Distr.: General
07 November 2019

English only

Economic Commission for Europe

Conference of European Statisticians

Work Session on Demographic Projections

Belgrade, 25–27 November 2019
Item No. 11 of the provisional agenda
Population projections at sub-national level

Ex-post Analysis of Sub-national Population Projection Errors for Ukraine and Data Quality Problems

Note by Ptoukha Institute for Demography and Social Studies of the National Academy of Sciences of Ukraine*

Summary

The last sub-national population projection for Ukraine was developed in 2009. The absence of more recent projection is mainly due to a long period passed since the last census which took place in December 2001. Since then current population estimates are based on the vital statistics and migration data. This situation may accumulate considerable inaccuracies mainly due to migration registration. Incompleteness of migration registration also leads to some bias in vital statistics data. If a woman gave birth in a region other than she registered (but actually lives) this birth can be ascribed to region of occurrence. This results in some numerator-denominator inconsistency in the fertility rates at the regional level. Thereby migratory attractive regions have overestimated fertility rates and vice versa. After Ukraine lost control on the Crimea and parts of Donetsk and Luhansk regions in 2014 the vital statistics data quality became worse.

The population projection 2009 revision was calculated for all regions of Ukraine: 24 oblasts, Autonomous Republic of Crimea and two cities Kyiv and Sevastopol. The method for estimating future population numbers is cohort component. The starting point of this projection is 01 Jan 2009 and the end is 01 Jan 2036. The discrepancy between sum of regions and the whole country was distributed proportionally by region, sex and 1-year age groups. Open age interval is 100 years. For calculation mortality and fertility to be consistent to Ukraine was used Brass models. Mortality and fertility rates of Ukraine were taken as a standard.

In this research is offered ex-post errors analysis of regional population projection 2009 revision in comparison to current estimations. The biggest errors account for children 0–4 years old due to fertility hypothesis errors and older 90 due to small numbers problem.

*Prepared by Pavlo Shevchuk

I. Introduction

1. Ukraine consists of the 27 administrative regions (24 oblasts, Autonomous Republic of Crimea and two cities with special status Kyiv and Sevastopol). Most regions in Ukraine fit within 800–3000 thousand people. Dnipropetrovsk (with more than 3 million people) and Donetsk (with more than 4 million) regions and the city of Sevastopol (about 400 thousand) are the exceptions. Thus, Ukraine's regions roughly correspond to NUTS 2 regional level.
2. Ukraine skipped 2010 census round. The last Ukrainian population census was carried out on 5 December 2001. Since then post-censal population estimates are based on annual vital statistics and migration data. With the long period passed from the year 2001, accumulated inaccuracies in reporting births, deaths and especially migration can result in biased estimates of population. This disadvantage is more serious for regional population estimates level and will be detailed hereafter and will be discussed further below.
3. In 2014, an armed conflict broke out in Ukraine, starting with the annexation of Crimea by the Russian Federation and intensifying in the Donbas region. This caused new obstacles which aggravated population data quality in Ukraine.
4. Regional population projections in Ukraine are needed by regional authorities, policymakers and planners to gauge future demand for food, water, and energy and infrastructure capacity. This paper proposes an ex-post analysis of regional population projection errors for Ukraine.

II. Data and method

A. Data issues

5. As mentioned before, one of the important issues in terms of reliability of population estimates for Ukraine is a long period passed from the last census. Post-censal population counts are estimated at the beginning of each calendar year based on the population structure by age and sex at the beginning of the previous calendar year, the number of vital events and migration registration during that year. Thus, the accuracy of current population estimates depends on reliability of these components.
6. Traditionally the most problematic and significant is undercounting of migrants (mostly emigrants). This problem exacerbated since 2014 when Ukraine faced with emigration flaws from regions of military conflict. Unfortunately, magnitude of these flaws remains uncertain. Since April 4, 2016 the residence registration was delegated from State Migration Service of Ukraine to executive bodies of local communities. As it turned out, some of them were not prepared for technical reasons which led to the worsening of the completeness registration of migrants.
7. Until 2014 the vital statistics registration in Ukraine was complete. Vital events can be registered at place of occurrence (birth or death), place of residence of the parents (or one of them) in case of birth, the most recent address or the place of burial in case of death. Generally this practice is reliable. But in regional aspect there was (and is) some peculiarities. If a woman gave birth in a region other than she registered (but actually lives) this birth can be ascribed to region of occurrence. This results in some numerator-denominator inconsistency in the fertility rates at the regional level. Thereby migratory attractive regions have overestimated fertility rates and vice versa. The similar approach applies to deaths. If an elderly parent lives with hers/his children not registered as permanent resident in this region hers/his death can be ascribed to this region. This situation causes deaths cases in ages in which there no population in this region. In these cases the State Statistics Service of Ukraine (SSSU) uses special procedure for recalculation the number of immigrants to ensure no negative values in population estimates in such regions. But it's just ad hoc arithmetic operation because in region of permanent residence such a person remains

"to live". This situation is flaring for senior ages with small number of people. But it is applicable to any other not registered persons. Obviously, it leads to mismatch of numerator and denominator in the mortality rates estimations at the regional level.

8. Since 2014 vital statistics and migration data can not be collected for the whole territory of Ukraine. SSSU does not collect data on non-government controlled areas (NGCA). It causes a time series inconsistency. Also military conflict made it impossible to get official birth and death certificates in part of Ukraine. Accordingly, such events do not get into statistical processing for the current population estimates. Citizens residing in non-government controlled areas can register vital events in any other government-controlled region of Ukraine. This is the only way to obtain official Ukrainian documents. These vital events relate to the territories of permanent residence in Donetsk or Lugansk, if they are registered in government controlled parts of these regions. If NGCA residents from the Donetsk or Luhansk regions register vital events in other regions of Ukraine, then these vital events assign to the region of registration. In this case, the data processing does not allow to check which event associates with the NGCA. Thereby this is numerator-denominator inconsistency all over again. Hence, the information on vital events in these areas is incomplete. Obviously registration of migrants in NGCA is uncertain entirely.

9. In this paper the results of regional population projection 2009 revision is compared to an official SSSU data. The two periods will be distinguished: until 2014 (including 01 January 2014) and after that. As mentioned above since 2014 Donetsk and Luhansk regions have troublesome population statistics. Concerning Crimean Peninsula the occupation administration was carried out population census on October 2014, thereby population count for Autonomous Republic of Crimea and the city of Sevastopol were corrected and time series become incomparable. So, for the first period will be analysed 27 regions and for the second period will be analysed 23 regions. Also it should be remembered the data quality problems listed above.

B. Projection methodology

10. The method used for estimating future population numbers is cohort component by 1-year age groups for each calendar year. Open age interval is 100 years. The launch year of this projection is 2009 (the population data as of 01 Jan 2009) and the target year is 2035 (the population as of 01 Jan 2036). The projection is deterministic and has 3 variants conventionally named "medium", "high" and "low".

11. The discrepancy between sum of regions and the whole country was distributed proportionally by region, sex and 1-year age groups.

12. For calculation mortality and fertility rates by region to be consistent to Ukraine was used Brass models. Mortality and fertility rates of Ukraine were taken as a standard. Models relational parameters for regions were estimated based on actual 2008 data.

13. Migration projection was developed by expert judgement. Firstly hypotheses concerning the main directions of future migration flaws by regions were worked out. Then it was synthesised to immigrants and emigrants numbers. These numbers were adjusted to be consistent to projection for the whole country by reciprocal balance matrix.

III. Analysis of Regional Population Projection Errors for Ukraine

14. Population projection for regions of Ukraine 2009 revision is available at https://idss.org.ua/forecasts/region_pop_proj_en

15. Following the approach proposed by S. Rayer and S. Smith (2010) the two indicators were calculated: mean absolute percent error (MAPE) and mean algebraic percent error

(MALPE). The first represents measure of precision and the second one represents measure of bias.

$\mathrm{MAPE} = \sum |PE_t| / n$, where *PE* denotes percent error: $PE_t = ((F_t - A_t)/A_t) \times 100$, *t* the calendar year *F* the population projection, *A* the actual population estimate, and *n* the number of areas. And $\mathrm{MALPE} = \sum PE_t / n$.

16. As it turned out the "Medium" variant underestimates the population dynamics, the MALPE is slightly below zero (Table 1). "High" variant is positive and was closer to zero but finally diverted too high. "Medium" variant become most close to SSSU estimations. Respectively the "Low" variant is least accurate.

Table 1
Mean absolute percent error (MAPE) and mean algebraic percent error (MALPE)

	2011	2012	2013	2014	2015	2016	2017	2018	2019
Variant	For 27 regions				For 23 regions				
	MAPE								
Medium	0.4	0.6	0.9	1.1	1.5	1.6	1.8	1.9	2.0
High	0.2	0.4	0.5	0.7	0.8	1.0	1.4	1.9	2.4
Low	0.5	0.9	1.4	1.9	2.4	3.0	3.4	3.9	4.3
	MALPE								
Medium	-0.2	-0.4	-0.6	-0.7	-0.8	-0.9	-0.9	-0.9	-0.7
High	0.1	0.2	0.2	0.4	0.5	0.8	1.2	1.6	2.2
Low	-0.5	-0.8	-1.4	-1.9	-2.4	-3.0	-3.4	-3.9	-4.3

17. It was not found correlation between accuracy of projection and population size of a region. Correlation coefficients for 2014 (as of 1 January, the last peaceful estimation) are -0.09, 0.05 and 0.00 for "Medium", "High" and "Low" variants respectively. For 23 regions as of 1 January 2019 correlation coefficients are -0.14, 0.45 and -0.09 for "Medium", "High" and "Low" variants respectively. Better coefficient for "High" variant can be explained by systematically overestimation which encountered with slightly better population dynamics in migratory attractive regions.

18. As expected errors by specific age are higher. The biggest errors account for 90 years and older due to small numbers problem (Table 2) and for children 0–4 years old due to fertility hypothesis error which will be discussed hereafter. Also errors are higher with projection horizon lengthened. The MAPE for any age 2019 is higher than 2014. Errors for 20–24 years old a bit higher than majority other age groups due to migration hypothesis error. Quite high maximum errors amongst regions (Table 2) depend on small population size at 5-years age groups by region.

19. One of the most valuable causes of projection errors is incorrect fertility hypothesis. It leads to substantial deviation in population size under 5 years old compared with current SSSU estimation only 5 and 10 calendar years later. At period 2009–2013 fertility was overestimated for 18 regions out of 27 and MAPE for "Medium" variant is 2.1%. At period 2014–2018 fertility projection accuracy falls rapidly. In that period fertility is overestimated for 22 regions out of 23 and MAPE for "Medium" variant is 9.1%. The only region with underestimated fertility at period 2014–2018 is Kyiv which brings us back to the problem of current population estimation accuracy. Kyiv probably has bigger actual population in reproductive age than by official estimation.

Table 2
Projection errors by age and variants, %

	2014, for 27 regions						2019, for 23 regions					
	MAPE			Maximum error			MAPE			Maximum error		
Age	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low
0–4	2.1	8.1	7.1	6.2	13.2	11.9	8.4	24.9	8.4	23.5	36.7	21.6
5–9	0.6	0.5	0.4	2.9	2.8	1.9	2.6	7.8	7.9	10.1	13.9	17.0
10–14	0.8	0.6	0.4	4.4	5.2	1.5	1.3	0.8	1.1	5.3	4.7	5.6
15–19	1.4	1.2	1.2	5.1	4.7	4.8	3.3	3.1	3.0	11.9	14.3	9.3
20–24	2.3	2.0	2.2	8.3	10.0	7.5	4.0	3.1	3.6	11.2	11.7	8.7
25–29	1.5	1.2	1.6	5.1	5.8	4.8	4.1	3.2	4.2	15.2	18.2	11.5
30–34	1.0	0.6	1.2	3.3	2.6	3.5	2.7	1.8	3.3	9.6	8.7	11.2
35–39	1.2	0.6	1.2	3.7	3.0	3.0	2.1	1.2	2.8	8.6	6.6	10.3
40–44	1.4	0.8	1.5	3.7	3.4	3.0	2.2	1.0	3.0	7.4	5.1	8.9
45–49	1.5	0.8	1.7	3.2	2.5	2.8	2.7	1.0	3.6	7.2	4.9	8.6
50–54	1.6	0.8	2.1	2.8	1.8	3.2	2.9	1.1	4.3	6.6	4.1	8.2
55–59	1.6	0.7	2.2	3.0	1.9	3.5	3.1	1.2	5.0	7.2	4.5	9.1
60–64	1.5	0.4	2.1	2.4	2.3	3.0	2.7	1.0	5.1	7.0	3.7	9.3
65–69	1.4	0.7	2.1	3.9	3.0	3.3	2.3	2.3	4.9	5.5	9.1	8.4
70–74	1.4	1.1	2.5	5.1	3.4	4.2	2.1	3.8	5.1	8.3	10.3	7.9
75–79	1.5	2.1	2.6	5.0	4.1	6.0	2.4	6.1	6.4	10.4	12.1	11.7
80–84	2.7	3.9	3.7	7.5	8.9	9.8	4.6	10.7	6.8	13.6	17.4	15.8
85–89	4.5	5.8	5.0	12.1	13.5	13.5	8.7	18.0	8.8	21.6	33.4	26.1
90+	11.6	12.3	11.7	31.6	32.7	34.2	14.9	21.1	17.9	39.6	64.2	45.9

20. "Medium" variant overestimated mortality at period 2009–2014 for all regions. At 8 regions were registered fewer deaths than even "High (life expectancy)" variant. MAPE for "Medium" variant is 11.4%. Maximum deviation observed for Kyiv. The explanation might be the same like for fertility. It seems strange but at period 2014–2018 mortality projection appears more precise. MAPE for "Medium" variant at this period is 6.3%. Mortality projection errors have less prominent impact on age structure than fertility projection errors because fertility errors apply to youngest cohorts immediately while mortality errors allocate through all age groups.

21. The future net migration is the hardest component to estimate. Registered net migration exceeds high/low bands of this projection for 20 out of 27 regions in period 2009–2013. For period 2014–2018 migration projection is slightly better. In that period net migration exceeds high/low bands of the projection for 15 out of 23 regions. It should be reminded that registration of migrants is least certain. After next census and due intercensal correction to population data it might appear that this projection actually is more precise.

IV. Conclusion

22. Given that arrivals are registered better than departures, it is more likely that the actual migration balance is lower than reported one. Therefore, it is expected that population size in most regions of Ukraine is lower than currently estimated. It has been suggested that the errors of the medium variant projection (Table 1), which seems to be lowering the population, is actually lower, and the population projection is closer to the real population number.

23. One of the shortcomings of this projection is that its starting point is 2009 that is more than 7 years after the last census. Therefore, the base period may not be very suitable and the

launch year may already contain errors. The Ptoukha Institute for Demography and Social Studies intends to update regional projections immediately after the new census is coming in 2020 and to revise them within 5 years after the census.

24. This ex-post analysis of population projection errors can be useful for the future sub-national population projections. After evaluation of the accuracy of the current population estimates and its adjustment to the upcoming 2020 Census results the sources of possible errors would be reduced and the regional population estimates would be revised.

References

Ptoukha Institute for Demography and Social Studies of the National Academy of Sciences of Ukraine (2009). Available at https://idss.org.ua/forecasts/region_pop_proj_en

Rayer, S., and S. Smith. (2010). Factors Affecting the Accuracy of Subcounty Population Forecasts. *Journal of Planning Education and Research* 30(2): 147–161. doi: 10.1177/0739456X10380056