A Bayesian Approach to Population Estimation with Administrative Data

John R. Bryant[1] *and Patrick Graham*[2]

The article describes a Bayesian approach to deriving population estimates from multiple administrative data sources. Coverage rates play an important role in the approach: identifying anomalies in coverage rates is a key step in the model-building process, and data sources receive more weight within the model if their coverage rates are more consistent. Random variation in population processes and measurement processes is dealt with naturally within the model, and all outputs come with measures of uncertainty. The model is applied to the problem of estimating regional populations in New Zealand. The New Zealand example illustrates the continuing importance of coverage surveys.

Key words: Bayesian; official statistics; demography; administrative data.

1. Introduction

Statistical agencies around the world are developing new methods for population estimation that make better use of administrative data. The long-term goal is often to do away with a traditional census and to rely on administrative data, perhaps supplemented by a coverage survey. This goal has already been attained in some countries (Coleman 2013).

The conceptually simplest approach to estimating population size and structure from administrative data is to maintain a highly accurate population register, and to read population estimates straight off the register. However, few countries have this option available to them.

The conceptually simplest alternative to a population register is to take a single administrative data source, such as a list of people enrolled within the health system, and to adjust for known deficiencies. In the absence of a census, a standard way to identify deficiencies is to conduct a survey collecting information on undercoverage, overcoverage, and misclassification errors such as faulty addresses. Using a single administrative data source plus a coverage survey is much like using a traditional census plus a coverage survey. Relying on a single administrative data source has important disadvantages, however. The statistical agency is unlikely to have the same degree of control over administrative data that it does over the census, and may therefore be unable to prevent changes in policy, information technology, or recording practices that affect the

[1] Statistics New Zealand – Population Statistics, Dollan House Private Bag 4741, Christchurch 8140, New Zealand. Email: john.bryant@stats.govt.nz
[2] Statistics New Zealand, Statistical Methods, Dollan House Private Bag 4741, Christchurch 8140, New Zealand. Email: patrick.graham@stats.govt.nz

quality and consistency of the data. Moreover, a single data source may not take in all groups within the target population.

Rather than rely on a single administrative data source, a statistical agency can combine several administrative data sources. The combining of data can occur at the individual level, via record linkage. Linking together multiple administrative datasets is more difficult than is generally realised, however, particularly in countries such as New Zealand where there is no universal personal identifier. Linkage errors complicate population estimation: when an individual appears in two datasets but the individual's records are not linked, he or she may be counted twice in population estimates. Large-scale record linkage also raises privacy and ethical concerns.

The combining of data sources can instead occur at the level of the cell count. Counts classified by age, sex, and region can be calculated for each dataset, and then population can be derived as some sort of weighted combination. This avoids many of the problems of individual linking, but poses problems of its own. Assigning weights to datasets is difficult, especially when there is no gold standard and there is random variation in the data and population. Moreover, different data sources typically include different variables, and cover different age groups or time periods (Bycroft 2013; Office for National Statistics 2013).

Statistics New Zealand has been developing a formal statistical approach to deriving population estimates from multiple administrative data sources (Bryant and Graham 2013). Data are combined at the level of the cell count. The overall model contains submodels describing regularities within the demographic processes, and describing the relationship between the demographic processes and the various available datasets. The approach is Bayesian, which provides the necessary flexibility and the ability to account for diverse sources of uncertainty. Coverage rates play a central role in the modelling, as a diagnostic, and as a source of implicit weights for the data.

This article provides an overview of our approach, and describes an application to the problem of estimating regional populations in New Zealand. The application illustrates the difficulty of inferring population from administrative data alone. The results suggest that, in the absence of a traditional census, it would be necessary to supplement administrative data with a carefully-targeted coverage survey.

2.   A Bayesian Framework for Population Estimation

Here we provide a brief introduction to our statistical model. More detail is available in Bryant and Graham (2013). The model is summarized in Figure 1.

At the core of the model is a demographic account Q (Rees 1979; Stone 1984). The account is a complete description of the demographic stocks and flows of interest. In Bryant and Graham (2013), the demographic account contains counts of births, deaths, migrations, and population, all disaggregated by age, sex, region, and time, and all linked by accounting identities. In the application below, however, we work with a simple account containing only population stocks. Whatever the level of detail, the account is treated as unobserved, and values for cells within the account must be inferred from available data.

Entries within an account typically exhibit strong regularities. For instance, age profiles for areas with universities typically have sharp peaks in the main student ages. The model of the demographic account, M_Q, captures these regularities. Often there are auxiliary data

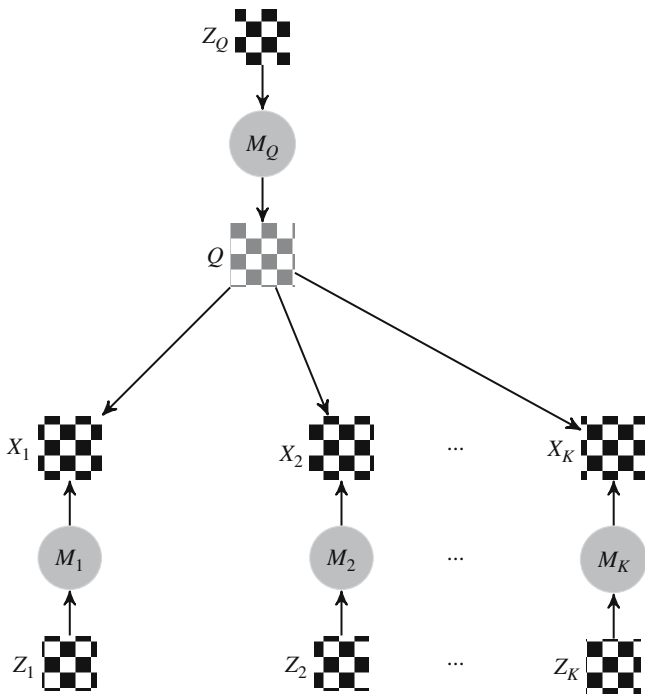Fig. 1.   *Our population estimation framework. Q is the demographic account, the Xs are data sources, and the Zs are covariates. Black denotes observed quantities and grey denotes unobserved ones. Hatched squares denote counts of people or events, and circles denote submodels. Arrows denote probabilistic relationships.*

Z_Q that can assist with the estimation of parameters within M_Q. Data on the location of universities, for instance, can predict the existence of age spikes.

Datasets $X_1, \ldots, X_K$ consist of counts of people or events, or proxies for these counts. No sharp distinction is made between administrative sources such as tax data and more traditional sources such as the census. Datasets can be added to or removed from the model easily.

The $M_1, \ldots, M_K$ denote data models. A data model M_k treats dataset X_k as a response and the demographic account Q as a predictor. The model describes the closeness and consistency of the relationship between the data and the underlying demographic process. The approach is similar to that of measurement error or latent variable models, in that the datasets are treated as reflecting a common unobserved construct.

The relationship between data and demographic process varies from data source to data source. With a highly reliable data source, there is essentially a one-to-one relationship. A reliable birth registration system, for instance, captures almost every birth. Some data sources are subject to undercoverage or overcoverage, but in a consistent way. For instance, a data source might cover only 80% of the target population, but maintain the same coverage level from year to year. Finally, some data sources are subject to fluctuating degrees of coverage, with no consistent relationship between coverage levels and variables such as age, sex, region, or time.

If a data source is known to be highly reliable, the data model can be designed accordingly: Section 3 gives an example. More typically, the analyst has some idea of

patterns in coverage, but does not know detailed coverage rates. Data models can incorporate the analyst's qualitative knowledge by, for instance, using age and sex as predictors if the analyst thinks that coverage varies along these dimensions. The data models then provide quantitative measures of the relationship between coverage rates and the predictors. Even more can be learned when the data model is hierarchical – that is, when the coverage rates are themselves treated as draws from distributions, the parameters of which vary with the predictors (Gelman and Hill 2007). Hierarchical models can distinguish between situations where variables such as age, sex, region, and time predict coverage rates precisely and situations where predictions are poor. In other words, hierarchical models provide quantitative measures of the consistency of a data source.

When the population estimation model generates proposed values for cells in the demographic account, proposals that fit the predictions of the relevant data models are more likely to be accepted. Models for consistent data sources make sharper predictions than models for inconsistent data sources. Departures from sharp predictions are penalized more heavily than departures from diffuse ones. The population estimation model thus implicitly gives greater weight to consistent data sources than to inconsistent ones.

The fact that each dataset X_k is 'predicted' from the corresponding data model M_k and the demographic account has important practical advantages. The demographic account, by construction, has at least as much detail as any of the individual datasets. If a dataset is missing a dimension that is present in the demographic account, then the account is aggregated across that dimension before it is supplied to the data model. Similarly, if a dataset has missing values for a given year or age group, then the corresponding years or age groups are removed from the account before it is supplied to the data model. This means that it is not necessary to place all the input data into the same format. The approach thus avoids one of the most difficult and time-consuming parts of traditional population estimation.

Inference is carried out via Markov chain Monte Carlo methods. A Gibbs sampler alternates between the full conditional distributions for Q, M_Q, and $M_1, \ldots, M_k$. Sampling from the distribution for Q is difficult. The accounting identities and non-negativity constraints in Q mean that cell values do not follow standard distributions, so that customized updating procedures are required. However, sampling from M_Q and $M_1, \ldots, M_K$ is generally straightforward (Bryant and Graham 2013). The model output consists of samples from the posterior distributions for the demographic account, the demographic model M_Q, and the data models $M_1, \ldots, M_K$. Subsections 3.2 and 3.3 provide some examples.

3.   Application to Subnational Population Estimation in New Zealand

3.1.   Data and Setting

We apply a simple version of the model to the problem of estimating population counts by five-year age group, sex, time, and 'territorial authority' in New Zealand. Territorial authorities range in size from a few hundred people to 1.5 million. We omit the smallest territorial authority, and estimate counts for the remaining 66. Two of our four data sources only have consistent data for the years 2012 and 2013, so we restrict the estimation to those years. Although a population census was carried out in 2013, we do not use data from the 2013 census except for a validation exercise.

The data sources are summarized in Table 1. The first three are all administrative sources. As discussed in detail in Statistics New Zealand (2013) and Gibb (2014), information about administrative processes and comparison of counts at the national level suggest that none of the three administrative data sources accurately reflect the number of people who live in New Zealand. The target population of the primary health care data, for instance, is more or less equal to the usually resident population, but people within the target population do not appear in the data if they do not visit the doctor. Comparisons of numbers at the national level indicate that many young adults, who tend not to visit the doctor, are indeed not included. The target population for the tax data is people who have tax deducted directly from wage or social welfare payments. The target population excludes most people who do not work or receive social welfare payments, and includes some people living outside New Zealand. The target population for the electoral roll data also does not quite align with the resident population. Moreover, national figures indicate that many young people who are part of the target population are not on the electoral roll.

The one nonadministrative data source, the national-level population estimates, is the most accurate of the sources in Table 1. It is constructed by adjusting census data (in this case 2006 census data) for coverage errors, and then updated using accurate data on births, deaths, and international migration.

3.2. Initial Model

3.2.1. Specification
We model population using

$$q_i \sim \text{Poisson}\left(\theta_i^Q\right)$$

$$\log \theta_i^Q \sim \text{N}\left((\text{H}^Q\beta^Q)_i, \sigma_Q^2\right)$$

Table 1. *Data sources used in the application*

Data source	Description	Expected relationship with population counts	Detail available
Health	Enrolment in primary health care providers	Good correspondence overall, but lower for young adults, particularly males	Age, sex, region, 2012-2013
Tax	People with taxable income from work or benefits	Some overcoverage and undercoverage, varying by age and sex	Age, sex, region, 2012-2013
Electoral	People enrolled to vote	Significant undercoverage at younger ages.	Ages 18 + and region, 2013. No sex.
National population estimates	National population by age and sex	Accurate, though with some uncertainty about young adults	Age and sex, 2012-2013

where q_i is the number of people in the ith age-sex-region-time cell of the demographic account, β^Q is a vector of coefficients, and H^Q is a design matrix. The model includes age, sex, and region effects, plus all second-order interactions between these terms, plus a time effect. Priors for the model are described in the Appendix.

We model the relationship between the tax data X^{tax} and demographic account Q using

$$x_i^{\text{tax}} \sim \text{Poisson}\left(\theta_i^{\text{tax}} q_i\right)$$

$$\log \theta_i^{\text{tax}} \sim N\left((H^{\text{tax}} \beta^{\text{tax}})_i, \sigma_{\text{tax}}^2\right)$$

Parameter θ_i^{tax} measures coverage in cell i. The model includes an age effect, a sex effect, and an interaction between the two. By not including region and time effects, we are implying that we expect age-sex profiles for coverage to be similar across regions and across time. Restrictions such as this are necessary to achieve identification. The restrictions are not completely binding, however. As is apparent in the results below, a sufficiently strong signal in the data pulls the posterior distribution away from the prior.

The σ^2 term in a Bayesian hierarchical model like the one for tax measures how well the variables in β are able to explain variation in θ. A posterior distribution for σ^2 that is concentrated near zero implies that the variables have substantial predictive power (Gelman and Hill 2007). In the model for the tax data, low values for σ_{tax}^2 would imply that age, sex, and the interaction between the two accurately predict coverage rates for the tax data. In other words, low values for σ_{tax}^2 would imply that the age-sex profile for coverage was approximately constant across regions and time. Conversely, high values for σ_{tax}^2 would imply inconsistent age-sex profiles.

The health and electoral data are modelled in the same way as the tax data, except that there is no sex effect in the model for the electoral data. The national population estimates need a different model. A Poisson distribution has too much variance to represent the close relationship that exists between national population estimates and the true population counts. Instead we use a Poisson-binomial mixture,

$$x_i^{\text{nat}} = u_i + v_i$$

$$u_i \sim \text{Poisson}((1 - \pi)q_i)$$

$$v_i \sim \text{Binomial}(q_i, \pi).$$

The Poisson-binomial mixture can be interpreted as a simple model of enumeration errors, in which v_i is the number of people correctly enumerated in cell i, and u_i is the number incorrectly enumerated. Parameter π is set to 0.98, based on discussions with Statistics New Zealand staff about the likely accuracy of the national estimates.

The results presented below were obtained from five independent chains with a burn-in of 10,000 and production of 10,000, recording one out of every 50 iterations. We monitored convergence by calculating potential scale reduction factors. The multivariate potential scale reduction factor (Brooks and Gelman 1998) for ten randomly chosen population cells was 1.02.

3.2.2. Results

Figure 2 shows the results for four selected regions. The first two regions are highly urban; the second two are a mix of rural areas and towns. The 95% credible intervals for the second two regions are wider, reflecting their smaller size and hence the greater relative importance of random variation. The first two regions have peaks beginning in the late teenage years, while the second two regions have troughs. These are the characteristic age profiles produced by the migration of young people out of rural areas and towns into cities.

Figure 3 shows estimates of coverage rates for the three administrative datasets. A rate of 1.0 implies that there is one person in the administrative dataset for each person in the true population; a rate higher than 1.0 implies overcoverage, and a rate lower than 1.0 implies undercoverage.

Each dataset has a characteristic age profile for coverage. The width of the credible intervals also varies across datasets. This reflects the consistency of the coverage profiles across regions and across time, or, equivalently, the value for σ. In the model for the health data, the median posterior estimate for σ is 0.013; in the model for the tax data it is 0.086; and in the model for the electoral data it is 0.055. When distributing population across regions, the model penalizes deviations from the pattern predicted by health data the most, and penalizes deviations from the pattern predicted by the tax data the least.

The results for Dunedin in Figure 3 are anomalous. The age group 20–24 appears to have coverage rates well over 1.0 in the tax and electoral data for Dunedin, but coverage rates of less than 1.0 in the tax and electoral data for other regions. The explanation for this anomaly is that the health data for Dunedin are idiosyncratic, resulting in population estimates that are too low. Dunedin has a large university and a large student population. However, the student health service in Dunedin does not belong to the standard primary health care system, so most young people there do not show up in the health data. The model has not been provided with information about the discontinuity in the relationship between health data and population. It therefore places its usual high weight on the health data and low weight on the tax and electoral data.

The tax data for the older ages shows a different sort of anomaly. Estimated coverage rises about 1.0, particularly in Auckland. The rise in apparent coverage can be explained by idiosyncrasies of the administrative data. It is clear from the metadata, and from the fact
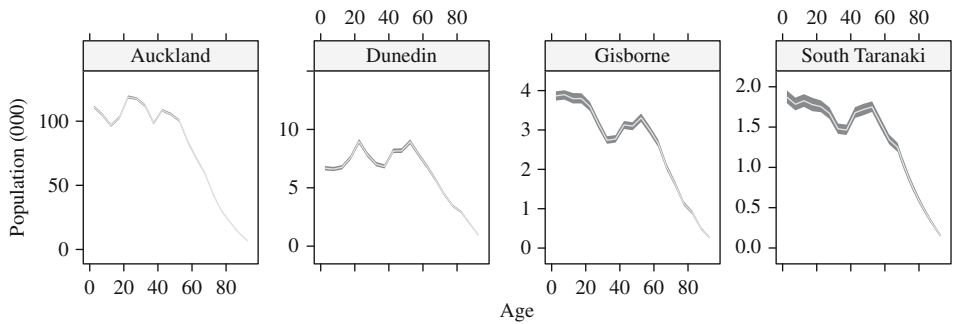

Fig. 2. *Population estimates from the initial model, for four selected regions, males and females combined, 2013. The dark bands are 95% credible intervals and the grey lines are medians.*
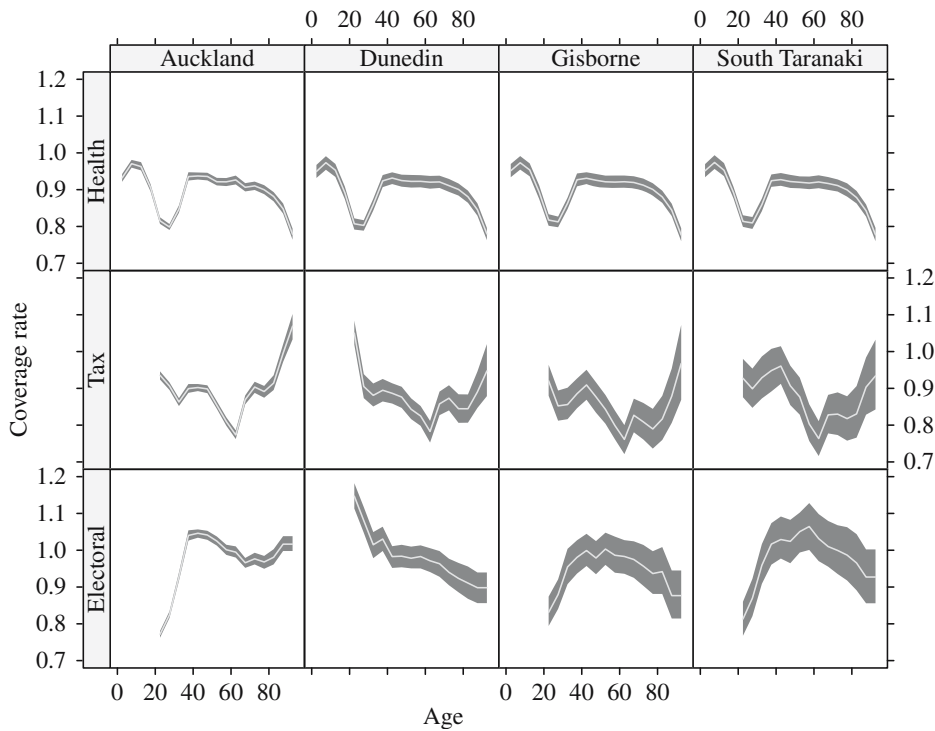
Fig. 3.   Coverage rates from the initial model, for the three administrative data sources (the rows) and four selected regions (the columns), males and females combined, 2013.

that the dataset contains many people aged 100 or more, that many people are not removed from the dataset after they have died.

3.3.   Revised Model

3.3.1.   Specification

We make two specification changes in response to the initial results. First, we add a covariate to the population model that takes a value of 1 if a cell refers to age groups 15–19 or 20–24 and to a main centre, and 0 otherwise. This covariate captures the systematic relationship between the type of region and the number of young people. Second, we delete the cells from the health dataset that refer to 15–24 year olds in Dunedin. In the absence of data on students enrolled in the student health service, the health dataset provides little guidance on numbers of young people in Dunedin.

3.3.2.   Results

Population estimates from the revised model are shown in Figure 4. The 'student spike' in Dunedin is substantially higher under the revised model than it was under the initial model. The credible intervals are much wider for the student ages in Dunedin than they are for other ages, which is appropriate, given that the estimates for the student ages are constructed using the two least-reliable datasets.
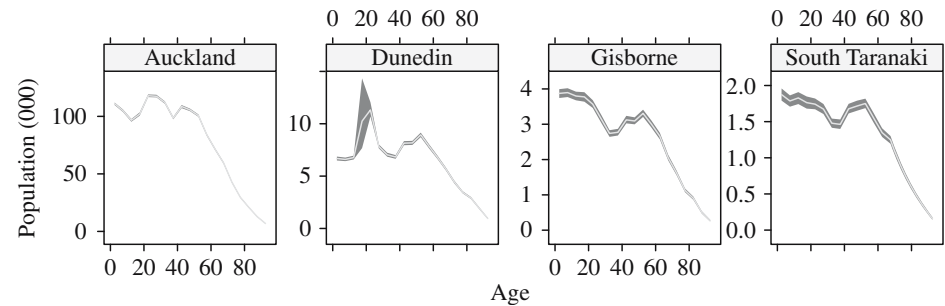
Fig. 4. *Population estimates from the revised model.*

Coverage rates from the revised model are shown in Figure 5. The coverage rates for the tax and electoral datasets in Dunedin look less anomalous than before, though they still differ from the other regions. We suspect that, even with the main-centre-by-age indicator variable, the population model is still pulling the Dunedin estimate down, closer to the age pattern for other regions.

Finally, Figure 6 shows results from a simple validation exercise. We take the population counts from the 2013 census and, within each age-sex combination, scale regional population numbers upwards so that they match the national population estimates described in Table 1. We subtract the model estimates from the scaled census estimates as a measure of errors in the output from the revised model. There is a clear pattern in the four
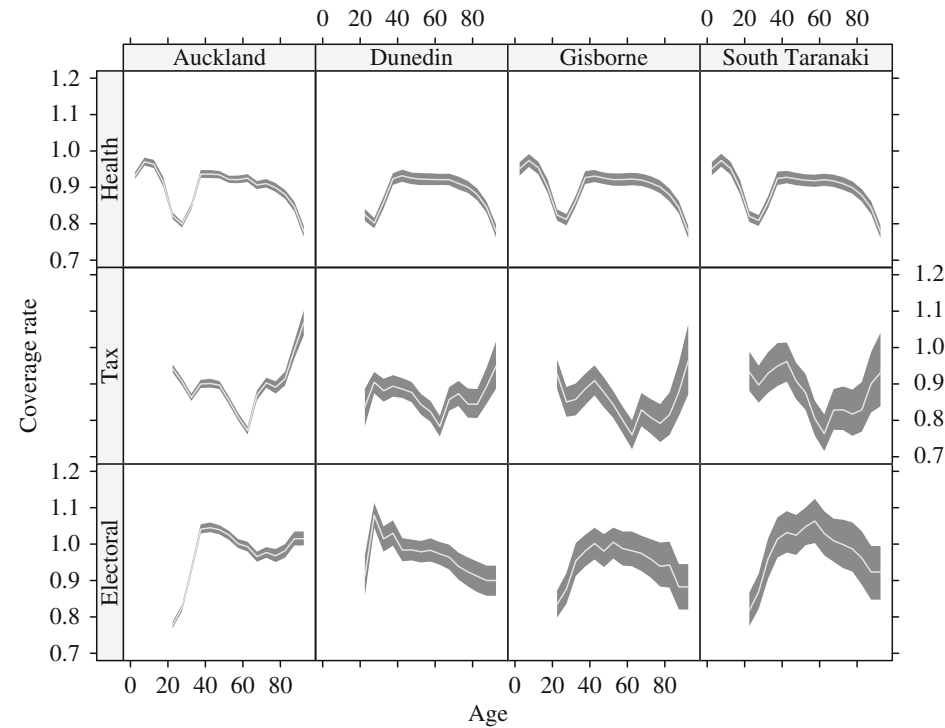

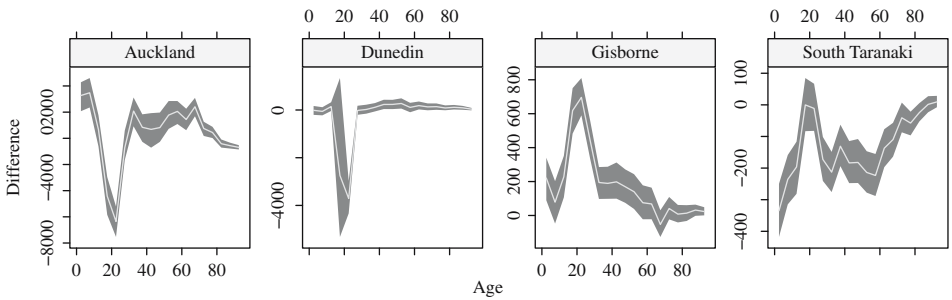Fig. 5. *Coverage rates from the revised model.*

Fig. 6.   Estimates from the revised model minus scaled 2013 census estimates. Values greater than 0 suggest that the model is overestimating the true population; values less than 0 suggest that it is underestimating.

selected regions (and in the remaining regions not shown here). The model consistently understates the number of young people in main centres, and overstates the number in towns and rural areas.

4.   Discussion

Statistics New Zealand is developing new methods for deriving population estimates by combining counts from multiple administrative data sources. The methods implicitly weight the various data sources in proportion to the consistency of their coverage rates. The process of deriving the implicit weights is automatic and data driven. Unlike traditional approaches to the study of coverage rates, no data source needs to be treated as the gold standard. Instead, the denominator for the coverage rates is generated within the model. The methods deal naturally with random variation in the population counts and data sources. All model outputs come with measures of uncertainty.

When weighting data sources, the model does not necessarily favour data sources with higher coverage rates. For instance, if one data source has an average coverage rate of 1.0 but is inconsistent across regions or time, while another data source has an average coverage rate of 0.4 but is highly consistent, then the model weights the second dataset higher than the first. If a higher coverage rate implies greater efficiency, then data sources with higher coverage rates will tend to be more consistent. However, the distinction between high coverage and consistent coverage is important. The ability to exploit data sources with low but consistent coverage rates is an advantage of cell-level approaches to population estimation.

A typical data model in our framework simply describes the empirical relationship between the data and the demographic process, without providing reasons for any discrepancies. For instance, none of the three models for administrative data in our application distinguish between discrepancies due to misaligned target populations, discrepancies due to reporting lags, and discrepancies due to processing error, though all three types of discrepancies are present in the data. Our approach to evaluating administrative data is thus complementary to approaches such as that of Zhang (2011) which seek to identify the specific sources of error. Results from such approaches are useful for our framework as a guide to the construction of data models. In return, our approach can provide estimates of the net effect of the various errors.

The application to New Zealand regional populations presented in this article is based on a relatively simple model. Models that were used for the production of official statistics

would typically be more elaborate. In particular, such models would typically be based on a full demographic account, containing births, deaths, and migrations, in order to exploit available data on these processes. The specify-estimate-evaluate cycle would be repeated many times, in the light of anomalies in coverage rates.

Nevertheless, the accuracy of any model, no matter how elaborate, is limited by the data available. If all data sources are subject to the same deficiencies, then data confrontation is unable to detect and correct for these deficiencies. An example is the overestimation of young people in rural areas and underestimation in urban areas in our modelling of regional populations in New Zealand. There is substantial evidence that administrative data systems in New Zealand miss many changes of address, or only capture them after a considerable lag (Statistics New Zealand 2013). Failure to update addresses has a greater effect on data for young people than on data for other age groups, because young people are much more mobile. The result is that administrative data for 'sending' regions contain too many young people, and administrative data for 'receiving' regions contain too few.

Such problems can be dealt with through a coverage or validation survey. The survey can be designed to respond to known problems with the administrative data. For instance, if administrative systems are failing to detect migrations by young people, then the survey can target these age groups, and ask questions about migration and the updating of addresses. The survey would yield data on true migrations versus reported migrations that could be supplied to the estimation model.

Data from the coverage survey could be included within the larger population estimation model as a special type of covariate. Models of the relationship between the coverage survey, the administrative data source, and the true population would need to include information about survey design and sample size, so that the survey data are given appropriate weight. The result would be coverage rates and population estimates that simultaneously took account of the survey data, the evidence from other data sources, and demographic plausibility.

Appendix: Further details on models

In the model for population,

$$\beta^Q = (\beta^0, \beta^{\text{age}}, \beta^{\text{sex}}, \beta^{\text{reg}}, \beta^{\text{age:sex}}, \beta^{\text{age:reg}}, \beta^{\text{sex:reg}}, \beta^{\text{time}}).$$

(For simplicity, we omit Q superscripts from the elements of β^Q.) Standard deviation σ_Q is given an improper uniform prior, as is intercept β^0, and the elements of sex effect β^{sex} and time effect β^{time}. The prior for the elements of age effect β^{age} is a second-order polynomial trend model, a type of dynamic linear model (Prado and West 2010, 119–120). The polynomial trend prior allows for the fact that neighbouring age groups are more likely to be similar than distant age groups. The standard deviations for the observation noise and state evolution noise in the age prior are assumed to be constant over time, and are given improper uniform priors. The elements of region effect β^{reg} are assumed to follow a Student-t distribution with a mean of 0 and 4 degrees of freedom.

In the initial version of the model, all interaction terms are given normal priors. The means of these priors are set to 0, and the standard deviations are given improper uniform priors. In the revised version of the model,

$$\beta_{ar}^{\text{age:reg}} \sim \text{N}(\delta + \gamma z_{ar}, \tau^2)$$

where Z_{ar} is 1 if a is age group 15–19 or 20–24 and r is "Auckland", "Christchurch", "Dunedin", "Hamilton", "Palmerston North", or "Wellington", and 0 otherwise. Parameters δ, γ, and τ are all given improper uniform priors. For identification, all subvectors within β are centered at 0 within the Gibbs sampler, with β^0 adjusted accordingly.

In the model for the tax data,

$$\beta^{\text{tax}} = (\beta^0, \beta^{\text{age}}, \beta^{\text{sex}}, \beta^{\text{age:sex}}).$$

Standard deviation σ_{tax} is given an improper uniform prior, as are intercept β^0, and the elements of sex effect β^{sex}. Age effect β^{age} is given a polynomial trend prior, identical to the prior for the age effect in the population model. Age-sex interaction $\beta^{\text{age:sex}}$ is given a normal prior with mean 0. The standard deviation for the age-sex prior is given an improper uniform prior.

The model for the health data is identical to the model for the tax data. In the model for the electoral roll data,

$$\beta^{\text{roll}} = (\beta^0, \beta^{\text{age}}).$$

Standard deviation σ_{roll} is given an improper uniform prior, as are intercept β^0, and the elements of sex effect β^{sex}. Age effect β^{age} is given a polynomial trend prior, identical to the prior for the age effect in the population model.

5. References

Brooks, S., and A. Gelman. 1998. "General Methods for Monitoring Convergence of Iterative Simulations." *Journal of Computational and Graphical Statistics* 7: 434–455. Doi: http://dx.doi.org/10.1080/10618600.1998.10474787.

Bryant, J. R. and P. J. Graham. 2013. "Bayesian Demographic Accounts: Subnational Population Estimation Using Multiple Data Sources." *Bayesian Analysis* 8: 591–622. Doi: http://dx.doi.org/10.1214/13-BA820.

Bycroft, C. 2013. *Options for Future New Zealand Censuses: Census Transformation Programme*, Technical report, Statistics New Zealand. Available at: http://www.stats.govt.nz/methods/research-papers/topss/options-future-nz-censuses.aspx (accessed 12 July, 2015).

Coleman, D. 2013. "The Twilight of the Census." *Population and Development Review* 38: 334–351. Doi: http://dx.doi.org/10.1111/j.1728-4457.2013.00568.x.

Gelman, A. and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/ Hierarchical Models*. Cambridge: Cambridge University Press.

Gibb, S. 2014. *Evaluating the Potential of Linked Data Sources for Population Estimates: IDI as an Example*. Technical report, Statistics New Zealand. Available at: http://www. stats.govt.nz/methods/research-papers/topss/evaluating-potential-linked-data-sources. aspx (accessed 12 July, 2015).

Office for National Statistics. 2013. *Beyond 2011: Options Explained 2*, Technical report, Office for National Statistics. Available at: http://www.ons.gov.uk/ons/guide-method/ census/2021-census/reports-library/beyond-2011-reports-archive/index.html (accessed 12 July, 2015).

Prado, R. and M. West. 2010. *Time Series: Modelling, Computation, and Inference*. New York: CRC Press.

Rees, P. 1979. "Regional Population Projection Models and Accounting Methods." *Journal of the Royal Statistical Society, Series A (General)* 142: 223–255. Doi: http:// dx.doi.org/10.2307/2345082.

Statistics New Zealand 2013. *Evaluating Administrative Sources for Population Estimates.* Technical report, Statistics New Zealand. Available at: http://www.stats.govt. nz/methods/research-papers/topss.aspx#censustransformation (accessed 12 July, 2015).

Stone, R. 1984. *The Accounts of Society*, Nobel Prize in Economics documents. Available at: http://www.jstor.org/stable/2951292 (accessed 12 July, 2015).

Zhang, L.-C. 2011. "A Unit-Error Theory for Register-Based Household Statistics." *Journal of Official Statistics* 27: 415–432.