National Records of Scotland

Beyond 2011

Population Estimates from Administrative Data Sources Using Bayesian Markov chain Monte Carlo (MCMC) Methods: A Follow Up Investigation

Published on 17 April 2014

Contents

1. Introduction

1.1 One alternative to the population wide census currently under investigation by National Records of Scotland (NRS) is to use pre-existing administrative data to construct small area population estimates (for both Data Zones (DZ) and Intermediate zones (IZ)). Previous work carried out in collaboration with Professor Peter Congdon of Queen Mary, University of London showed that it was possible to construct such estimates. In this report, it is assumed that the reader is familiar with the report of this work 'Population estimates for Data and Intermediate Zones Using Administrative Data' as there are several references to it and, in the interests of avoiding repetition, details are not given here.

1.2 Two major limitations were present in the production and evaluation of the above population estimates. The first was that in order to protect the anonymity of individual data all records used were subject to disclosure control procedures. This resulted in the exclusion of particularly small population groups from the production of estimates. Whilst many such values were themselves estimated from other sources the procedure still resulted in a decreased volume of primary data. The second was the necessary use of the 2011 Mid-Year Estimates (MYEs) as 'gold standard' estimates. Without up-to-date census data available the administrative data estimates could only be compared against local authority MYEs which may only provide approximate indications of the accuracy of administrative estimates.

1.3 This report addresses these two limitations. As these estimates were produced internally to the NRS the disclosure-control requirements for release of such data were unnecessary, allowing full administrative data to be used in the construction of population estimates. Furthermore, the release of the census data for 2011 in Scotland allowed for a better assessment of the accuracy of administrative population estimates to be carried out.

2. Conclusions and Recommendations

2.1 The investigation found that the use of data which had undergone disclosure-control procedures did have an overall negative effect on the accuracy of population estimates. However any improvements due to use of full data were minimal (aggregate benefits usually representing improvements of at most one percentage point in the root mean square percentage discrepancy).

2.2 Whilst the method was very successful in integrating information from a number of sources it also displayed susceptibility errors in the provided datasets. This is illustrated by the large decrease in accuracy of estimates for those aged over 85 as a result of including non-disclosure controlled data.

2.3 Given the minor effects of non-disclosure controlled data and census based evaluation the recommendations of this paper do not differ from those of the original article.

3. The Follow-Up Investigation

3.1 The first step in assessing possible improvements to the administrative estimates produced was to attempt to replicate the estimates previously produced using disclosure controlled data compared to 2011s MYEs. The estimates were found to be reliably reproducible with both estimates showing the same percentage discrepancy from the MYEs when aggregated to the local authority level (Table 1). When aggregated over age groups predictions were again successfully replicated (Table 2).

Inclusion of non-disclosure-controlled data

3.2 The estimates produced were based on five sources of aggregate administrative data. These were the National Health Service Central Register (NHSCR), Customer information from the Department of Work and Pensions, Her Majesty Revenues and Custom Child Benefit data, the annual 'Pupils in Scotland' census and the Electoral Register. The effects of including non-disclosure data was evaluated at both the local authority level and at the level of nationwide aggregate age group populations.

3.3 Use of non-disclosure controlled data in creating local authority population estimates produced an overall decrease in the percentage discrepancy between the produced estimates and the MYE. Using non-controlled data caused the Root Mean Square percentage Discrepancy (RMS%D) to fall from 3.8% to 2.9% (Table 1). The overall improvement was largely driven by improved estimates in large urban areas (most notably in Glasgow and Edinburgh) which were consistently underestimated by predictions using disclosure-controlled data.

3.4 Whilst unconstrained local authority estimates were improved by the inclusion of non-disclosure controlled data the overall effect of including such data in making nationwide age group estimates was negative, with the RMS%D rising from 4.9% to 6.7% (Table 2). However this increased error is primarily a result of the extremely poor estimate produced for the group aged 85 and older. Ignoring this group, the estimates produced using full data improve very slightly (the RMS%D falling from 4.9% to 4.7%). One possible cause of the poor performance on the over 85 age group is the inclusion of complete NHSCR data, which differs significantly in its estimates for this group from the MYE (refer to the original report for breakdown of population figures across the sources employed).

3.5 Excluding the extreme case of the over 85 age group estimates produced using full data as opposed to data which had undergone disclosure control produced slightly improved estimates at both the local authority and aggregate age group levels. However, these estimates were not constrained within the MYE as proposed in the previous report. Evaluating the estimates as above is also still limited by the use of the MYEs as a benchmark as opposed to census figures.

Comparisons of unconstrained estimates with 2011 Census figures

3.6 Comparisons of the unconstrained administrative estimates with aggregate local authority data reported by the 2011 Census showed slightly decreased errors for estimates using full-administrative data, but errors for estimates produced using disclosure-controlled data showed little to no improvement from the MYE comparison. For local authorities, estimates using disclosure-controlled data went

from a RMS%D of 3.8% compared to the 2011 MYE to errors of 3.7% in comparison with census data from that year. In contrast, the errors of the full data estimate decreased from 2.9% in comparison with the MYE to 2.6% in comparison to census figures.

3.7 For age groups, the effects of census based comparisons, rather than MYE, appear to be minimal with regards to the errors in estimates. The disclosure-controlled estimates show little change in discrepancy level (RMS%D of 4.9 compared to MYE, and 4.6% compared to census data). Estimates produced using non-disclosure controlled data also show little change regardless of the inclusion of the 85+ age group whose estimate is uniquely poor (RMS%D falls from 6.7% to 6.1% including 85+ group, 4.7% to 4.2% excluding 85+ group).

Comparisons with scaled data in the Glasgow area

3.8 To examine the relationship between estimates using full-data and disclosure-controlled sources that had undergone the recommended population scaling, scaled estimates were produced for the City of Glasgow whose results were then compared to the 2011 Census data for each IZ in the area.

3.9 The results of this comparison can be seen in Figure 1. The equivalent map from the original report is included for comparison (Figure 2).

3.10 The general pattern of results remains the same as in the previously reported work. Whilst the use of non-disclosure controlled data and comparison to census figures, rather than MYEs, results in moderate improvements in accuracy, the final estimates still suffer from consistent discrepancies with population figures.

3.11 The Glasgow area estimates also saw one of the largest improvements from inclusion of non-disclosure-controlled data and the move to the census as the benchmark (-9% to -1% discrepancy). Maps of other areas would show less of an improvement.

Conclusion

3.12 The conclusion from the follow-up work reported above reinforces that reached in the original work undertaken by Professor Congdon. This is that, while Bayesian Markov chain Monte Carlo (MCMC) methods represent a powerful method of integrating data and allowing information contained in different formats in different data sets to be brought together in a single output, they are not design to identify and allow for bias in these data sets and therefore any bias is taken through to the output. It is concluded therefore that some method of addressing bias will require to be applied prior to the use of Bayesian methods for data integration.

Table 1 LOCAL AUTHORITY	2011 MYE	Original Population Estimate	Replicated Population Estimate[1]	% Discrepancy of Both Estimates with MYE	2011 Census Figure	Full Data Estimate	% Discrepancy of full data Estimate with MYE	% Discrepancy of full data Estimate with Census
Aberdeen City	220,420	227,056	227,137	3.0%	222,793	226,119	2.6%	1.5%
Aberdeenshire	247,600	249,115	249,170	0.6%	252,973	251,121	1.4%	-0.7%
Angus	110,630	115,386	115,409	4.3%	115,978	115,770	4.6%	-0.2%
Argyll & Bute	89,590	85,827	85,885	-4.1%	88,166	87,294	-2.6%	-1.0%
Clackmannanshire	50,770	52,902	52,803	4.0%	51,442	52,528	3.5%	2.1%
Dumfries & Galloway	148,060	149,548	149,510	1.0%	151,324	150,888	1.9%	-0.3%
Dundee City	145,570	141,856	141,949	-2.5%	147,268	146,171	0.4%	-0.7%
East Ayrshire	120,200	123,868	123,828	3.0%	122,767	123,406	2.7%	0.5%
East Dunbartonshire	104,570	109,503	109,513	4.7%	105,026	109,288	4.5%	4.1%
East Lothian	98,170	99,135	99,114	1.0%	99,717	99,483	1.3%	-0.2%
East Renfrewshire	89,850	92,113	92,234	2.7%	90,574	92,345	2.8%	2.0%
Edinburgh, City of	495,360	447,820	448,142	-9.5%	476,626	467,058	-5.7%	-2.0%
Eileen Siar	26,080	25,689	25,642	-1.7%	27,684	26,117	0.1%	-5.7%
Falkirk	154,380	157,600	157,517	2.0%	155,990	156,929	1.7%	0.6%
Fife	367,370	371,304	371,455	1.1%	365,198	370,938	1.0%	1.6%
Glasgow City	598,830	543,604	544,444	-9.1%	593,245	588,817	-1.7%	-0.7%
Highland	222,370	226,103	226,115	1.7%	232,132	227,504	2.3%	-2.0%
Inverclyde	79,220	79,872	79,905	0.9%	81,485	81,482	2.9%	0.0%
Midlothian	82,370	85,257	85,114	3.3%	83,187	84,842	3.0%	2.0%
Moray	87,260	89,018	88,971	2.0%	93,295	88,656	1.6%	-5.0%
North Ayrshire	135,130	142,101	142,020	5.1%	138,146	141,675	4.8%	2.6%

Footnote

1) Whilst small discrepancies exist between the original and replicated estimates, this is a typical result of the Monte Carlo methods employed. These discrepancies can be made arbitrarily small by increasing the computing time available to produce the estimates.

Table 1 (continued)

LOCAL AUTHORITY	2011 MYE	Original Population Estimate	Replicated Population Estimate[1]	% Discrepancy of Both Estimates with MYE	2011 Census Figure	Full Data Estimate	% Discrepancy of full data Estimate with MYE	% Discrepancy of full data Estimate with Census
North Lanarkshire	326,680	347,251	347,170	6.3%	337,727	344,701	5.5%	2.1%
Orkney Islands	20,160	19,390	19,386	-3.8%	21,349	19,842	-1.6%	-7.1%
Perth & Kinross	149,520	145,997	146,036	-2.3%	146,652	146,583	-2.0%	0.0%
Renfrewshire	170,650	172,661	172,562	1.1%	174,908	175,461	2.8%	0.3%
Scottish Borders	113,150	110,770	110,807	-2.1%	113,870	111,863	-1.1%	-1.8%
Shetland Islands	22,500	21,496	21,515	-4.4%	23,167	21,841	-2.9%	-5.7%
South Ayrshire	111,560	116,278	116,127	4.1%	112,799	115,333	3.4%	2.2%
South Lanarkshire	312,660	324,462	324,484	3.8%	313,830	324,347	3.7%	3.4%
Stirling	90,770	88,811	89,071	-1.9%	90,247	89,990	-0.9%	-0.3%
West Dunbartonshire	90,360	92,335	92,307	2.2%	90,720	93,238	3.2%	2.8%
West Lothian	172,990	178,390	178,494	3.2%	175,118	177,349	2.5%	1.3%

Footnote

1) Whilst small discrepancies exist between the original and replicated estimates, this is a typical result of the Monte Carlo methods employed. These discrepancies can be made arbitrarily small by increasing the computing time available to produce the estimates.

Table 2 Age group	2011 MYE	Original Population Estimates	Replicated Population Estimates	% Discrepancy of replicated Estimates with MYE	2011 Census Figure	Full Data Estimate	% Discrepancy of full data Estimate with MYE	% Discrepancy of full data Estimate with Census
0-4	297,741	270,654	270,792	-9.1%	293,000	278,724	-6.4%	-4.9%
5-9	273,374	269,002	269,345	-1.5%	270,000	284,132	3.9%	5.2%
10-14	282,776	277,398	277,830	-1.7%	292,000	283,758	0.3%	-2.8%
15-19	317,880	308,527	308,468	-3.0%	331,000	314,955	-0.9%	-4.8%
20-24	367,138	351,263	351,038	-4.4%	364,000	359,218	-2.2%	-1.3%
25-29	358,433	361,432	361,861	1.0%	346,000	369,048	3.0%	6.7%
30-34	322,100	338,374	338,492	5.1%	322,000	347,572	7.9%	7.9%
35-39	321,711	350,027	350,320	8.9%	340,000	356,382	10.8%	4.8%
40-44	384,643	403,707	403,801	5.0%	394,000	410,188	6.6%	4.1%
45-49	402,552	414,972	415,027	3.1%	411,000	420,728	4.5%	2.4%
50-54	374,031	376,401	376,280	0.6%	376,000	380,211	1.7%	1.1%
55-59	328,047	326,869	326,876	-0.4%	331,000	329,742	0.5%	-0.4%
60-64	331,987	329,356	329,307	-0.8%	337,000	330,720	-0.4%	-1.9%
65-69	261,533	252,157	252,193	-3.6%	262,000	255,318	-2.4%	-2.6%
70-74	217,780	206,983	206,882	-5.0%	221,000	211,208	-3.0%	-4.4%
75-79	177,999	165,318	165,314	-7.1%	178,000	170,325	-4.3%	-4.3%
80-84	124,845	115,145	115,117	-7.8%	123,000	119,488	-4.3%	-2.9%
85+	110,230	114,936	114,894	4.2%	108,000	87,260	-20.8%	-19.2%
All Ages	5,254,800	5,232,521	5,233,837	-0.4%	5,299,000	5,308,978	1.0%	0.2%

Figure 1: Intermediate Zone differences, full data estimates versus Census 2011, Glasgow

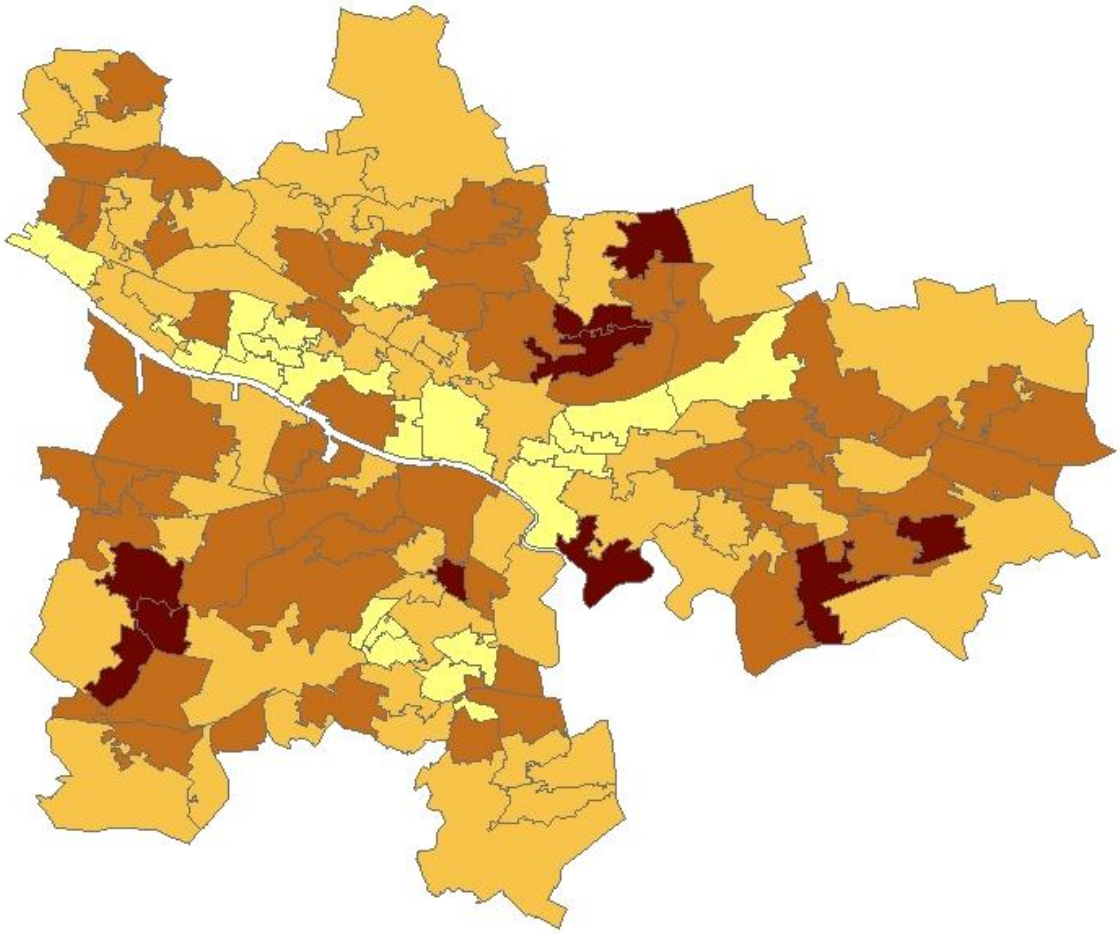

Figure 2: Intermediate Zone differences, disclosure-controlled data estimates versus Mid-Year Estimates, 2011, Glasgow