Supporting material for "A system of population estimates compiled from administrative data only"

John Dunne

Central Statistics Office (CSO), Cork, Ireland.

E-mail: John.Dunne@cso.ie

Li-Chun Zhang

University of Southampton, Southampton, United Kingdom.

Summary. This document provides a comparison of the PECADO estimates from the use of two alternative list B data sources: the Driver Licence based dataset (DLD) and the Quarterly National Household Survey (QNHS) dataset.

Evaluation of DLD as list B using QNHS as an alternative list B

Figure 1 presents a comparison of two sets of population estimates, where one set has been compiled using DLD as list B (blue) and the other using QNHS as list B (green). The PAR is list A in both cases.

The confidence intervals for the population estimates compiled with QNHS have been estimated assuming independent capture (in list B) across the individuals. In reality this is not the case, due to the two stage design of the survey and the fact that all individuals in a house are captured if one person in that house is captured for the survey. The design effect for the QNHS when measuring unemployment rates has previously been estimated in the region of 1.5 to 2. If we assume the same design effect also carries over to these population estimates, then the confidence intervals should be adjusted by a factor equal to the square root of the design effect. For instance, if we have a design effect of 2 then the confidence intervals should be widened by a factor of $\sqrt{2}$.

It can be seen that the differences between the two set of estimates are not statistically significant in most of the age-gender groups. For age categories under 18, the list A counts are in close agreement with the QNHS-based population estimates, which means that the non-coverage of these people by DLD is not an issue in practice.

The differences between the two sets of estimates peak for both males and females around the age 30, plus an odd peak for males at age 20. Whether or not these differences are deemed significant statistically, the exhibited pattern suggests it is likely that there are some systematic differences between the two sources in terms of these age groups, where migration or mobility otherwise is high in the population.

One possible contributing factor is that the DLD contains some erroneous records, i.e., there may be persons renewing their driving licence who reside outside the State, even though there is a requirement for providing evidence that they reside in the State before renewing their driver licence. Erroneous DLD records would cause over-estimation of the target population size. This may be an arguable explanation as there is a cost to letting a drivers licence lapse - a person may have to resit their driving test. However,

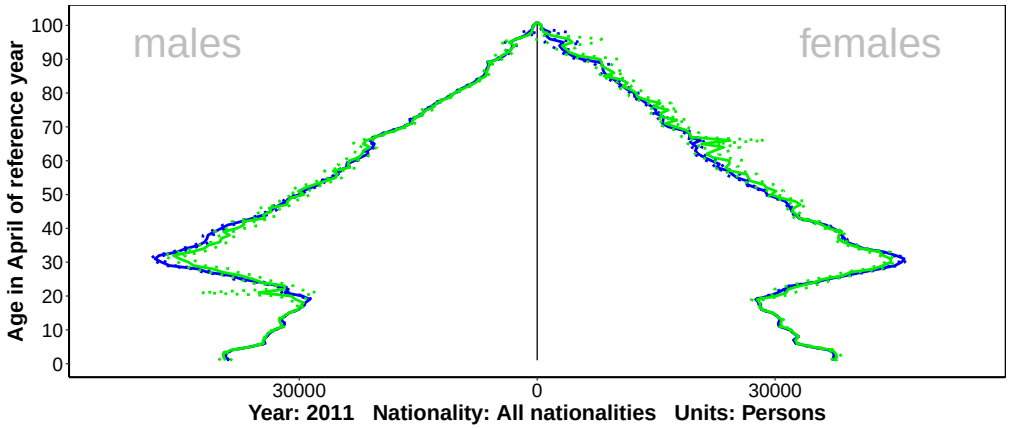

Fig. 1. Comparison of population estimates for 2011 using two different data sources as list B: QNHS (green) or DLD (blue), 95% confidence intervals derived from QNHS marked with dots.

in practice the burden of obtaining evidence (i.e., a utility bill or bank statement with address) is considered a significant deterrent.

Meanwhile, survey refusals and hard-to-contact persons could also potentially cause under-estimation using the QNHS. Heuristically, suppose the population size (within each estimation block) is given as $N = N_0 + N_1$, where N_1 is the number of people with zero (or nearly zero) probability of response in the QNHS and N_0 is that of the others. For instance, no matter when or how the survey is carried out, there will always be some people who are refusals or cannot be contacted at all. Accordingly, let the list A capture be given as $x = x_0 + x_1$, and the QNHS capture be n_0 (given $n_1 = 0$), such that the joint list capture is m_0 (given $m_1 = 0$). For the corresponding DSE, we would have

$$(x_0 + x_1)E\big(\frac{n_0}{m_0}\big) = (1 + \frac{x_1}{x_0})E\big(\frac{x_0 n_0}{m_0}\big) = \ (1 + \frac{x_1}{x_0})N_0$$

where

$$\frac{(1 + \frac{x_1}{x_0})N_0}{N_0 + N_1} = \frac{1 + \frac{x_1}{x_0}}{1 + \frac{N_1}{N_0}} < 1 \quad \Leftrightarrow \quad \frac{x_1}{N_1} < \frac{x_0}{N_0}$$

That is, under-estimation could be expected if the people who are hard-to-catch in the QNHS are also under-represented in the PAR (as list A).

Hence, heterogeneous QNHS survey nonresponse is a likely cause for the observed differences, although one cannot completely rule out the possibility of erroneous records in the DLD either. With the long-term objective of census transformation in mind, the choice of DLD as list B over QNHS is preferable unless there is sufficient evidence to dismiss the DLD based estimates. An obvious reason for this preference is that the DLD list is significantly larger and therefore provides much more precise estimates. Of course, one should also aim to strengthen the administrative routines of license renewing, which have been introduced in order to eliminate the potential erroneous records in DLD.