Journal of Official Statistics, Vol. 19, No. 4, 2003, pp. 421–448

Estimating Census Undercount by Demographic Analysis: New Approaches to the Emigrant Component

Philip Redfern[1]

This is a method of jointly measuring net undercount in a census of population and the stock of emigrants as at census date (that is, the stock of persons born here and resident abroad). It is a variant of demographic analysis but does not require the data on migratory flows that are an essential component of the conventional analysis. The method is applied here to the 1991 Census of England and Wales (E&W).

The calculations start from statistics of births and deaths. Part of the stock of emigrants is measured by data from contemporaneous censuses in a few foreign countries – perhaps 3 or 4 foreign countries or, to avoid the delay in assembling foreign data, even 0 foreign countries. The rest of the emigrant stock is *estimated* from data extracted from the E&W Census on persons who have returned from abroad, supplemented (in some versions of the method) by historical data from foreign censuses. Each of the uncertain elements in the calculations is given an a priori error distribution. Empirical constraints, based on evidence from comparable countries, are imposed on the sex-age profile of percentage net undercount.

The method is inexpensive**.** A Monte Carlo process generates results in the form of frequency distributions. Most importantly, the near congruence of the results from different versions of the model demonstrates the robustness of the methodology.

Key words: Census of population; demographic analysis; emigration; international migration; overcount; population statistics; post-enumeration survey; undercount; under-enumeration.

1. Introduction

Because of the important rôle of the census of population in social and economic policy-making, increasing effort has been made in the last half century to measure the accuracy of census counts. A main objective has been to measure the net undercount of persons of each sex and age-group, both at national and sub-national levels; the net undercount is the number of persons missed from the census less the number counted twice and less the number wrongly counted.

Methods of measuring undercount in a conventional census are of two kinds: micro methods exemplified by the post-enumeration survey (PES) and macro methods described as demographic accounting or demographic analysis (DA). The DA method involves the comparison of the census counts with independent estimates of population compiled from data on births, deaths and migrants.

But estimates of undercount are themselves subject to substantial errors, though these errors are rarely quantified. Errors in the PES stem from incomplete response to surveys and difficulties in matching persons. Errors in DA stem mainly from the poor quality of

[1] 1 Greenhills Close, Charlton Kings, Cheltenham, Glos. GL53 9EY, UK. E-mail: philipredfern@btinternet.com

Table 1. Net undercount in the U.S. 1990 Census, per cent

	PES	DA	Difference DA-PES
Both sexes, all ages	1.58	1.85	0.27
Males 0–29	3.16	2.16	−1.00
30–49	1.85	3.83	1.98
50 and over	−0.57	2.72	3.29
Females 0–29	3.03	1.66	−1.37
30–49	0.88	0.50	−0.38
50 and over	−1.20	0.25	1.45

Sources: PES: U.S. Census Bureau (1992). DA: Data elaborating Table 3 of Robinson et al. (1993).

data on migrants. The extent of the errors can be gauged by the wide differences between the estimates of undercount relating to a given census produced by different methods. Table 1 compares the estimates of net undercount in the 1990 Census of the United States produced by the PES and DA methods. Among the six sex-age groups shown, the average difference between the alternative estimates (ignoring its sign) is 1.6 percentage points.

In England and Wales (E&W) the PES carried out after the 1991 Census failed to yield a valid estimate of undercount, so that undercount could only be estimated by demographic analysis (OPCS and GROS 1995). However, uncertainty about the reliability of the data on migratory flows in the decennium 1981–1991 – a main element in the analysis – led me to seek an alternative way of measuring our 1991 undercount. My alternative was a variant of demographic analysis and appears in ''A Bayesian model for estimating census undercount, taking emigration data from foreign censuses'' published in The International Statistical Review (ISR) (Redfern 2001). A summary is given in Sections 2 to 6 below. I shall refer here to that paper in a form such as ''ISR Section 1.''

In this article the term native means a person born in E&W. A native who is resident in E&W at the time of the 1991 Census is a native-at-home. A native who is resident outside E&W at that time is an emigrant. And a person born outside E&W and resident in E&W at the time of the 1991 Census is an immigrant.

The ISR methodology, as applied to the 1991 Census of E&W, involved the following key stages: (1) data were extracted from the 1990–1991 Censuses of 24 other countries on numbers of persons born in E&W who were then resident there (that is, emigrants); and estimates were made of emigrants in the 200 or so ''residual'' countries using, as indicators, data from our own E&W Census on returners, that is, persons resident here in 1991 whose usual address one year before the census date had been in one of the residual countries; (2) these estimates of emigrants, in conjunction with data on births and deaths and our own census count of natives-at-home, led to an estimate of net undercount among natives worldwide; (3) by making appropriate assumptions about the ratios of rates of undercount among three categories – natives-at-home, emigrants and immigrants – the result of stage (2) could be converted into an estimate of undercount in the 1991 Census of E&W; (4) each of the uncertain elements in the calculations – including of course the assumptions made at (3) – were given an a priori error distribution based on available

evidence; and (5) the resulting sex-age profile of percentage net undercount was subject to three a priori constraints based on evidence from comparable countries. The results in the form of a frequency distribution are described as Model A.

The ISR methodology was seen to have shortcomings. Critics noted that other countries' censuses may employ methods, timing and definitions of residence that are different from UK practice. And, though the ISR methodology had attempted to make appropriate allowances for such differences, the critics claimed that the estimates of undercount would be subject to an unknown error. Moreover, the ISR method could not be applied until data from the foreign censuses had been received. The timetable for measuring undercount therefore depended on the census timetables of countries which received substantial numbers of British emigrants and their willingness to give priority to UK requests for data.

The models developed in this article are again applied to the 1991 Census of E&W. To mitigate the demerits of the ISR model, this article's models use data on emigrants extracted from the 1990–1991 Censuses of only a very small number of foreign countries: 3 in the model labeled P, and 3, 1 and 0 respectively in Models R3, R1 and R0. (I use the term foreign to refer to countries outside the UK.) In place of the data on 1991 emigrants that, in the ISR model, had been extracted from other foreign censuses, this article's models P and R substitute estimates of emigrants using the same technique and data on returners as had been applied in model A to estimate emigrants in the 200 or so residual countries. In addition the R models use emigrant totals extracted from earlier (1980–1981) censuses in 20 or so countries as a supplementary indicator of emigrant levels in 1991. The a priori constraints on the sex-age profile of undercount now play an even greater rôle in controlling uncertainty than was the case with the ISR model.

Sections 7 to 14 describe the new models, give the results (each in the form of a frequency distribution) and make comparisons with the results from the ISR Model A. To test the robustness of the new models, Sections 15 and 16 examine a number of variant models and conclude that the new models have the qualities of consistency and robustness in good measure. The near congruence between the results of all the models leads to another conclusion: one can reject the argument that the different methods, definitions, and timing of foreign censuses seriously affect ISR-style estimates.

Section 17 refers to the need to test the new methods in another country or at another time, and suggests the 2001 Census of E&W as an obvious test bed.

A key conclusion (Section 18) is that the initial estimates of undercount can be made without using any emigrant data from contemporaneous foreign censuses, and therefore to a timetable comparable to that of conventional methods of measuring undercount. But these estimates ought to be refined later as data from a few contemporaneous foreign censuses become available. That may not happen until after the census agency has already published its first estimates of undercount using other methods, and in that case the agency may face conflicts of evidence and possible revision of its published estimates.

Finally Section 18 draws attention to three main strengths of the new methods: low cost; presentation of results in the form of frequency distributions rather than central estimates or ''best guesses;'' and robustness.

2. Underlying Principles

The starting point of the methodologies in both the ISR paper and this article is the Demographic Equation (1) below, which refers to all persons born in E&W (natives). I restrict the application of this equation to those born in the 60 years to mid-1991, which is taken for the purposes of this article as the date of the 1991 Census in the UK. These persons may be divided into 12 birth cohorts ($i = 1, 2, \ldots, 12$), each comprising a 5-year age-group. Thus, those born between mid-years 1986 and 1991 (cohort $i = 1$) were aged 0 and under 5 (written 0–4) at census date; and those born between mid-years 1931 and 1936 (cohort $i = 12$) were aged 55 and under 60 (written 55–59). A further division by sex (males, $x = 1$, and females, $x = 2$) produces 24 sex-age groups (x, i). Those natives surviving to mid-1991 may be divided into four categories: those resident in E&W (natives-at-home) who were either recorded in our census or missed by the census; and those resident outside E&W – whether in Scotland, Northern Ireland or outside the UK – (emigrants) who were either recorded (in principle) by a contemporaneous census taken in their country of residence or missed by such census. Natives with no usual residence will be included in the category ''missed from the E&W census'' if they were previously resident in E&W or in the category ''missed from another country's census'' if they were previously resident in another country. (I return to this last point at the end of Section 6.)

The Demographic Equation (1) reflects this fourfold categorisation of surviving natives. For convenience the symbols are taken from the ISR paper (and so retain the prime notation used there to indicate that corrections have been applied to the basic data).

$$B'(x,i) - D(x,i) = C'_N(x,i) + U_N(x,i) + C'_E(x,i) + U_E(x,i) \tag{1}$$

where $B'(x,i)$ is the number of births in cohort (x,i); $D(x,i)$ is the number in cohort (x,i) who died – whether at home or abroad – before mid-1991; $C'_N(x,i)$ is the count of natives-at-home in sex-age group (x,i) in the 1991 Census of E&W; $U_N(x,i)$ is the corresponding net undercount of natives-at-home; $C'_E(x,i)$ is (in principle) the count of emigrants in sex-age group (x,i) in 1991 Censuses taken outside E&W; and $U_E(x,i)$ is the corresponding net undercount of emigrants.

Table 2 gives illustrative figures for each of the items of Equation (1) for the age-group 40–44 ($i = 9$). All the elements of Table 2 are estimates subject to a margin of error. For births (B') the margin of error is very small. For deaths (D) error is greater, but still small; thus, some error is introduced when the available cohort life tables relating to persons living in E&W are converted to the basis needed for estimating D, that is, to life tables relating to persons born in E&W. (Incidentally, it is because death rates at ages 60 and over are much higher than at younger ages, and therefore introduce greater errors, that I applied the methodology only to age-groups under 60.) The census count of natives-at-home (C'_N) is a firm figure apart from an adjustment (included in it) to correct for census respondents' misstatements of their countries of birth. The less reliable elements of Table 2 are the count of emigrants in censuses taken outside E&W (C'_E) and the estimates of undercount (U_N and U_E).

In the ISR paper, the count of emigrants (C'_E) was built up under four headings: (1) counts of emigrants in Scotland and Northern Ireland extracted from those countries'

Table 2. Natives of E&W born in mid-year 1946 to mid-year 1951, as at mid-1991: an illustration. Thousands

		Males	Females
Births	$B'(x,9)$	1,998	1,883
less deaths before mid-1991	$D(x,9)$	−149	−100
Survivors to mid-1991		1,849	1,783
made up of:			
Natives-at-home			
Count in census of E&W	$C'_N(x,9)$	1,611	1,604
Census undercount	$U_N(x,9)$	53	5
Emigrants			
Count in censuses outside E&W	$C'_E(x,9)$	179	174
Census undercount	$U_E(x,9)$	6	1

censuses, which, though separate from the census of E&W, follow the same methods, definitions and timing; (2) counts of emigrants extracted from the 1990–1991 Censuses of 22 countries outside the UK which received substantial numbers of British emigrants (the countries are listed in Table 4); (3) an estimate of emigrants in the other 200 or so countries outside the UK (the residual countries) using the methodology outlined in Section 5 below; and, so far as not counted under other headings, (4) Ministry of Defence data on members of the UK Armed Forces and their families resident outside E&W (the AFs etc). Errors in the estimate of C'_E are due to several factors, including (1) the uncertainty of the estimate relating to the residual countries; (2) errors in converting foreign data on persons born in the UK to the basis needed here – namely, persons born in E&W; (3) an allowance for differences between foreign countries and the UK in census methods, definitions and timing; and (4) an allowance for census respondents' misstatements of their countries of birth.

The two remaining items in Table 2 taken together, undercounts (U_N and U_E), are calculated simply as a residual. This residual then has to be divided between the two items by making an appropriate assumption: in the illustrative example it is assumed that the rate of undercount among emigrants in censuses outside E&W was the same as the rate of undercount among natives-at-home. (Compare Section 3 below.)

Having thus made an estimate of the undercount of natives-at-home (U_N), the remaining step in estimating the total undercount in the census of E&W is to estimate the undercount of immigrants (U_I). This is illustrated in Table 3.

Table 3. The 1991 Census of E&W: illustrative totals for ages 40–44. Thousands

		Males	Females
Census counts			
Natives-at-home	$C'_N(x,9)$	1,611	1,604
Immigrants	$C'_I(x,9)$	207	229
Total		1,818	1,833
Undercount			
Natives-at-home	$U_N(x,9)$	53	5
Immigrants	$U_I(x,9)$	14	1
Total		67	6
Population of E&W		1,885	1,839

The census counts of natives-at-home (C'_N) and of immigrants (C'_I) include equal and opposite adjustments for respondents' misstatements of their countries of birth. The estimate of net undercount among natives-at-home (U_N) is taken from Table 2. The estimate of net undercount among immigrants (U_I) rests on an appropriate assumption about the rate of undercount: in this illustrative example it is assumed that the rate was twice the rate of undercount among natives-at-home.

3. Dealing with Errors in the ISR Model

Errors, uncertainties and assumptions are introduced at every stage in the calculations illustrated in Tables 2 and 3. A list of these errors was compiled and each error was modeled in the calculations by introducing a random variable to which I gave the name random value parameter (RVP). Each RVP was assumed to be drawn at random from a rectangular distribution ranging between a chosen lower limit for that RVP and a chosen upper limit. The RVPs, of which there were 30, were assumed to be distributed independently of one another. Independence was assumed partly because, for many but not all pairs of RVPs, it was intuitively reasonable: for example for a pair of RVPs representing respectively errors in emigrant numbers and errors in rates of undercount of immigrants. A more important reason was that any other assumption in respect of 435 correlation coefficients would have been impracticable. But the effect of ignoring correlation may be to understate the width of uncertainty intervals – a form of correlation bias.

Details of the RVPs and an indication of how the limits were assigned to each are given in ISR, Section 4 and Annex 1. I give an example. For sex-age group (x, i) the rate of undercount among emigrants in censuses outside E&W was expressed as a ratio, $V(x, i)$, of the rate of undercount among natives-at-home in the census of E&W. This ratio was taken to be

$$V(x, i) = 10^v \tag{2}$$

where $v = v_1 + v_2 \cdot s + v_3 \cdot a$; v_1, v_2 and v_3 are three RVPs distributed independently within chosen limits; s is a sex variable which takes the value $+1$ for males aged 20 and over, -1 for females aged 20 and over, 0 for children aged under 15, and at ages 15–19 the arbitrary values $+0.5$ for males and -0.5 for females; and a is an age variable in the range $(-1, 1)$ defined by the equation

$$a = (2i - 13)/11 \tag{3}$$

The term $v_2 \cdot s$ is a sex differential whose ''slope'' (v_2) is randomly chosen, and similarly the term $v_3 \cdot a$ is an age differential. The range of v_1 was taken to be $(-0.301, +0.301)$. Thus, ignoring the sex and age differentials, the net rate of undercount of emigrants was taken to be between $\frac{1}{2}$ and 2 times the rate among natives-at-home.

To generate the frequency distribution of the estimates of undercount, the calculation was repeated many times, giving the 30 RVPs a different set of 30 random numbers each time. Each such calculation may be termed a replication and the construction of the set of replications is a Monte Carlo process.

4. Introducing *a priori* Constraints into the ISR Model

With the aim of narrowing the frequency distribution of undercount generated by Monte Carlo, the ISR model embodied a second Bayesian feature. I discarded those of the replications which gave rise to a sex-age profile of net undercount that was infeasible in view of evidence from countries comparable to E&W. Writing $u(x, i)$ for the rate of net undercount in sex-age group (x, i) and u_F for the rate of net undercount among women aged 35–59 (an age-group among whom undercount is low), the Base Model A of the ISR paper was built up from replications which satisfied all of the following three criteria:

a) A positive net undercount among women aged 35–59, that is:

$$u_F > 0 \tag{4}$$

b) u_F is less than the rate of net undercount among children aged 10–14 (age group $i = 3$), or, more precisely:

$$\tfrac{1}{2}u(M, 3) + \tfrac{1}{2}u(F, 3) - u_F > 0 \tag{5}$$

c) at ages 55–59 (age group $i = 12$), the rate of net undercount among men is greater than among women:

$$u(M, 12) - u(F, 12) > 0 \tag{6}$$

The profiles in Australia and Canada, and in most instances in the U.S. too, conform to these criteria. The profile of the official estimates of net undercount in the 1991 Census of E&W also conforms. About 58 per cent of the original Monte Carlo replications were discarded as a result of imposing these three constraints. The resulting frequency distribution of undercount for Model A is described in Section 6.

The constraints must be chosen to reflect the demography and census methods of the country whose undercount is being estimated – in this case E&W. For example, evidence from the U.S. 2000 Census about the prevalence of double counting implies that inequality (4) would have to be relaxed if this article's methods were to be applied to that census.

5. The ISR Model: Estimating Emigrants in the Residual Countries

This section explains the novel method used in the ISR model to estimate emigrants in the 200 or so residual countries. Though peripheral to the ISR model, the method is central to the models developed later in this article where it has also been applied to estimate emigrants in most or all of the 22 foreign countries which received substantial numbers of British emigrants.

Let us revert to the application of the method to the residual countries. As an indicator of the number of emigrants resident in each country of the world outside E&W, data were extracted from the 1991 Census of E&W on numbers of returners from each country. A returner from country k is defined as a person born in E&W who was counted as resident in E&W in the 1991 Census and who, in that census, reported that his or her usual address one year before the census date was in country k. (But note that the formulation of the census question on previous residence varies from one country to another; thus, the U.S. Census has asked about place of residence five years earlier.) Some data on emigrants and returners are in Table 4, Cols. 1 and 2.

Table 4. Returners and emigrants, E&W, aged 0–59, 1991. Base model A

Country k	(1) Returners $R(k)$ *	(2) Emigrants $C_E(k)$ *	(3) Regression coefficients $f_1(k)$	(4) $f_2(k)$	(5) $f_3(k)$	(6) RMS ε	(7) Sum of squares of ε	(8) ξ
A	(nos.)	(000s)						
United Kingdom								
Scotland	14,977	295	0.08	−0.03	−0.02	0.12	0.32	0.51
Northern Ireland	3,411	43	−0.14	−0.08	0.26	0.34	2.47	3.69
Major emigrant destinations								
Australia	14,912	700	1.13	0.02	0.42	0.33	2.30	34.78
Canada	2,701	327	1.63	0.05	0.58	0.38	2.99	69.77
United States	10,827	332	0.28	−0.10	0.69	0.30	1.85	8.27
British Isles								
Guernsey & Jersey	1,467	32	0.29	−0.05	0.00	0.32	2.10	4.18
Isle of Man	591	19	0.57	−0.08	0.06	0.29	1.77	9.61
Republic of Ireland	2,610	110	0.98	−0.03	−0.70	0.24	1.21	28.75
Europe								
Belgium	901	17	−0.10	0.22	−0.13	0.52	5.59	6.77
Denmark & Sweden	601	17	0.59	0.04	−0.40	0.33	2.28	12.09
France	4,478	43	−0.45	−0.12	−0.31	0.33	2.25	8.32
Germany	7,487	80	−0.67	0.23	0.70	0.57	6.84	23.12
Greece	1,400	8	−0.88	−0.09	−0.35	0.57	6.72	26.63
Italy	1,708	39	0.27	−0.08	−0.50	0.35	2.57	6.74
Netherlands	1,492	31	0.06	−0.04	−0.16	0.38	3.04	3.40
Portugal	728	5	−1.14	−0.07	−0.35	0.50	5.20	37.55
Spain	5,433	27	−1.55	−0.03	0.15	0.24	1.22	59.29
Switzerland	801	17	0.17	0.04	−0.13	0.36	2.76	3.63
Other countries								
Hong Kong	2,806	19	−1.16	0.19	−0.31	0.38	3.11	36.65
India	884	3	−1.75	0.04	0.29	0.11	0.27	74.94
New Zealand	2,333	129	1.12	−0.03	0.24	0.24	1.17	31.81
South Africa	2,931	132	0.69	0.01	−0.03	0.31	2.04	13.51
Total 24 countries	85,479	2,425					60.06	504.00
B Residual countries	22,767	160						
C Armed Forces etc	18,400	129						
Total A + B + C	126,646	2,714						
Mean regression coefficient			0.00	0.00	0.00			
Standard deviation			0.90	0.10	0.38	0.36		
Standard deviation × $\sqrt{3}$			1.56	0.17	0.67	0.62		
Analysis of variance of ξ^2			411.24	3.37	29.33	60.06		504.00
Same in percentage terms			81.6	0.7	5.8	11.9		100.0

*Details of the entries in cols. 1 and 2 appear as footnotes to ISR Table 4. In estimating emigrants, the geographical RVPs take the central values of their ranges.

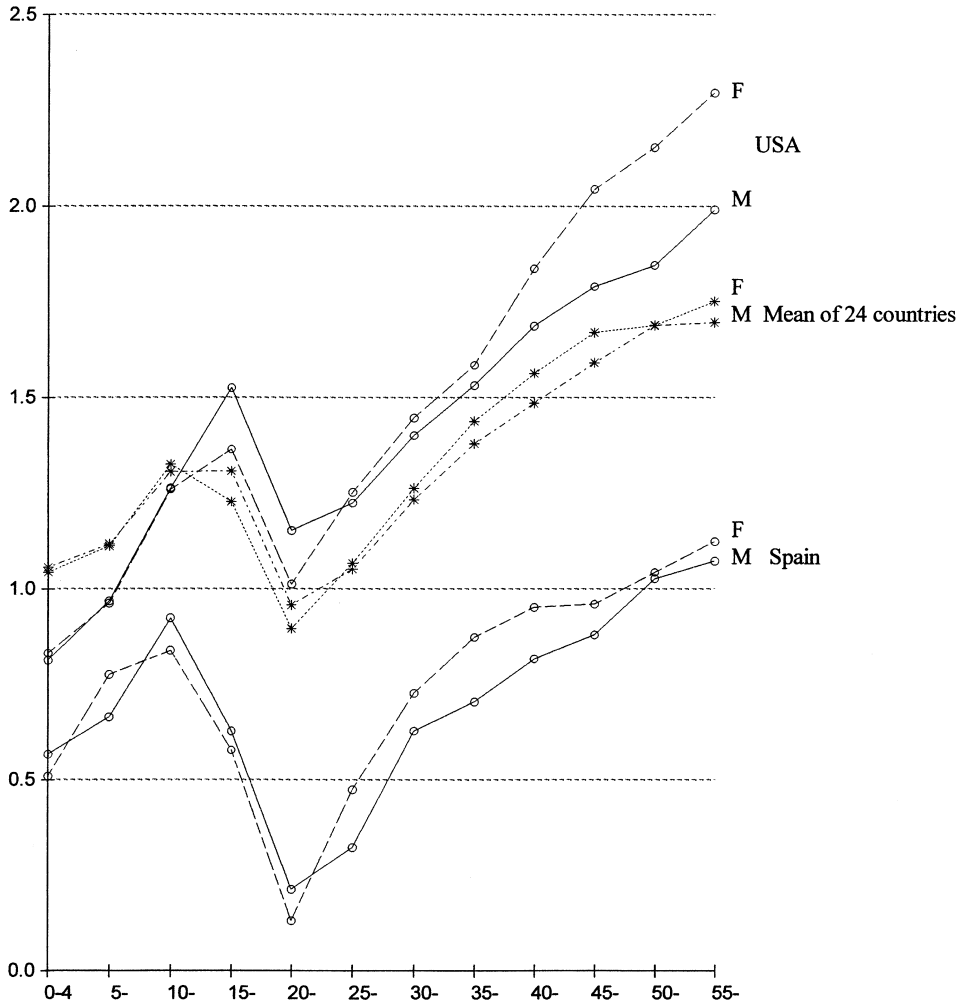

Fig. 1. $\lambda_0(k, x, i)$ *by age-group for Spain and the US, and the mean of 24 countries,* $\overline{\lambda}_0(x, i)$

The novel method was (1) to calculate the ratio of emigrants to returners (call this ratio μ) for each of the 24 countries that had provided emigrant data for the ISR paper (that is, Scotland and Northern Ireland and 22 foreign countries), and then (2) to estimate emigrants in the residual countries taken together by multiplying the number of returners from the residual countries by a value of μ that had been extrapolated from the values in an appropriate subset of the 24 countries. (In the ISR paper the chosen subset comprised 14 countries: 11 in Europe outside the UK plus Hong Kong, India and the U.S.) The procedure was carried out separately for each of the sex-age groups (x, i).

The analysis may be expressed algebraically as follows. Write $C_E(k)$ for the census count of emigrants resident in country k, converted where necessary to the basis ''born in E&W,'' $(k = 1, 2, ..., K;\ K = 24)$; and write $C_E(k, x, i)$ for the component in sex-age group (x, i). Write $R(k, x, i)$ for the number of returners from country k in sex-age group (x, i). Let us treat the residual countries taken together as the 25th, or $(K + 1)$th, country,

Table 5. Estimated numbers of emigrants and of net undercount, England & Wales, ages 0–59, 1991

Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	No. of replications	Emigrants* (000s)				Net undercount in census of E&W			
		Total		In 12 age-groups		Total (000s)		Pc in 24 sex-age gps	
		Mean	Std. devn	RMS of differences of means from A-2	RMS of std. devns	Mean	Std. devn	RMS of differences of means from A-1	RMS of std. devns
ONS est.	..	..	..	..	..	1,048	..	1.69	..
Base A-1	250	..	..	..	..	1,543	199	–	0.66
Base A-2	100	2,787	119	–	12.0	1,546	192	0.02	0.64
P	100	2,685	163	14.5	16.5	1,681	246	0.50	0.80
R3	250	2,699	119	14.1	12.8	1,681	191	0.52	0.67
R1	100	2,716	109	18.6	13.1	1,656	189	0.76	0.71
R0	150	2,715	163	18.8	22.1	1,652	237	0.94	0.99
Variants									
R3-1	50	2,641	114	17.8	12.6	1,706	182	0.56	0.64
R3-2	50	2,709	107	14.2	11.5	1,660	156	0.46	0.62
R3-3	50	2,675	123	15.6	12.7	1,726	214	0.61	0.73
R3-4	50	2,723	114	12.4	13.3	1,680	186	0.52	0.70
R3-5	50	2,747	111	11.5	11.9	1,635	205	0.46	0.67
R0-1	50	2,605	136	24.4	20.2	1,757	225	1.02	0.95
R0-2	50	2,697	131	19.2	20.4	1,660	232	0.90	1.04
R0-3	50	2,843	126	20.1	18.7	1,540	206	1.00	0.87
R0-NZ	100	2,631	162	27.5	24.1	1,759	219	0.94	1.00
R0-v2	100	2,676	144	19.6	22.5	1,724	255	1.08	1.13

*Including an estimate of undercount in censuses outside E&W.

Note. The following models have been built up in sets of 50 replications. In a given set the proxy total of emigrants in country k in 1981 is taken to be the 1991 total *times* the factor shown below.

R3 is made up of 5 sets: R3-1 1/1.1; R3-2 1/1.05; R3-3 1; R3-4 1.05; R3-5 1.1.
R1 is made up of 2 sets: R1-1 1/1.05; R1-2 1.05.
R0 is made up of 3 sets: R0-1 1/1.1; R0-2 1; R0-3 1.1.
R0-NZ and R0-v2 are each made up of 2 sets on the pattern of model R1.

so that $R(K+1,x,i)$ is the number of returners from the residual countries in sex-age group (x,i). Write:

$$\mu(k,x,i) = C_E(k,x,i)/R(k,x,i) \quad (k = 1,2,\ldots,K) \tag{7}$$

The value of the ratio μ varies widely, so it is appropriate to work with $\log \mu$ (to the base 10) $= \lambda$. Write $\lambda(k,x,i) = \log \mu(k,x,i)$. Because the conversion of $C_E(k)$ to the basis ''born in E&W'' involves uncertainties represented by ''geographical RVPs'' (ISR, p. 299 at [5]), write λ_0 and μ_0 for the values taken by λ and μ when the geographical RVPs are set at the midpoints of the ranges assigned to these RVPs.

Figure 1 plots $\lambda_0(k,x,i)$ against age-group i, separately for males and females, for two countries – the U.S. and Spain – and also shows the unweighted mean of $\lambda_0(k,x,i)$ among

the 24 countries, labeled $\overline{\lambda}_0(x, i)$. The figure demonstrates (i) that the curves have a rough cubic shape with a maximum in the age-range 10–20 followed by a minimum in the early 20s; and (ii) that there are wide variations between the curves for different countries.

The next step was to extrapolate from the known values of the λs in the subset of 14 countries to give postulated values of λ for the residual countries taken together, $\lambda(K + 1, x, i)$. To do this, it was necessary to quantify the pattern of the λs, that is, the extent to which the profile of λ_0 (as in Fig. 1) varied between one country and another among the 14 in terms of its absolute level, the sex differential, age differential, etc. ISR Section 5 describes the method used to establish this pattern and gives a formula for $\lambda(K + 1, x, i)$ that embodies RVPs to reflect the errors and uncertainties. (The models developed in this article employ similar methods to establish the pattern of the λs among the 24 countries, and then to give postulated values of the λs for many of the 22 foreign countries as well as for the residual countries taken together – see Sections 8 and 9 below.)

Numbers of emigrants in the residual countries, $C_E(K + 1, x, i)$, were then estimated as follows.

$$\mu(K + 1, x, i) = 10^{\lambda(K+1,x,i)} \tag{8}$$

$$C_E(K + 1, x, i) = R(K + 1, x, i) \times \mu(K + 1, x, i) \tag{9}$$

The $C_E(K + 1, x, i)$ are proxies for data that might have been collected in censuses held in the residual countries.

6. ISR Model (Base Model A): Results

Table 5 summarises estimates of emigrants as at mid-1991 and of net undercount in the 1991 Census of E&W, as generated by different models. The estimates are restricted to persons aged under 60. The 2nd line of the table refers to the original 250 replications of ISR Base Model A labelled A-1 (for which emigrant numbers are not available); and the 3rd line to a further 100 replications A-2. Cols. 2 and 3 show the mean and standard deviation of the frequency distribution of numbers of emigrants (in thousands); these figures include an allowance for undercount in censuses outside E&W. Cols. 6 and 7 show the mean and standard deviation of the distribution of net undercount (in thousands). The estimates are subject to sampling errors because they are based on a limited number of valid replications (Col. 1). A full statement of the ISR results in the form of a Demographic Account of births, deaths, emigrants, immigrants and total population at mid-1991 appears in Redfern (2004, Appendix A, Tables 4 and 5). This is based on a further 200 valid replications of Base Model A – which we may label A-3.

For Base Model A-2 the age profile of numbers of emigrants (males and females combined) is shown in Fig. 2A as the pair of dotted lines linking the asterisks. The lower of the two lines is the 1st percentile of the distribution and the upper is the 99th percentile. The width of the band bounded by the two lines is a measure of the uncertainty of the profile. The standard deviations of the distributions of numbers of emigrants in each of the 12 age-groups are summarised as a Root Mean Square (RMS) value in Table 5, Col. 5 (12.0 thousand). This is an alternative measure of the uncertainty.

For Base Model A-1 the age profile of percentage net undercount, $u(x, i)$, is shown in Figs 4A (males) and 4B (females) as the pair of dotted lines linking the asterisks. As

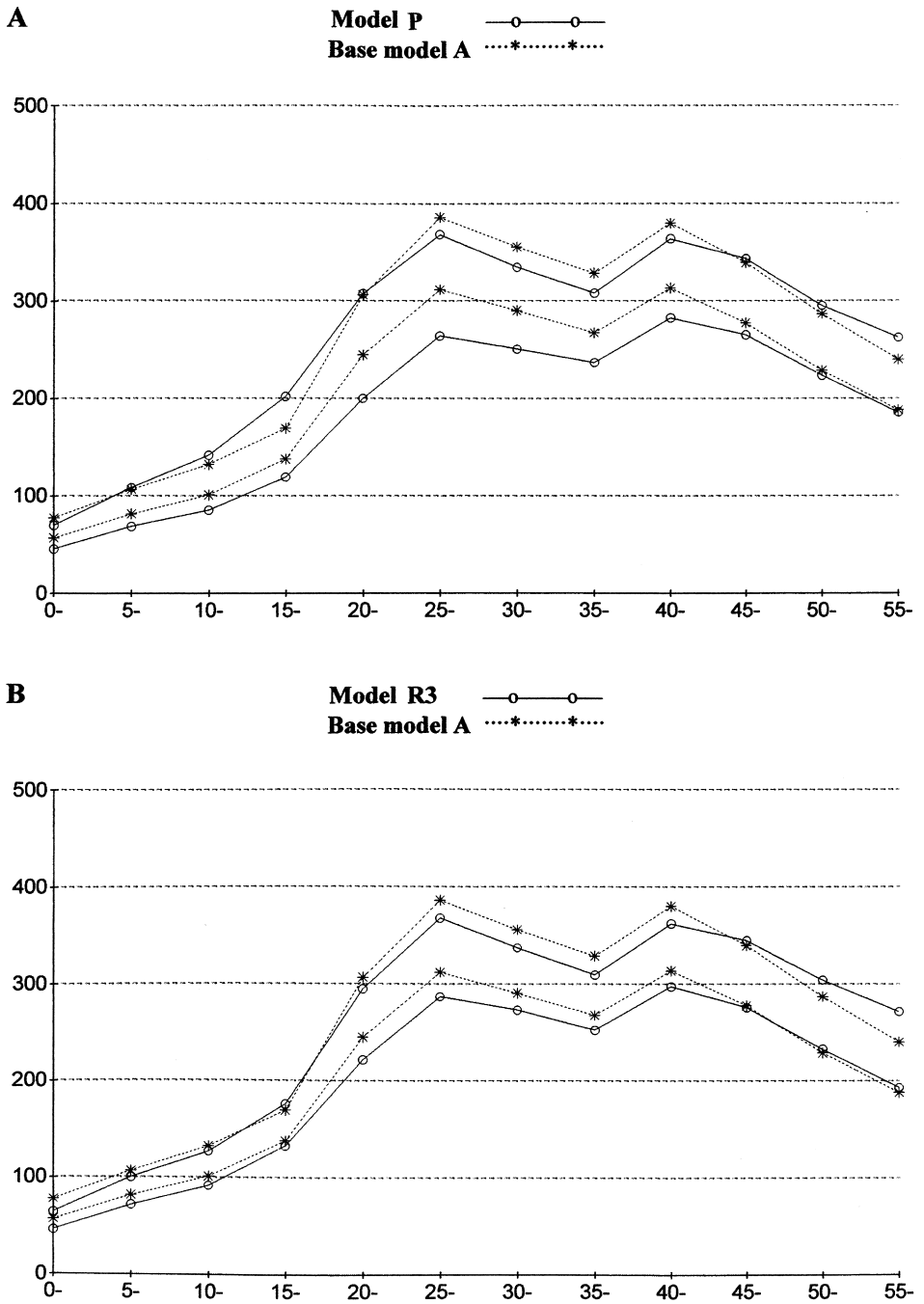

Fig. 2. Emigrants by age-group
E&W 1991 (thousands)
The two lines for each model are the 1st and 99th percentiles of the distributions.

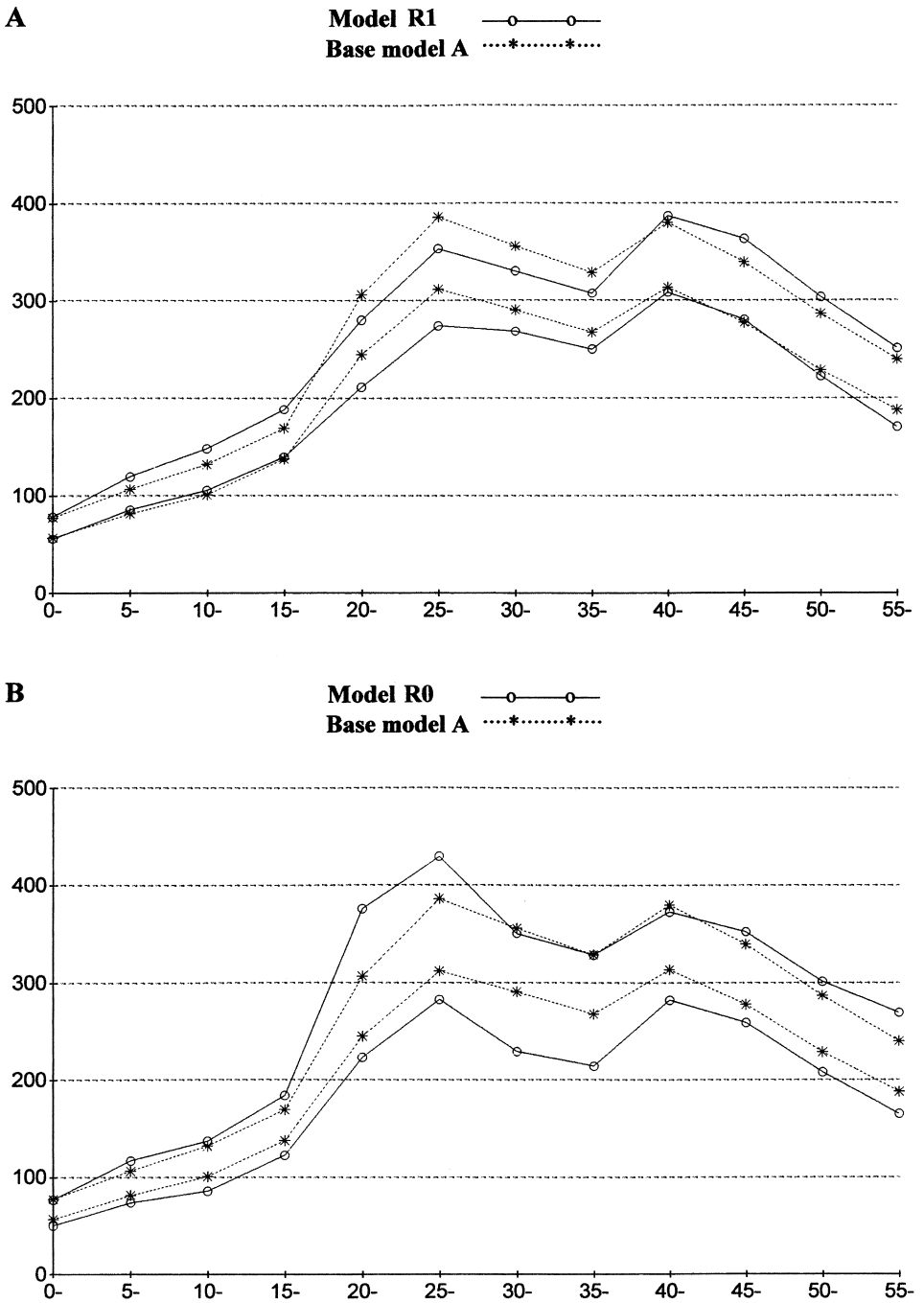

Fig. 3. Emigrants by age-group
E&W 1991 (thousands)
The two lines for each model are the 1st and 99th percentiles of the distributions.

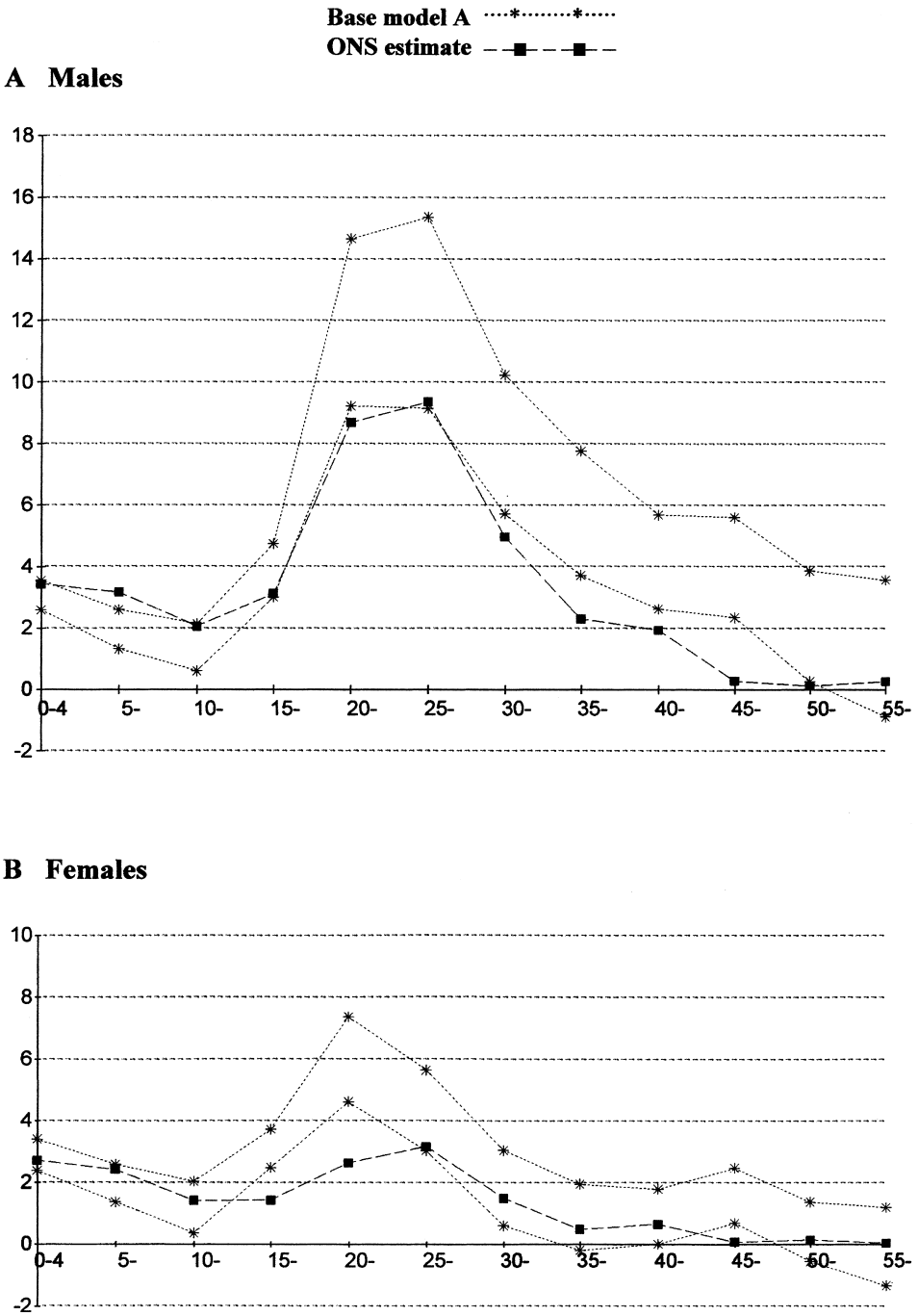

Fig. 4. Percentage net undercount by age-group, E&W 1991
The two lines for Base model A are the 1st and 99th percentiles of the distributions.

Table 6. Numbers of foreign countries whose census data have been used in the models

	No. of foreign countries whose contemporaneous census data have been used	No. of foreign countries whose earlier (1980–81) census data have been used	Of emigrants in foreign countries, the percentage for whom estimates have been made using data on returners
Base model A	22	–	7*
Model P	3 (Australia, Canada, U.S.)	–	40
Model R3	3 (Australia, Canada, U.S.)	19	40
Model R1	1 (Australia)	21	69
Model R0	–	22	100

*Refers to emigrants in the 200 or so ''residual'' countries.

before, the lower of the two lines is the 1st percentile of the distribution and the upper is the 99th percentile. Again, the width of the band bounded by the two lines is a measure of the uncertainty of the profile. Table 5, Col. 9, shows the RMS value of the standard deviations of the distributions of percentage net undercount in each of the 24 sex-age groups (0.66 percentage points); it is an alternative measure of the uncertainty.

The robustness of the ISR model was examined in ISR Section 8. This gave results from a variant model in which the 3 constraints specified at inequalities (4), (5) and (6) above were relaxed and from two variant models in which the ranges of values assigned to the RVPs were widened.

In Figs 4A and 4B the dashed lines linking the square point markers represent the official estimates of 1991 undercount as originally published by the Office of National Statistics (ONS). For several sex-age groups the ONS estimates lie outside the band bounded by the ISR 1st and 99th percentiles. Other measures of the differences between the ONS and ISR estimates are given in Table 5, Cols. 6 and 8. Thus, the ONS estimate, 1,048 thousand (Col. 6), is well below the mean of the the A-1 distribution (1,543 thousand). The figure of 1.69 percentage points in Col. 8, line 1 is the RMS value, taken over the 24 sex-age groups, of the difference between the official estimate of percentage undercount in sex-age group (x, i) and the mean of the distribution of the A-1 estimate $u(x, i)$.

Two factors contribute to the differences between the Model A estimates and the ONS estimates. The first is errors in the data on migrant flows in the decennium 1981–1991 that are a component of the official estimates. The second factor is one of definition. The ISR estimates of undercount include (in principle) all persons who had previously been resident in E&W but had no usual residence at the time of the 1991 Census. I call these people wanderers; examples are seamen and round-the-world backpackers. Census-takers at home and abroad would exclude most wanderers from their counts of the resident population. We have to ask ourselves, therefore, whether the difference between the A estimates and the ONS estimates is a credible estimate of wanderers. The difference has a mean value of 495 thousand (Table 5, Col. 6), with a 1st percentile of 96 thousand and a 99th percentile of 980 thousand. There are virtually no independent data on wanderers to

Table 7. The mean, $\overline{\lambda}_0(x,i)$, and standard deviation, $\sigma_0(x,i)$, of $\lambda_0(k,x,i)$ among 24 countries, 1991

Ages	(1) Mean*	(2)	(3) Standard deviation*	(4)	(5)
	Males	Females	Males	Females	Postulated M&F
0–	1.06	1.04	0.26	0.25	0.255
5–	1.12	1.11	0.30	0.30	0.298
10–	1.31	1.33	0.34	0.34	0.342
15–	1.31	1.23	0.38	0.39	0.385
20–	0.96	0.90	0.41	0.38	0.398
25–	1.05	1.07	0.41	0.36	0.410
30–	1.23	1.26	0.38	0.40	0.423
35–	1.38	1.44	0.43	0.41	0.435
40–	1.49	1.56	0.43	0.41	0.448
45–	1.59	1.67	0.46	0.46	0.460
50–	1.69	1.69	0.49	0.48	0.473
55–	1.70	1.75	0.48	0.48	0.485

*unweighted

provide guidance, but a figure near to the 1st percentile would seem the most plausible. Put another way, the ONS estimate of undercount is probably too low, and certainly so in some sex-age groups, notably women aged 20–24 (see Fig. 4B).

7. The New Models P and R

All the new models presented here have been applied to the 1991 Census of E&W and they differ from the ISR model (Base Model A) only in the calculations of emigrants in foreign countries. All require emigrant data from many fewer contemporaneous foreign censuses than were used in the ISR model, as Table 6 shows; in this context contemporaneous means ''of the 1990–1991 round of censuses.''

8. Establishing the Pattern of the λs

In order to extrapolate λ, that is, infer values of λ for countries $A, B, \ldots, Q$ (say) from known values for countries $R, S, \ldots, Z$ (say), we need an understanding of the variability or pattern of the λs among the whole set A to Z. So, in an exercise to estimate undercount in the 1991 Census of E&W, I ought to have established the pattern of the λs from analysis of an earlier round of censuses, for example 1981, and then used that pattern, with or without modification, in the 1991 model. But I lacked the 1981 data and had to use, as a proxy, the pattern revealed by the 1991 data. Sections 15 and 16 discuss whether changes in the pattern of the λs over a 10-year period might invalidate my conclusions.

The pattern of the λs in 1991 has been analysed by fitting a regression to the values of $\lambda_0(k,x,i)$ for country k, repeating this for each of the 24 countries k whose emigrant data contributed to the ISR model. (Strictly I should refer to 22 rather than 24 countries because, as the 22 entries under head A of Table 4 show, in two cases (Guernsey and

Jersey, Denmark and Sweden) data for two countries have been combined. But for simplicity I shall speak of 24 countries.) The regression equation is:

$$\lambda_0(k,x,i) = \overline{\lambda}_0(x,i) + \sigma_0(x,i) \cdot [f_1(k) + f_2(k) \cdot s + f_3(k) \cdot a + \varepsilon(k,x,i)] \quad (k = 1,2,\ldots,24) \tag{10}$$

where $\overline{\lambda}_0(x,i)$ and $\sigma_0(x,i)$ are the unweighted mean and standard deviation of $\lambda_0(k,x,i)$ among the 24 countries; s and a are the sex and age variables; $f_1(k)$, $f_2(k)$ and $f_3(k)$ are regression coefficients; and $\varepsilon(k,x,i)$ is the residual term in the regression. $f_1(k)$ indicates the extent by which the values of $\lambda_0(k,x,i)$ in country k differ in absolute level from the values averaged over all the 24 countries. $f_2(k)$ indicates how the values of $\lambda_0(k,x,i)$ in country k differ in terms of sex differentials from the average in the 24 countries. $f_3(k)$ describes a country's age differentials in a similar way.

The values of the mean $\overline{\lambda}_0(x,i)$ and the standard deviation $\sigma_0(x,i)$ are shown in Table 7. For each country k, the values of the regression coefficients $f_1(k)$, $f_2(k)$ and $f_3(k)$ are shown in the top part of Table 4, Cols. 3 to 5; the RMS value of the residual ε in Col. 6; and, in Cols. 7 and 8, the sum over the 24 sex-age groups (x,i) of $[\varepsilon(k,x,i)]^2$ and $[\xi(k,x,i)]^2$ respectively, where:

$$\xi(k,x,i) = [\lambda_0(k,x,i) - \overline{\lambda}_0(x,i)]/\sigma_0(x,i) \tag{11}$$

The bottom part of Table 4 shows (i) the standard deviation among the 24 countries k of the regression coefficients $f_1(k)$, $f_2(k)$ and $f_3(k)$ and also the standard deviation of the residual element $\varepsilon(k,x,i)$; (ii) (in the next line) the same set of standard deviations multiplied by $\sqrt{3}$ (referred to in Section 9 below); and (iii) the analysis of variance of the variable $\xi(k,x,i)$ in absolute and percentage terms. This last shows that the factors $f_1(k)$, $f_2(k)$ and $f_3(k)$ explain respectively 81, 1 and 6 per cent of the total variance, and that the remaining 12 per cent is accounted for by the residual $\varepsilon(k,x,i)$.

9. Estimating Emigrants in Model P

Model P uses data on emigrants from the censuses of two countries within the UK – Scotland, referenced as [1], and Northern Ireland [2] – and three countries outside the UK – Australia [3], Canada [4] and the U.S. [5]. As Table 4, Cols. 1 and 2 show, these countries (Northern Ireland excepted) head the list in terms of the numbers of emigrants and/or numbers of returners. Emigrants must now be estimated in the other 19 countries [referenced as $6,7,\ldots,K$ where $K = 24$] and in the residual countries taken together (referenced as $(K+1)$), using the figures of returners, $R(k,x,i)$, as indicators.

As an approximation we may write for all the K countries:

$$\lambda(k,x,i) = \overline{\lambda}(x,i) + \sigma(x,i) \cdot [f_1(k) + f_2(k) \cdot s + f_3(k) \cdot a] \quad (k = 1,2,\ldots,K) \tag{12}$$

Equation (12) is derived from Equation (10) by omitting the residual element ε (which accounts for only a small part of total variance) and generalising the equation to a small degree by substituting λ and σ in place of λ_0 and σ_0. From Equation (12) we have

$$\begin{aligned}\lambda(k,x,i) = {} & [\lambda(1,x,i) + \ldots + \lambda(5,x,i)]/5 + \sigma(x,i) \cdot \{\{f_1(k) - [f_1(1) + \ldots + f_1(5)]/5\} \\ & + s \cdot \{f_2(k) - [f_2(1) + \ldots + f_2(5)]/5\} \\ & + a \cdot \{f_3(k) - [f_3(1) + \ldots + f_3(5)]/5\}\} \quad (k = 6,\ldots,K)\end{aligned} \tag{13}$$

This apparently complicated equation has a simple interpretation. It states that an estimate of $\lambda(k,x,i)$ for one of the $(K-5)$ $(=19)$ countries for which emigrant data are not available, say Belgium, is the unweighted average of the corresponding observed values of $\lambda(k,x,i)$ for the 5 countries (Scotland, etc) – call this the starting sex-age profile $\Lambda(x,i)$, or simply Λ – adjusted for the differences between the country-specific factors f_1, f_2 and f_3 for Belgium and the unweighted average of the factors f_1, f_2 and f_3 for the 5 countries.

In applying Formula (13), some limited smoothing was done to the starting sex-age profile $\Lambda(x,i)$ to iron out irregularities. Next, values must be given to $\sigma(x,i)$ for all 24 sex-age groups. I have taken the values in Table 7, Col. 5; they are a rough approximation to the values observed in 1991 (Cols. 3 and 4) and are a proxy for 1981 values. Values must also be given to the country-specific factors f_1, f_2 and f_3 for each of the 24 countries. Note that the values of f_1, f_2 and f_3 for the 5 countries providing emigrant data are unknown, as of course are the values for the other 19 countries. The values of f_1, f_2 and f_3 for any country k $(k = 1, 2, ..., K)$ are therefore represented by 3 RVPs, which are assumed to have rectangular distributions within the limits ± 1.56, ± 0.17 and ± 0.67, respectively; these limits are taken from the line headed ''Standard deviation $\times\sqrt{3}$'' near the foot of Table 4 and reflect the fact that the standard deviation of a rectangular distribution in the range $\pm x$ is $x/\sqrt{3}$. The $3K$ RVPs for f_1, f_2 and f_3 for the K countries are assumed to be distributed independently of one another except that an approximate adjustment was made to Equation (13) to allow for correlation among the 5 components of the sum $[f_1(1) + \ldots + f_1(5)]$, and similarly among the components of $[f_2(1) + \ldots]$ and among the components of $[f_3(1) + \ldots]$.

For the residual countries taken together, treated as the $(K+1)$th country, the formula at (13) has again been applied. But the limits of $f_1(K+1)$ have been set at $(-2.5, 0)$ on grounds of the kind discussed in ISR, Section 5, p. 290, and the limits for $f_2(K+1)$ and $f_3(K+1)$ have been set at ± 0.17 and ± 0.67 respectively as for the 24 countries.

From the estimated values of $\lambda(k,x,i)$ for the 19 countries $(k = 6, \ldots, K)$, and for the residual countries taken together, and the corresponding census figures of returners, $R(k,x,i)$, an estimate of emigrants resident in each country k, $C_E(k,x,i)$, is given by the equation:

$$C_E(k,x,i) = R(k,x,i) \times 10^{\lambda(k,x,i)} \quad (k = 6, ..., K, K+1) \tag{14}$$

10. Model P: Results

In Model P the mean of total emigrants is some 100 thousand lower than in Base Model A (Table 5, Col. 2), counterbalanced by net undercount that is some 140 thousand higher (Col. 6). Not surprisingly, the frequency distributions of the Model P estimates are about 30 per cent wider than those of Base model A (Cols. 3, 5, 7 and 9).

The Model P age profile of emigrants appears in Fig. 2A as the pair of solid lines linking the circles; these lines are the 1st and 99th percentiles of the distribution. The profile follows closely the Base Model A profile (the dotted lines) though dipping a little below at ages 20–44. The divergence between the two profiles is measured in another way in Table 5, Col. 4; this gives the RMS value, taken over the 12 age-groups i, of the difference

between the mean of the Model P estimates of emigrants in age-group i and the corresponding Base Model A-2 mean: namely 14.5 thousand. In a similar way Col. 8 measures the divergence between the Model P profiles of percentage undercount (not illustrated) and the Base Model A-1 profiles: an RMS value taken over the 24 sex-age groups of 0.50 percentage points.

11. Models R3 and R1: Description

In Model P the estimate of total emigrants, $C_E(k)$, in any of the 19 countries k ($k = 6, 7, \ldots, 24$) varies very widely; thus switching the value of $f_1(k)$ from one end of its assumed range (-1.56) to the other ($+1.56$) multiplies emigrants by a factor of about 22. As a result only about 1 in 50 of the replications satisfies the three constraints of Section 4. (Such a low yield of valid replications puts a heavy burden of computing on a researcher working with a basic PC and software of 1990 vintage!)

In Model R3 the method of estimating emigrants in the 19 countries k is varied: I place limits on the range of values that $C_E(k)$ can take. The aim is to make $C_E(k)$ equal to the total of emigrants recorded in country k's census a decade earlier, that is in 1980–1981, multiplied by two factors: (1) Π, a factor that is the same for each country k and lies in the range 1/1.1 to 1.1 (representing a percentage change in the range (-9.1, $+10$)); (2) a second factor $\Omega(k)$ that varies from country to country and lies in the range 1/1.3 to 1.3 (a percentage change in the range (-23.1, $+30$)). The limits placed on these factors should encompass all or most changes in emigrant numbers over the decade to 1991. Thus, the 1980–1981 total of emigrants in country k is treated as an indicator of the 1991 total; it supplements the use of numbers of returners, $R(k, x, i)$, as indicators of emigrant numbers.

Implementation of this is approximate. It is achieved by replacing the term $\{f_1(k) - [f_1(1) + \ldots + f_1(5)]/5\}$ in equation (13) by the term $\{\pi + \omega(k)\}$, where π and the set $\omega(6), \ldots, \omega(24)$ are RVPs distributed independently of one another within suitably chosen limits. The rest of equation (13) is unchanged. For the residual countries we have of course no 1981 figures of emigrants as an indicator for setting 1991 limits, but these limits can be narrowed, as compared with model P, by reference to the narrower limits imposed on emigrant numbers in the 19 countries.

I faced a practical difficulty in implementation: most of the foreign census data for 1981 were not available to me. This lack was however turned to advantage. Any estimate of 1991 emigrants and undercount ought to be invariant with respect to figures of 1981 emigrants. Would my model demonstrate such invariance? Section 15 examines this point.

Model R1 is constructed on the same principles as R3. It uses emigrant data from only one contemporaneous foreign census: Australia's. The starting sex-age profile Λ is now the average of the observed λs of Scotland, Northern Ireland and Australia with, again, limited smoothing.

12. Models R3 and R1: Results

The first point to note is that the ratio of all replications to valid replications – about 4:1 for R3 and 14:1 for R1 – is much lower than the 50:1 experienced with Model P.

The means of the estimates of emigrants and undercount differ little between the R3, R1

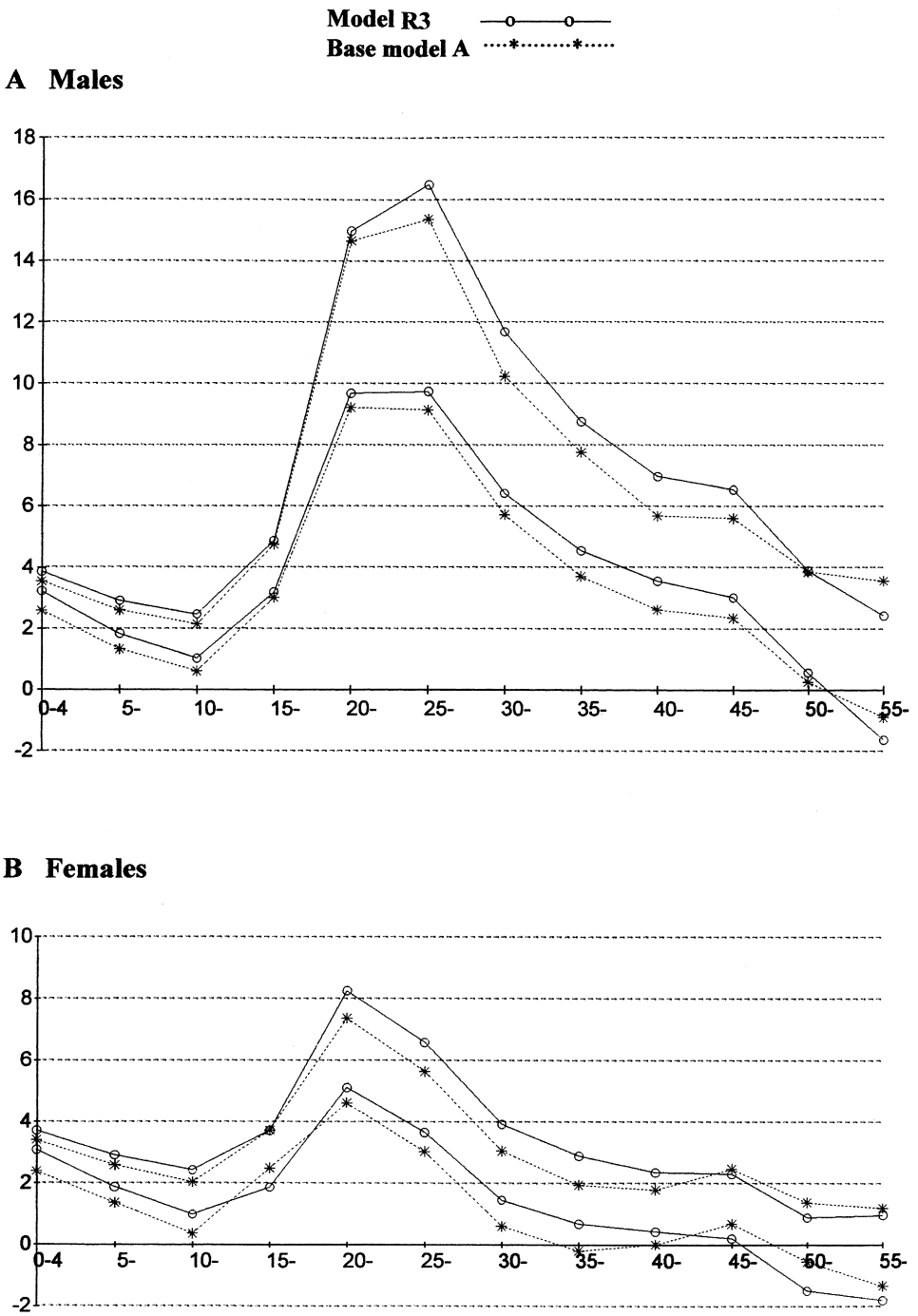

Fig. 5. Percentage net undercount by age-group, E&W 1991
The two lines for each model for the 1st and 99th percentiles of the distributions.

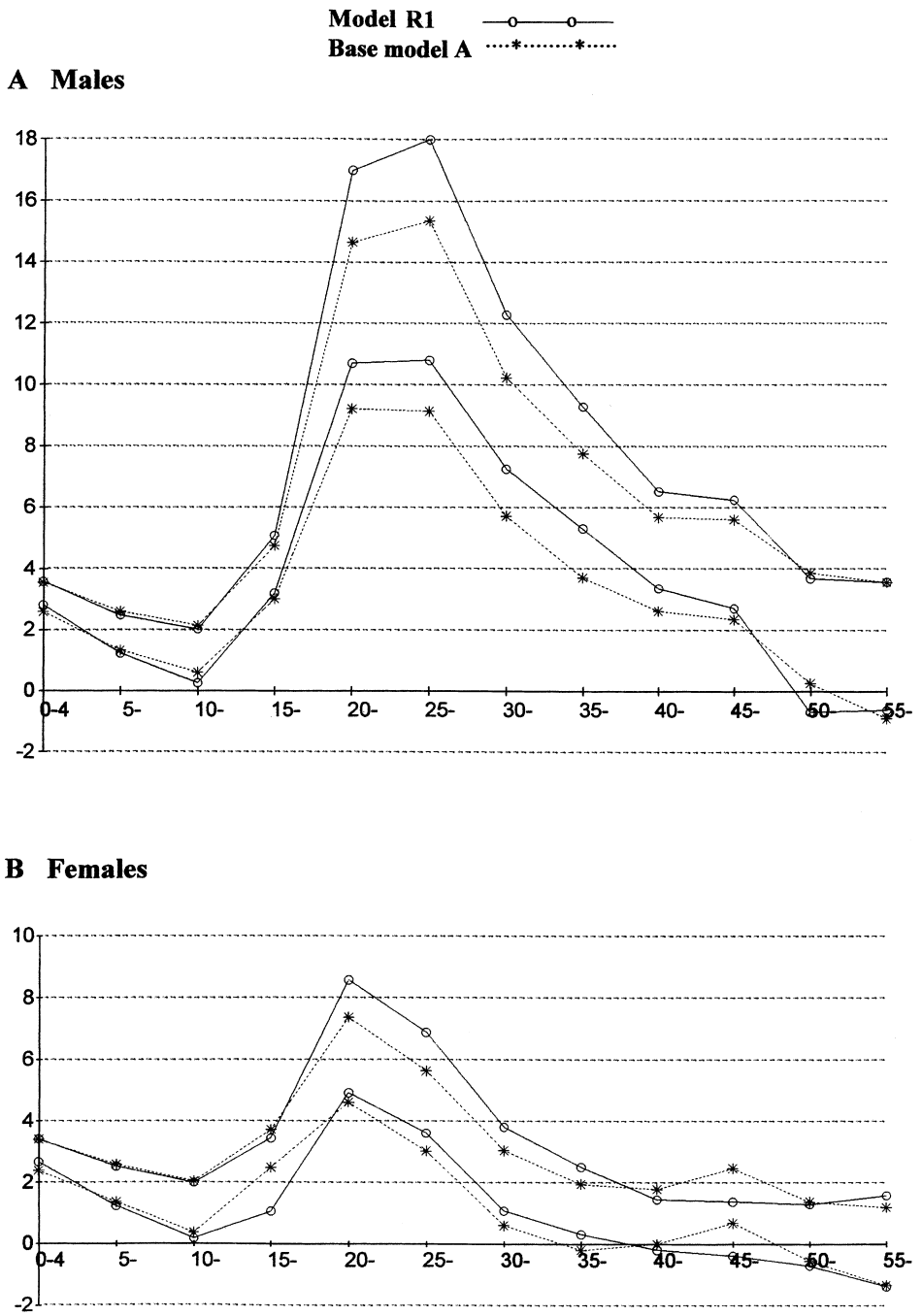

Fig. 6. Percentage net undercount by age-group, E&W 1991
The two lines for each model are the 1st and 99th percentiles of the distributions.

and P Models (Table 5, Cols. 2 and 6). But the R1 profiles of emigrants and percentage undercount mirror the Base Model A profiles less well than do the P and R3 profiles (Cols. 4 and 8). The R Models use more information than the P Model (namely, 1981 data on emigrants) and this is reflected in the reduced widths of their frequency distributions which are not much different from the A widths (Cols. 3, 5, 7, and 9).

The age profiles of emigrants for the R3 and R1 Models are plotted by solid lines in Figs 2B and 3A respectively. The profiles of percentage undercount are in Figs 5 and 6. In each case the dotted lines show the A profiles for comparison. There is a satisfactory degree of similarity between the R and A profiles, with a good degree of overlap between their respective bands.

13. Model R0: Description

The final stage was to attempt a Model R0 that used no data on emigrants extracted from a contemporaneous foreign census. For this model, the data for Scotland and Northern Ireland were combined (as though they were a single country) to provide the starting sex-age profile Λ; this was again smoothed. The method of estimating emigrants employed in Model R3 for 19 foreign countries and then in R1 for 21 countries would now be extended to the 22nd: Australia. But this did not work without substantial modifications to the model. In 1991 more British emigrants were recorded in Australia than in any other country (Table 4, Col. 2) and there were more returners from Australia than from any other country outside the UK (Col. 1). This meant that, though the residual element $\varepsilon(k, x, i)$ of Equation (10) could be ignored in formulating Equation (12) as applied to 21 countries, ε could not be ignored in estimating emigrants in Australia. I found that, in tests made before introducing ε terms for Australia, a substantial proportion of valid Model R0 replications exhibited a bimodal age profile of undercount for females, with one mode at ages in the 20s and a second mode in the 30s. This was implausible on the evidence of other countries – as also of the official estimates of undercount in our 1991 Census.

For Model R0 therefore, a further change was made to the right hand side of Formula (13) in the case of Australia only ($k = 3$): namely, to add the term $\sigma(x, i) \times \varepsilon(3, \cdot, i)$ but only for the age-groups for which ε had a major impact, namely $i = 4$ to 8 inclusive (ages 15–39). The element $\varepsilon(3, \cdot, i)$ denotes an RVP that takes the same value for females as for males; limits were set on the basis of the analysis described in Section 8: ± 0.85 for $i = 4, 5$ and ± 0.6 for $i = 6, 7, 8$.

Simultaneously I introduced new a priori constraints on the profile of percentage net undercount to supplement the three existing constraints in Section 4. The aim was to reject replications with: (1) a bimodal profile; or (2) very negative undercount at higher ages. The chosen constraints were:

$$u(x, 4) > u(x, 3) \qquad (x = 1, 2) \tag{15}$$

$$u(x, 5) > \tfrac{3}{4} u(x, 6) \quad (x = 1, 2) \tag{16}$$

$$u(x, 6) > u(x, 7) \qquad (x = 1, 2) \tag{17}$$

$$u(x, 7) > u(x, 8) \qquad (x = 1, 2) \tag{18}$$

$$u(x, i) > -0.02 \qquad (x = 1, 2; i = 9, \ldots, 12) \tag{19}$$

These inequalities state that for each sex: (15) the rate of net undercount at ages 15–19 exceeds the rate at 10–14; (16) the rate at 20–24 exceeds $\frac{3}{4}$ times the rate at 25–29; (17) and (18) the rate decreases as age increases through the range 25–39; (19) any net overcount at ages 40 and over cannot exceed 2 per cent.

Probably too many new a priori constraints were introduced – a case of ''overkill.'' Thus, all or nearly all of the replications that satisfied the three original constraints also satisfied certain of the new constraints. Among the new constraints, the one which invalidated the greatest number of replications was the female version of inequality (17): this inhibited a bimodal female profile.

14. Model R0: Results

A substantial computing effort was needed to generate results because the ratio of all replications to valid replications was of the order of 400:1.

The R0 means of emigrants and undercount are little different from the values generated by R3 and R1 (Table 5, Cols. 2 and 6). But the R0 profile of mean percentage undercount follows the A profile less closely than R3 and R1 do (Col. 8). And, not surprisingly, the standard deviations of the R0 distributions are between 25 and 75 per cent greater than those of R3 and R1 (Cols. 3, 5, 7, and 9). Some degeneration of the profiles is also seen in Fig. 3B (emigrants) and Fig. 7 (undercount), mainly at ages 20–44, though there remains a substantial area of overlap between the R0 and A bands. The profile of female undercount is perhaps the weakest part of the R0 analysis; it is not very informative to estimate undercount among women aged 20–24 as between 1.3 and 6.3 per cent.

15. Tests of Robustness

To examine robustness, I varied some (but not all) of the main elements of Models R3 and R0. The limits of π and $\omega(k)$ (see Section 11, 3rd paragraph) were not varied because they were already generously wide.

Models R0–1, R0–2 and R0–3

These models test whether the estimates of 1991 emigrants and undercount are invariant with respect to the (proxy) figures of 1981 emigrants that have been entered into the models. Each of the three models comprises 50 replications which, taken together, constitute the 150 replications of Model R0 (see the note at the foot of Table 5). The proxy figure of total 1981 emigrants in the 22 countries is 400 thousand higher in R0–3 than in R0–1. The results (Table 5, Col. 2) show that the mean estimate of emigrants in 1991 is 238 thousand higher in R0–3 than in R0–1. To state that in a different way, Model R0 understates the change in emigrant numbers between 1981 and 1991. There are counterbalancing differences in the estimates of mean undercount, which are 217 thousand lower in R0–3 than in R0–1 (Col. 6).

Models R3–1 to R3–5

These models provide another test of the impact of 1981 emigrant numbers on the estimates of emigrants and undercount in 1991. Each of the five models comprises 50

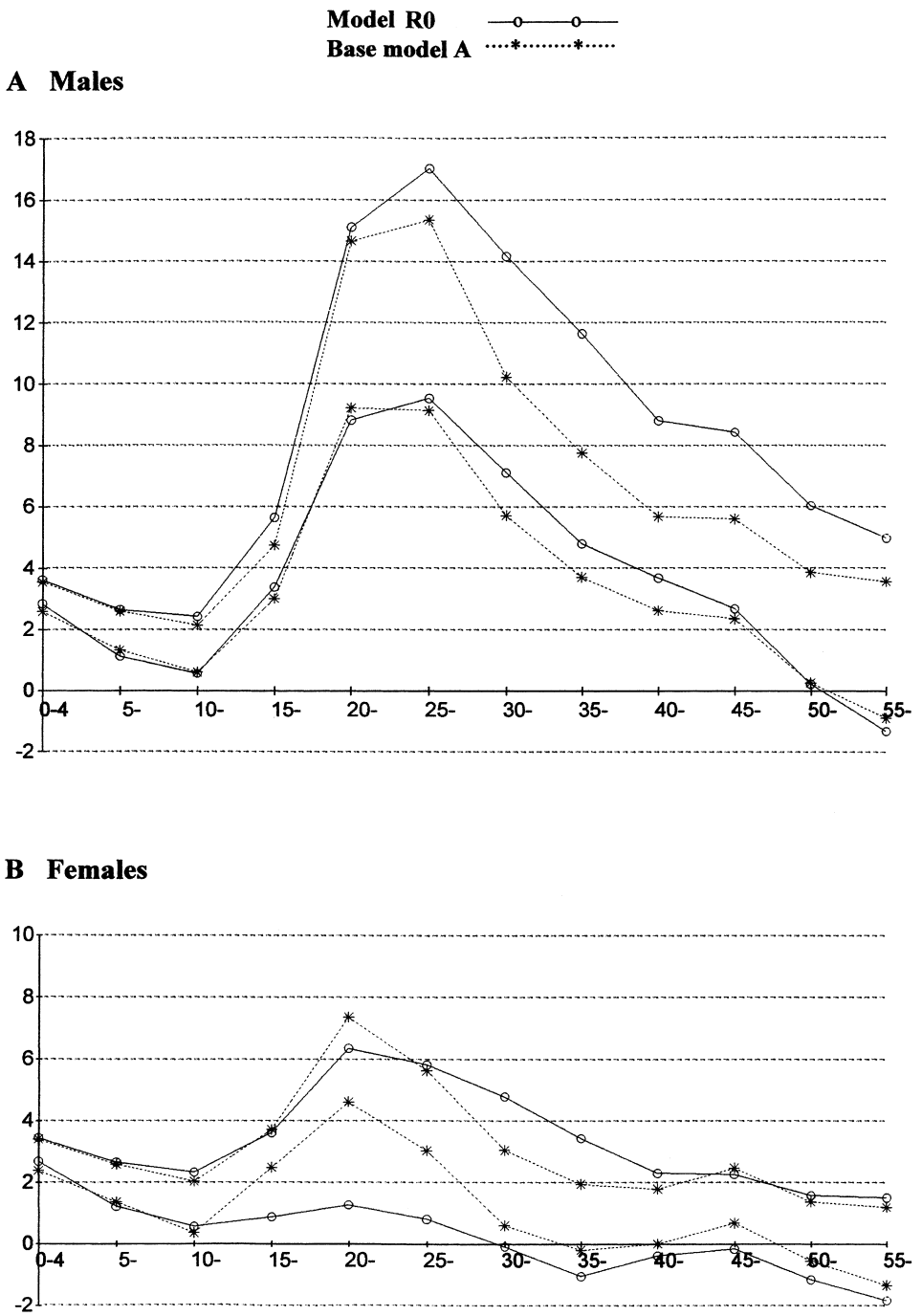

Fig. 7. Percentage net undercount by age-group, E&W 1991
The two lines for each model are the 1st and 99th percentiles of the distributions.

replications which, taken together, constitute the 250 replications of Model R3 (see the note to Table 5). The proxy figure of total 1981 emigrants in 19 countries is 140 thousand higher in R3–5 than in R3–1. The impact on 1991 estimates is substantially less than in Model R0.

Model R0-NZ

How far do the results of Model R0 reflect the use of data from Scotland and Northern Ireland to provide the starting sex-age profile Λ? To test this, a hypothetical Model R0-NZ has been created in which Λ takes the New Zealand values of $\lambda(k, x, i)$ (from Model A). As compared with Model R0 the main differences in the results are: a lower estimate of emigrants (Table 5, Col. 2); undercount that is some 100 thousand higher (Col. 6); and a profile of emigrants that mirrors the A profile less well (Col. 4).

Model R0-v2

The results of applying Equation (13) depend on the ranges assigned to the RVPs f_1, f_2 and f_3 for each country (and, in the case of Australia only, to the RVPs ε) and on the 24 values given to the $\sigma(x, i)$. These parameters might change over time, so that a model that used f-ranges and σs observed in the previous census might be flawed. Model R0-v2 tests the combined impact of varying many of the parameters. This is done by giving the variance $[\sigma(x, i)]^2$ values double those used in Model R0; that is, the values of the standard deviation $\sigma(x, i)$ are $\sqrt{2}$ times the values in Table 7, Col. 5. The resulting mean estimate of total undercount is 72 thousand greater than in Model R0 (Table 5, Col. 6).

16. Conclusions on Robustness

Table 5 presents 7 different models and, within two of them (R3 and R0), variant assumptions about the proxy numbers of emigrants in 1981. The lowest of the figures of mean undercount in Col. 6 is 1,540 thousand and the highest 1,759 thousand; this is a range of 219 thousand, which is of the same magnitude as the standard deviations shown in Col. 7 (which have an RMS value of 210 thousand). Hence, the part of the variability of the estimate of undercount that is attributable to differences between the seven models in terms of model structures, data sources and assumptions is very substantially less than the part of the variability attributable to the RVPs. We may conclude that the models have the qualities of consistency and robustness in good measure.

As Table 6 shows, the different models use emigrant data from foreign censuses to widely varying degrees. Thus, Model A uses data from 22 contemporaneous foreign censuses; Model P uses data from only 3; whilst model R0 uses no data from a contemporaneous foreign census but uses data from the previous censuses of 22 countries (as indicators of current emigrant numbers). Moreover, the models use our own census figures on returners to widely varying degrees (Table 6, last col.). Given the near congruence of the results of the different models (as described in the previous paragraph), it becomes very difficult, in my view, to argue that the validity of the methodology is in serious doubt because of the following factors:

(1) the different methods, definitions and timing of foreign censuses.
(2) weaknesses in the quality of responses to our own census question on usual address one year before the census date (that is, weaknesses in the data on returners).
(3) (in the case of the R models) changes in patterns of emigration in the period since the previous census round.

That is not to say that the methodology of the models developed in this paper would be immune to shock due to war, economic depression or drastic changes in regulations affecting emigration.

17. Next Steps

The conclusions of this article are based on the study of one country, England and Wales, and one census, 1991. They need to be tested in other countries and at other times.

An obvious first test would be in the 2001 Census of E&W because, at the time of writing, the validity of the census results is being questioned. The public response to the census was poorer than in any previous UK census, with only 94 per cent of the population appearing on completed census forms. Using the results from a large-scale PES, the Office for National Statistics (ONS) have made adjustments to all published census tables with the aim that they should cover 100 per cent of the population: a One Number Census (ONC) (Brown et al, 1999). However, at national level the ONC figures of population show a substantial deficiency of men in the age range 20–44 as compared with the population estimates which had been ''rolled forward'' from 1981 and which had been intended as a demographic check on the ONC results. Again at national level, the ONC sex ratio M:F at ages 20–44 is at an unprecedented low level. And at local level too, there is evidence of a downward bias in the sex ratio. But the ONS are deeply committed to the extensive range of figures that have already been published and have so far resisted a check on the 2001 estimates of undercount by means of what I call the ISR/JOS method. Their stated grounds for adopting this stance are set out, and then rebutted, in Redfern (2004), Discussion.

If this article's methodology were to be applied in another country, then, as already noted, the details should be reviewed to ensure that they fitted that country's demography and census practices.

18. Summing Up: Strengths and Weaknesses

The methods described here are empirical. They are most unlikely to be optimal, and so may be capable of further development. The methods will probably be effective only in certain demographic situations, for example when numbers of emigrants and immigrants are not too large compared with numbers of natives-at-home. (If we take natives-at-home in E&W in 1991 as 100, emigrants and immigrants were respectively 8 and 11.)

The methods do not help to solve the universal problem of counting illegal immigrants. Nor could the methods yield estimates of undercount at sub-national – that is, regional – level unless the foreign censuses of most relevance were able to analyse emigrants by region of birth; that kind of analysis is sometimes available but not often.

The methods developed in this article go a substantial way – but not the whole way – towards overcoming the most serious demerit of the ISR method: namely, that its implementation would have to await the availability of census data from the principal foreign countries which received emigrants (– the ISR Model A used data from 22 countries). By contrast, this article's Model R0 does not use data from any contemporaneous foreign census, so that it could produce ''broad brush'' estimates of undercount to a timescale similar to that of traditional methods of measuring undercount.

But, as noted in Section 14, the R0 results are degenerate in some respects. More reliable estimates would emerge from models that used emigrant data from a small number of contemporaneous foreign censuses: for example, Models P and R3 used data from 3 countries. The task of assembling the foreign data needed for these models would be much less than that needed for the ISR model and the speed of doing so greater. Even so, results might not emerge until after the census agency had already published its first estimates of undercount using other methods. That might pose problems: ''Shall we revise our estimates? How do we present the conflict of evidence?'' That challenge must be confronted squarely.

The methods of this article have important strengths. First, their cost is low. They do not require a PES with all its fieldwork and processing costs. Nor do they depend on machinery for measuring migrant flows, which is costly and, especially in countries without a population register, notoriously unreliable.

Second, the results are in the form of frequency distributions. They present explicit statements of error margins. In this respect they are unlike most other analyses of undercount which merely present central estimates or ''best guesses'' and which mask their substantial unreliability. An all-too-rare example of estimates of undercount with explicit statements of error appears in Robinson et al (1993) which refers to the 1990 Census of the U.S.; the ideas underlying that paper have many similarities to those deployed here. It is worth recalling the comment on my ISR paper by William Bell (1999), made at the ISI meeting in Helsinki: ''The uncertainty assessments are *subjective*'' (his italics).

Third and most importantly, the results from a variety of models demonstrate consistency and robustness.

19. References

Bell, William R. (1999). Discussion of Papers on Population Estimation and Forecasting. Bulletin of the International Statistical Institute, 52nd Session. Proceedings, LVIII, 3, 110.

Brown, J.J., Diamond, I.D., Chambers, R.L., Buckner, L.J., and Teague, A.D. (1999). A Methodological Strategy for a One-number Census in the UK. Journal of the Royal Statistical Society, Series A, 162, 247–267.

Office of Population Censuses and Surveys and General Register Office for Scotland (1995). 1991 Census, General report, Great Britain, Chapter 10. London: Her Majesty's Stationery Office.

Redfern, P. (2001). A Bayesian Model for Estimating Census Undercount, Taking Emigration Data from Foreign Censuses. International Statistical Review, 69, 277–301.

Redfern, P. (2004). An Alternative View of the 2001 Census and Future Census-taking. Journal of the Royal Statistical Society, Series A, 167, 209–248.

Robinson, J.G., Ahmed, B., Das Gupta, P., and Woodrow, K.A. (1993). Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis. Journal of the American Statistical Association, 88, 1061–1071.

U.S. Census Bureau (1992). Report of the Committee on Adjustment of Postcensal Estimates. August 7. Attachment 3. Washington: U.S. Census Bureau.

Received August 2002
Revised April 2003