.000	1.000
CY	1.000	1.000	1.000	1.000
LV	1.000	1.000	1.000	1.000
LT	1.000	1.000	1.000	1.000
LU	1.000	1.000	1.000	1.000
HU	1.000	1.000	1.000	1.000
MT	1.000	1.000	1.000	1.000
NL	1.004	1.001	0.802	0.706
AT	1.000	1.000	0.719	0.603
PL	1.000	1.000	2.769	3.441
PT	1.000	1.000	1.000	1.000
RO	1.079	1.003	0.952	0.970
SI	1.005	1.001	0.993	0.991
SK	1.000	1.000	1.000	1.000
FI	1.000	1.000	1.000	1.000
SE	1.000	1.000	1.000	1.000
IS	1.000	1.000	0.829	0.566
LI	0.997	1.000	1.005	0.994
NO	1.000	1.000	0.912	0.790
CH	1.030	1.007	0.844	1.032
MAE	**0.004**	**0.001**	**0.123**	**0.172**