A NEW METHOD OF ESTIMATING UNDERCOUNT IN THE CENSUS OF POPULATION, USING A BAYESIAN APPROACH

Philip Redfern
c/o 12 Hartley Close, Charlton Kings, Cheltenham, Glos., GL53:9DN, UK.

1. Summary

The new method of measuring net undercount in a census of population is a variant of demographic analysis and is founded in the belief that the most reliable counts of people are the counts of births and deaths. It has been applied to the 1991 census of England and Wales (E&W). The method has two distinctive features. The first is that it bases its estimates of numbers of emigrants on figures of persons born here and recorded as residents in the 1990-91 censuses of other countries. The corresponding data on immigrants are taken from our own census. The method does not therefore require the data on migratory flows that are an essential component of conventional demographic analysis. The second feature is the Bayesian approach in which (1) each of 30 uncertain elements in the calculations is given an a priori error distribution and (2) three constraints on the sex-age profile of percentage net undercount are imposed a priori. This, in conjunction with a Monte Carlo process, generates an error distribution for net undercount. These merits of the method are offset by the demerit that the calculations must await results from the censuses of the other countries in which substantial numbers of our emigrants reside.

2. Outline of the method

People born in E&W will be called natives, of whom those resident in E&W are natives-at-home and those resident outside E&W emigrants. People born outside E&W and resident in E&W are immigrants. Table 1 gives illustrative figures and is a summary of separate analyses for each sex for each of 12 age groups 0-4, 5-9, ..., 55-59. The calculations are limited to persons under 60 because of uncertainties about death rates at higher ages.

Table 1: An estimate of net undercount at ages under 60: 1991 census of E&W

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Births in E&amp;W, mid-1931 to mid-1991</td>
<td>41,867</td>
<td></td>
</tr>
<tr>
<td>Census counts at ages under 60, mid-1991</td>
<td>39,602</td>
<td></td>
</tr>
<tr>
<td>Census of E&amp;W 1991: all residents = (10)+(11)</td>
<td>1,386</td>
<td></td>
</tr>
<tr>
<td>Immigrants (census of E&amp;W)</td>
<td>254</td>
<td></td>
</tr>
<tr>
<td>Natives-at-home (census of E&amp;W)</td>
<td>1,132</td>
<td></td>
</tr>
<tr>
<td>Emigrants (other countries' censuses)</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>All natives = (2)-(7) = (9)+(10)</td>
<td>1,220</td>
<td></td>
</tr>
<tr>
<td>Census count of all natives = (5)+(6) [P1]</td>
<td>38,383</td>
<td></td>
</tr>
</tbody>
</table>

Estimation of net undercount is in two stages, see Table 1.

Stage 1: Line 1 shows births in E&W in the 60 years to mid-1991. By applying rates of survival (from cohort life tables) separately to each of the 12 male and 12 female 5-year birth cohorts, we estimate the numbers who survived to mid-1991 wherever in the world they lived (line 2). Lines 3 to 5 give data from the 1991 census of E&W 'rolled forward' from Census Day (21 April) to mid-year. The census count of emigrants (line 6) is from the 1990-91 censuses of 24 main receiving countries (adjusted, in principle, to mid-1991 timing), plus estimates of (i) emigrants in some 200 residual countries (see Section 4) and (ii) (so far as they are not already counted) UK Armed Forces and their families overseas. The difference between the census count of natives worldwide (line 7) and the estimated number of native survivors (line 2) is the net undercount of natives (line 8); this undercount occurs mainly in the census of E&W but partly in other countries' censuses.
Stage 2: We make an inference about undercount in 1991 in the population resident in E&W [P2] on the basis of the Stage 1 findings on undercount in the population born in E&W [P1]. This is feasible because the two populations have a common element that exceeds 90 per cent of each. (See the census counts of P2 and P1 in lines 3 and 7 of Table 1; the common element is natives-at-home in line 5.) To make the inference, assumptions must be made about the ratios of the rate of undercount of emigrants (in other countries' censuses) to the rate of undercount of natives-at-home to the rate of undercount of immigrants. For the illustrative purpose of Table 1, these ratios are assumed to be 1:1:2 in all sex-age groups, which leads to the estimates of undercount in lines 9 to 12.

3. The error model

What I call a random value parameter (RVP) is introduced into the calculations when the true value of a parameter is not known. For each RVP lower and upper limits have been chosen on the basis of available evidence (often by extrapolation or inference) or, failing that, 'informed guesswork'. A value has then been given to the RVP by drawing at random from a rectangular distribution between the chosen limits. The RVPs have been drawn independently of one another.

The 12 topics for which RVPs represent unknown quantities include:
(a) Undercount in births.
(b) Errors in the death rates used in compiling the cohort life tables.
(c) Conversion of the published cohort lifetables (which embody the experience of natives-at-home plus immigrants) into lifetables embodying the experience of natives-at-home plus emigrants.
(d) Adjustments to the census data for the 24 main receiving countries in respect of definitions, timing and respondents' misstatements of country of birth; and analogous adjustments to our own census data on immigrants.
(e) Errors in the estimate of emigrants in the residual countries (see Section 4).
(f) Differences in the undercount rates between emigrants, natives-at-home and immigrants (see Section 2, Stage 2).

Errors in each of 12 topics in each of 24 sex-age groups could be represented by 12x24 (=288) RVPs. This number has been reduced to 30 by introducing into each topic (with a few exceptions) an error term of the form $t = t_1 + t_2x + t_3a$, where $t_1$, $t_2$ and $t_3$ are three RVPs distributed independently; where $x$ is a sex variable (+1 for males and -1 for females); and where $a$ is an age variable (ranging from -1 in the age group 0-4 to +1 in the age group 55-59). Thus, $t_2$ is a sex differential and $t_3$ is an age differential.

4. Emigrants in the residual countries

To estimate emigrants in the residual countries an indicator was needed which was available for every country. I chose the number of returners as recorded in our 1991 census. A returner from country k is a person born in E&W who was counted as resident in E&W in the 1991 census and who reported a usual address one year before the census in country k. The basic method is as follows: (1) for each of the 24 main receiving countries calculate the ratio of emigrants to returners ($u$); and (2) estimate emigrants in the residual countries by multiplying the number of returners from the residual countries by a value of $u$ that has been extrapolated from the values in the 24 countries. This procedure is done separately for each of the 24 sex-age groups.

5. The Monte Carlo process

By replicating the calculation, using a different set of 30 random numbers each time, we get a distribution of the estimate of undercount. This model has similarities to, and differences from, the model of undercount in the 1990 US census developed by Robinson et al (1993). The two models exhibit major differences of demographic structure and equations. In addition, their model assumed a normal distribution of each unknown parameter, whereas I used a rectangular distribution because I did not feel confident in assuming a priori that the probability peaked sharply at the centre of its range.

As a second Bayesian step, accept a replication only if it generates a sex-age profile of net undercount that is a priori feasible, that is, if the profile satisfies three criteria: (1) The rate of net undercount among women aged 35-59 exceeds zero; (2) the rate among children aged 10-14 exceeds the rate among women aged 35-59; and (3) at ages 55-59 the rate
among men exceeds the rate among women. Just over 40 per cent of the replications satisfy these criteria. Note that the sex-age profiles of net undercount in the 1990-91 censuses of Australia, Canada and the US conform to the criteria; so do the profiles from our 1991 census, as estimated by the Office for National Statistics (ONS).

Table 2. Percentage net undercount at ages 0-59 combined: five models
Each model shows the results from 250 acceptable replications.

<table>
<thead>
<tr>
<th>ONS est.</th>
<th>2 B ONS</th>
<th>2 B 3 criteria not applied</th>
<th>A Base model</th>
<th>C Range of 18 RVPs mult'd by 1.5</th>
<th>D Range of all RVPs mult'd by 1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st pc'ile</td>
<td>1.99</td>
<td>2.5/</td>
<td>2.82</td>
<td>2.83</td>
<td>2.48</td>
</tr>
<tr>
<td>Median</td>
<td>3.36</td>
<td>3.69</td>
<td>3.71</td>
<td>3.73</td>
<td>3.70</td>
</tr>
<tr>
<td>99th pc'ile</td>
<td>4.91</td>
<td>4.92</td>
<td>4.90</td>
<td>5.26</td>
<td>6.46</td>
</tr>
<tr>
<td>Width of 98 pc interval</td>
<td>2.92</td>
<td>2.35</td>
<td>2.08</td>
<td>2.43</td>
<td>3.98</td>
</tr>
</tbody>
</table>

The 18 RVPs representing sex and age differentials.

6. The results: the base model A
Table 2, col.A, giving percentage net undercount for ages 0-59 combined, summarises the distribution generated by 250 replications. The comparison between cols.Z and A confirms the marked narrowing of the distribution as a result of introducing the three criteria detailed in Section 5. Figs.1a and 1c (males) and 1b and 1d (females) show the age profiles; the shaded area is bounded by the 1st and 99th percentiles of the distributions. The dotted lines in Figs.1a and 1b represent the ONS estimates.

Fig. 1. Percentage net undercount by age, 1990-91.
The shaded area is bounded by the 1st and 99th percentiles of estimated undercount in E&W (model A).

7. A review of the error model
The robustness of the base model A (Table 2, col.A) has been tested by developing three alternative models. In model B (col.B), each of the three criteria of Section 5 is relaxed by 0.25 per cent (that is, insert 'plus
0.25 per cent' in front of the word 'exceeds' wherever it appears). Model B is a staging post between Z and A.

The limits chosen for the RVPs may be too narrow even though I, lacking expert knowledge, aimed to err towards wide limits. In model D (col.D) the ranges assigned to each of the 30 RVPs are widened by a factor of x1.5, leaving the central point of each range unchanged. The evidence available for assigning limits to the 18 RVPs that represent sex and age differentials (t1 and t3 in Section 3) is weak, and so in model C (col.C) the ranges assigned to each of just these 18 RVPs are widened by a factor of x1.5; thus, model C is a staging post between models A and D. By comparison with base model A, model D has a broader distribution, most marked in the 99th percentile.

The a priori assumption that the correlations among the 30 RVPs are zero is important and, I believe, reasonable. Robinson et al say (1993, p.1069) that it is impossible to estimate the correlations empirically. A more serious criticism of the methodology may be that some of the formulae for modelling error (detailed in the full version of this paper) may be inappropriate.

8. Some international comparisons

Superimposed in Figs.1c and 1d are the age profiles of net undercount in the 1990-91 censuses of three other countries: Australia, whose figures are based primarily on a PES; Canada, based on the Reverse Record Check; and USA, based on demographic analysis. Two striking features of Figs.1c and 1d are the similarity of shape of the profiles of Australia, Canada and E&W and the dissimilarity of shape of the US profiles. This last is a paradox.

9. Concluding remark

The new method of estimating net undercount should be applicable in countries whose frontiers have remained substantially unchanged for 60 years if (i) they have reliable statistics of births and deaths and (ii) numbers of emigrants and immigrants are not large compared with numbers of natives-at-home. Because the method cannot be applied until (unadjusted) census data are obtained from other countries, its main uses should be to validate estimates of undercount that have already been made by other methods, and to revise annual population estimates which appear to have been wrongly rebased at the census.

Acknowledgements appear in the full version of this paper.

REFERENCE


RÉSUMÉ

UNE NOUVELLE MÉTHODE POUR ESTIMER LE SOUS-DÉNOMBREMENT DANS LE RECENSEMENT DE LA POPULATION, EN SE SERVANT D'UNE APPROCHE BAYESIENNE.

La nouvelle méthode pour mesurer le sous-dénombrement net dans un recensement est une variante de l'analyse démographique qui est fondée sur la notion que les dénombements les plus fiables de personnes sont les dénombements de naissances et de morts. On a appliqué cette méthode au recensement de l'Angleterre et du Pays de Galles de 1991. La méthode a deux particularités distinctives. La première est que ses estimations des nombres d'émigrés sont basées sur les chiffres pour les personnes nées ici et enregistrées comme domiciliées dans les recensements de 1990-91 d'autres pays. Les données correspondantes sur les immigrés viennent de notre propre recensement. La méthode ne nécessite pas donc les données sur les flux migratoires qui sont un élément essentiel de l'analyse démographique conventionnelle. La deuxième particularité est l'approche bayésienne où (1) une distribution d'erreurs a priori est donnée à chacun de 30 éléments incertains dans les calculs, et (2) trois contraintes sont imposées à priori sur le profil sexe/âge du pourcentage de sous-dénombrement net. Conjointement avec un processus de Monte Carlo, ceci produit une distribution d'erreurs pour le sous-dénombrement net. Ces mérites de la méthode sont contrebalance par le désavantage que les calculs doivent attendre des résultats des recensements des autres pays dans lesquels un grand nombre de nos émigrés résident.