View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Explore Bristol Research

Jones, K., Owen, D., Johnston, R., Forrest, J., & Manley, D. (2015). Modelling the occupational assimilation of immigrants by ancestry, age group and generational differences in Australia: a random effects approach to a large table of counts. Quality and Quantity, 49(6), 2595-2615. 10.1007/s11135-014-0130-8

Peer reviewed version

Link to published version (if available):
10.1007/s11135-014-0130-8

Link to publication record in Explore Bristol Research
PDF-document

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/pure/about/ebr-terms.html

Take down policy

Explore Bristol Research is a digital archive and the intention is that deposited content should not be removed. However, if you believe that this version of the work breaches copyright law please contact open-access@bristol.ac.uk and include the following information in your message:

• Your contact details
• Bibliographic details for the item, including a URL
• An outline of the nature of the complaint

On receipt of your message the Open Access Team will immediately investigate your claim, make an initial judgement of the validity of the claim and, where appropriate, withdraw the item in question from public view.

Modelling the occupational assimilation of immigrants by ancestry, age group and generational differences in Australia: a random effects approach to a large table of counts

Abstract
A novel exploratory approach is developed to the analysis of a large table of counts. It uses random-effects models where the cells of the table (representing types of individuals) form the higher level in a multilevel model. The model includes Poisson variation and an offset to model the ratio of observed to expected values thereby permitting the analysis of relative rates. The model is estimated as a Bayesian model through MCMC procedures and the estimates are precision-weighted so that unreliable rates are down-weighted in the analysis. Once reliable rates have been obtained graphical and tabular analysis can be deployed. The analysis is illustrated through a study of the occupational class distribution for people of different age, birth-place origin and generation in Australia. The case is also made that even where there is a full census there is a need to move beyond a descriptive analysis to a proper inferential and modelling framework. We also discuss the relative merits of Full and Empirical Bayes approaches to model estimation.

Keywords: tabular analysis of counts; log-Normal Poisson model; random effects; shrinkage; precision-weighted estimation; Bayesian estimation; Australian immigrant occupations

1. Introduction

The very large literature on the economic integration of international immigrants has identified substantial inter-generational differences in their occupational structures within particular countries and or cities, differing both from that of their host population and from each other. First-generation immigrants are most likely to be concentrated in relatively low status occupations, as suggested for the US in research initiated by Portes and Zhou (1993), for example, whereas studies of economic and social mobility show that greater proportions of second- than first-generation migrants occupy higher status roles and that the occupational structure of immigrant groups moves closer to the national mean, and hence to each other, in later generations (e.g. Boyd and Grieco, 1998, on the Canadian experience). Within that general pattern, however, there may be differences across immigrant groups reflecting variations in human, social and ethnic capital; those from markedly different backgrounds from the receiving society's may take longer to assimilate and achieve the levels of social mobility attained by those who differ less (Borjas, 1992).

In this paper we explore, using an innovative modelling procedure, whether such inter-group differences are characteristic of the occupational assimilation of first, second and third (or third-plus) generations of immigrants to Australia, which has attracted large numbers of settlers from a variety of cultural backgrounds since World War II (Jupp, 2001). The underpinning null hypothesis of our analyses is that, holding constant both generation and age, there should be no difference across a selection of different immigrant groups there in their occupational structures.

To evaluate that null hypothesis we model the dependent variable of rates of the number of individuals in each ancestry, age and generational group in each occupational category relative to the expected value if national rates prevailed (i.e. across all of the immigrant groups), using a bespoke table derived from the 2011 Australian census on the occupational structures of eight major immigrant groups. The cells in this table vary greatly in their size, so to evaluate differences between groups we develop a novel random-effects modelling approach, based on Poisson variation estimated as a Bayesian model using MCMC procedures. The resultant estimates are automatically precision-weighted so that those which are relatively unreliable – i.e. are based on relatively small numbers – are down-weighted in the analysis. This analysis then allows us to see the underlying patterns untroubled by underlying uncertainty.

2. International migration to Australia

Australia has experienced several, clearly distinguishable waves of immigration. Until the Second World War those arriving were dominated – in large part through a combination of cultural and political ties – by migrants from the United Kingdom and Ireland. After 1945, to meet the labour demands of a booming industrial economy a wider range of migrants was encouraged, but the 'White Australia' policy constrained this to European origins, of which the two main groups were Greeks and Italians. The 'White Australia' policy was abandoned in the early 1970s (Ho, 2013, 31), and substantial numbers of non-European migrants were attracted (Forrest et al., 2006: 443-446), many to the now burgeoning increasingly service-industry based economy from Northeast and Southern Asia (especially China and India). Most were highly skilled and educated (Hawthorne, 2005). At the same time, Australia

welcomed refugees, with substantial numbers from the war-torn areas of Lebanon and the Former Yugoslavia, both groups being divided into those from Christian and Muslim religious groups; many of these too were both skilled and well educated (Forrest et al., 2013, 190-194).

The data used here are taken from the 2011 Australian census, using the bespoke TableBuilder facility based on 100 per cent of the population.[1] For ten of the country's main migrant groups, we created a 5 (occupation) x 3 (generation) x 3 (age group) contingency table for all those (males and females aged 20-69) in the labour force. The five occupational groups identified were: managerial and professional; routine white collar (personal service, clerical and sales); skilled blue collar; semi- and un-skilled blue collar; and unemployed. The three generations were: first – born outside Australia; second – born in Australia to first generation immigrants; and third – born in Australia to second generation immigrants. The three age groups were: 20-29; 30-49; and 50-69.

The ten birthplace/ancestry groups differ substantially in their generational and age structures (Table 1); for the remainder of the paper we refer simply to ancestry groups. Generations have been variously defined; we deploy definitions suggested by Sweetman and Ours (2014, 3). The first generation comprises the respondent and both parents all born overseas, using birthplace data; second generation respondents were born in Australia but with both parents born overseas; in the third-plus generation the respondent and both parents were born in Australia. Information on respondent's ancestry was used to determine the ethnicity of the second (parents born overseas but where not stated) and third generations.

Those with UK and Irish backgrounds are predominantly third generation and less than one-fifth of them are aged 20-29; they are relatively older and well-established Australians (Table 1). Those from Greece and Italy, on the other hand, are mainly second generation settlers and middle-aged; and those from China and India are predominantly first-generation settlers with many more of them than average young adults – as are those from Yugoslavia and Lebanon, especially the relatively small numbers of Muslims in each group.

If the different generations and age groups have different occupational distributions – with members of the later generations more likely to be in the higher status, white-collar occupations, for example – this should be apparent in the occupational structures for the separate ancestry groups. But that does not appear to be the case as shown in Tables 2-3.

Fully two-thirds of the 4.5million individuals were in the two white-collar occupations, with slightly more in the managerial and professional than the clerical and sales category, and just over one-quarter were in the blue collar categories (almost equally divided among them); 4.5 per cent were unemployed. There were very few variations from this pattern by generation, the main one being the larger percentage unemployed in the first relative to the two subsequent generations (Table 2). There were slightly more substantial differences by age group (Table 3), notably in levels of unemployment (7.3 per cent for those aged 20-29 compared to 3.4 per cent for those aged 50-69) and presence in the managerial-professional group (a 10 percentage points difference between the youngest and oldest groups). The young and the recent arrivals were more likely to be unemployed; the old and the longest-

[1] For details on the TableBuilder facility see http://www.abs.gov.au/websitedbs/ censushome.nsf/ home/ tablebuilder – accessed July 29 2014.

established were more likely to be in the higher status occupations, but in general the differences were insubstantial.

Table 4 shows the occupational distributions for the ten ancestry groups. Again, there were few substantial differences in those percentages. In only three cases – both groups from the Former Yugoslavia plus Lebanese Muslims – were less than 60 per cent of individuals in the two white-collar occupational groups. Yugoslavs and Lebanese were more likely to be either unemployed or in blue-collar occupations than settlers from the four European countries, however, and there were also higher unemployment rates among Chinese and Indian settlers.

3. Statistical modelling of a large table of counts

These initial cross-classifications provide substantial support for our null hypothesis that despite very considerable differences in their age and generational structures there were few substantial differences across the ten ancestry groups in their occupational status. Whatever their origin, immigrants were equally likely to obtain a white-collar job. To evaluate that tentative conclusion further, we have undertaken statistical modelling of the data – for which, given the small numbers in the later generations, the two Former Yugoslav and two Lebanese groups have been combined, giving eight ancestry groups in total. The purpose of the modelling is to assess whether there are statistically significant differences across the ancestry by age by generation groups

The aim of this model-based exploratory approach to the analysis of a large table of counts is to discover the underlying patterns without being misled by unreliable rates based on small absolute counts. We have borrowed the approach from the field of disease mapping which is similarly concerned with finding pattern after first stabilizing incidence rates (Clayton and Kaldor (1987), although we use a full Bayesian (FB) analysis and not empirical Bayes (EB) (Owen and Jones, 2014). Our approach also has similarities to the method proposed for handling 'massively categorical data' (Steenburgh *et al* 2003). Importantly, we cast the method as multilevel or hierarchical random-coefficient models so that it can readily be implemented in general purpose software.

We first consider the need for modelling, and the standard approach of the log-linear model following Breslow and Day (1975). This we characterise as a fixed effects approach, and we then outline an alternative random-effects alternative and consider its advantages (Bell and Jones, 2014). The estimates generated by this approach are precision weighted so that rates based on small underlying counts are down-weighted during the modelling. Having secured more reliable rates we can then proceed to a graphical and tabular exploration of the differences.

3.1 Why model?

The dependent variable is the occupational class achievement ratio for each of the 360 cells or combinations based on 8 ancestry groups (the Christian and Muslim groups have been combined for Lebanon and the Former Yugoslavia) by 5 occupations by 3 generations by 3 age groups. The ratio is based on two numbers, the numerator being the observed count in each cell, and the denominator the expected number if the all-group rate applies. Thus 24,110 adults are unemployed, have UK ancestry, are aged 50-69 and belong to the third generation. If there are no differential occupational class effects we would expect 35,639 people to be in this category; this number equates to the proportion in that age group and generation who are

unemployed across all eight ancestry groups. This gives a ratio of 24,110/ 35,639=0.67; as it is below 1.0 this group enjoys a low relative risk of being unemployed. In contrast the ratio for first generation Indians aged 20-29 is 2.02, so that there is a doubling of the relative risk of unemployment in that group compared to the situation across all eight.

An important feature of these data is the considerable range in the observed and expected values that form the ratio – from 1 to over 375,000. Rates derived from the smaller numbers will be quite unreliable as a small change in absolute value could result in a large change in relative risk (Jones and Kirby, 1980). Consequently, it is important that an explicit modelling framework is developed in which the underlying stochastic variation due to small absolute counts is taken into account.

It is worth stressing that there is a need for modelling and inference even when we have a complete census of the Australian working population. This is contrary to the recent argument of Gorard (2013, 54), who contends that such an approach is not needed because 'there is no sampling variation'. But the observed count should be considered to be the outcome of a stochastic process which could produce different results under the same circumstances. It is this underlying process that is of interest and the actual observed values give only an imprecise estimate of this. The aim of the analysis therefore is not the descriptive statistic – the observed relative rate – but rather the parameter of the underlying rate in relation to the underlying uncertainty. That is we are interested in the weight of the evidence supporting a certain size of effect. In some cases even with a census there is so much natural variation around if we have a fine-grained table that we have to be careful about our judgments. The problem will only increase when further dimensions (for example Australian states) define the table. Consequently, and *contra* Gorard, it is important that an explicit modelling framework is developed to estimate confidence intervals in which the underlying stochastic variation due to small absolute counts is taken into account.

3.2 Specifying the fixed effects Poisson model for relative risk

We now outline a series of steps for how such a model can be specified and estimated. The first aspect that has to be dealt with is the chance element. If the process behind the distribution of counts in the cells of the table is entirely random, and the number of random events per cell has a constant mean rate of occurrence (π), and each event is independent, a Poisson distribution will result. A fundamental property of the Poisson distribution is that its mean is exactly equal to its variance, which is formulated as a Poisson regression model:

$$O_i \sim Poisson(\pi_i)$$
$$O_i \sim \pi_i + e_i$$
$$\pi = e^{\beta_0}$$
$$Log_e(\pi_i) = \beta_0$$
$$Var(O_i|\pi_i) = \pi_i$$

where O_i, the observed number of counts for cells of the table indexed by subscript *i*, is distributed as a Poisson distribution with an underlying mean rate of occurrence of π, plus a stochastic random term e_i. The mean rate is non-linearly related to any predictors as an exponential relationship that is transformed to a linear model by taking the natural logarithm (the log link). The variance of the observed values conditional on the underlying rate is equal to the underlying rate. The only parameter that needs to be estimated in this model is therefore β_0, the natural log of the rate of occurrence.

The second aspect is that we want to model the counts given the different numbers at risk of being in each occupational category. In a descriptive analysis, we would calculate the relative risk simply as the observed number divided by the expected, or the ratio of two counts:

$$RR_i = O_i / E_i$$

This can be achieved in the Poisson model by using an offset (McCullagh and Nelder, 1989). The relative risk is cast as a non-linear regression model:

$$\mathrm{E}(RR_i) = \mathrm{E}(O_i/E_i) = e^{\beta_0}$$

where the Expectation operator shows that we are averaging across all cells. Transforming this to a linear model, the division becomes a subtraction:

$$Log_e(O_i) - Log_e(E_i) = \beta_0$$

We can move $Log_e(E_i)$ to the right-hand side of the model and treat it like a predictor variable, but instead of estimating an associated regression-like coefficient, it is constrained to be 1:

$$Log_e(O_i) = Log_e(E_i) + \beta_0$$

This use of an offset treats the expected value effectively as a nuisance and allows us to model the underlying relative risk with the response simply being the log of the observed count. Using the log ensures that we cannot estimate a negative relative risk.

The third step in the modelling involves estimating the relative risk for different groups. We could do this by putting fixed-effects terms in to the model which are regression-like coefficients that are not constrained to the value 1 but are estimated from the data. Thus, to take a simple example, we could estimate the model for different age groups:

$$\begin{gathered} O_i \sim Poisson(\pi_i) \\ O_i \sim \pi_i + e_i \\ \pi_i = e^{(\beta_0 x_{0i} + \beta_1 x_{1i} + \beta_2 x_{2i})} \\ Log_e(\pi_i) = Log_e(E_i) + \beta_0 x_{0i} + \beta_1 x_{1i} + \beta_2 x_{2i} \\ Var(O_i|\pi_i) = \pi_i \end{gathered}$$

In this model the three age groups are represented by a constant (x_{0i}) which is just a set of 1s to represent the base category which we can arbitrarily select as the 20-29 age group, and two dummies (x_{1i} and x_{1i}) where a 1 represents the 30-49 and 50-69 age group respectively. The β_0 term, once exponentiated, gives the relative risk for the youngest category while exponentiating $\beta_0 + \beta_1$ gives the relative risk for the 30-49 group and $\beta_0 + \beta_2$ gives the relative risk for the oldest age group. Importantly the standard errors as well as the coefficients can be estimated in this generalized linear model taking account of the Poisson nature of the underlying counts (McCullagh and Nelder, 1989).

This specification can be readily extended to include further dummies and their interactions so that in the most complex or saturated model there are 360 fixed coefficients. This is however, somewhat unwieldy and an innovation of this paper is to use random not fixed effects to model the differences (Bell and Jones, 2014). We first examine the specification of

this multilevel model (Goldstein, 2011) and then outline its differences and advantages compared to the standard specification.

Specifying the random- effects Poisson model for relative risk

The model to be used is equivalent to a two-level Poisson model in its log-Normal form:

$$
\begin{gathered}
O_{ij} \sim Poisson(\pi_{ij}) \\
O_{ij} \sim \pi_{ij} + e_{ij} \\
\pi_{ij} = e^{(\beta_0 + u_j)} \\
Log_e(\pi_{ij}) = Log_e(E_{ij}) + \beta_0 + u_j \\
u_j \sim N(0, \sigma_u^2) \\
Var(O_{ij}|\pi_{ij}) = \pi_{ij}
\end{gathered}
$$

where individuals i are conceived as being in cells j which represent types of people (Subramanian *et al*., 2001). The key changes are in the line of the log link where the Log_e of the underlying rate (not the observed counts) is related to the offset, an overall intercept term β_0, and the u_j are the allowed-to-vary differential for each type of person. If this differential is positive, the cell has a higher risk than that expected; if negative, it is below that expected. Assuming that these differentials – the random effects - come from a Normal distribution[2] they can be completely summarized by their variance,σ_u^2, which measures the overall differences between cells having taking account of Poisson variation. We anticipate that, because the sum of the expected number of individuals in any occupational class is equal to the sum of the observed values, the β_0 term will be zero which when exponentiated becomes 1. The all-group underlying rate has therefore been standardized to 1. The u_j differentials are on the Log_e scale; the exponent of these values gives us the relative risk.

Unfortunately, we do not have individual data, due to confidentiality constraints, but only aggregate counts for cells. However, we can use the device of a 'pseudo-level' where cells are both the i's and j's in the model. Consequently, there is exactly the same set of units at level 1 and level 2, and each level 2 unit has exactly one level 1 unit. This may appear rather strange, but simply views the aggregate counts at level 2 as consisting of replicated responses for individuals at level 1. Indeed, as we know nothing else about these individuals except the values for the four defining variables we could reproduce the underlying individual data. We could create the individual data from the aggregate and vice versa and no information whatsoever is lost in going from one scale to the other. This device allows for extra Poisson variation in the same manner as Browne *et al* (2005) and Leckie et al (2012) achieved for over-dispersed Binomial multilevel models. In essence there are two sources of variation that need to be separated – variation due to true between cell variation and that due to natural Poisson variability. Equivalently the lower level of the model is used to model the natural

[2] The Normality assumption of the cell differentials is obviously a key assumption for the validity of the variance in summarising the differences in the relative risk. This can be informally assessed with a Normal probability plot. In practice we have found that this assumption is generally met; no doubt due to using the log transform. Moreover, McCulloch and Neuhaus (2011) have found model results are generally robust to the shape of the random-effects distribution. An exception to this would be marked outliers for particular cells which could be accommodated by specifying separate fixed effects for these cells which would make them immune to shrinkage.

variation of a Poisson variable while the higher level is used to model the extra Poisson variation of the true rate.

3.3 Comparing fixed and random effects

To appreciate conceptually the properties of the precision-weighted multilevel estimates it is useful to compare the fixed- effects model with their random-effects equivalent. The estimates from the saturated fixed- effects model are given by

$$Log_e(\pi_i) = Log_e(E_i) + \beta_1^* D_{1i} + \beta_2^* D_{2i} + \cdots \beta_{360}^* D_{360i}$$

where there are 360 dummy variables, one for each cell. The β_j^* is the log relative rate for each cell which when exponentiated will be exactly equivalent to the simple ratio ($\frac{O}{E}$), the observed to expected rate for each and every cell. The equivalent random effects model is

$$Log_e(\pi_{ij}) = Log_e(E_{ij}) + \beta_{0j}$$
$$\beta_{0j} = \beta_0 + u_j$$
$$u_j \sim N(0, \sigma_u^2)$$

so that β_{0j} is again the log relative rate but this time it is assumed to come from an overall distribution with a mean of β_0 and a variance of σ_{u0}^2.

The two sets of estimates can be related as follows:

$$\beta_{0j} = \beta_0 + u_{j=} w_j \beta_j^* + (1 - w_j)\beta_0$$

where the weight $w_j$is given by the reliability of a particular cell

$$Reliability\ of\ cell\ j = w_j = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + (\sigma_{e0j}^2)}$$

The σ_{u0}^2 is the between cell variance of the log differentials which gives the true differences between cells, while σ_{e0j}^2 gives the natural or stochastic variation of the rate for a particular cell based on a Poisson variable. The approximate[3] standard error (Breslow and Day, 1987, equation 2.9) of the log rate of each cell is given by

[3] For ease of exposition (and as per normal practice) the imprecision in the ratio is dependent only on the imprecision of the observed count. It is being assumed that the expected count is precise. A more realistic formulation is given in Talbot et al (2011). The specific nature of the weighting for this log-Normal model is considered by Papageorgiou and Ghosh (2012, equations 1 to 3); albeit in an empirical Bayes formulation.

$$SE\left(Log\left(\frac{O_j}{E_J}\right)\right) = \frac{SE\left(\frac{O_j}{E_J}\right)}{\left(\frac{O_j}{E_J}\right)} = \frac{\frac{\sqrt{O_j}}{E_j}}{\left(\frac{O_j}{E_J}\right)} = \frac{1}{\sqrt{O_i}}$$

so that the stochastic variance of the log rate is:

$$Var(Log\left(\frac{O_j}{E_J}\right)) = \frac{1}{O_j}$$

Substituting this stochastic variance of the log rate into the formula for the weights gives

$$Reliability\ of\ cell\ j\ = w_j\ = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + (\frac{1}{O_i})}$$

This equation which formulates a signal-to-noise ratio has strong intuitive appeal.[4] The reliability is equal to the proportion of the variance in the observed log rates that we could explain if we knew the true rates. Weights are a function of the between cell variation across all cells (the variance[5] is a summary of the differences between types of people) and the stochastic uncertainty of a particular cell which is dependent on the absolute size of the counts. Reliability will therefore be at a maximum when there are true sizeable differences between cells and when the count is large to give a precise estimate of the log rate in any particular cell. The reliability becomes higher as the proportion of stochastic variance in the observed log rates becomes lower and vice versa.

We can now see what happens in the equation relating the random to the fixed effects. If there large differences between cells (σ_{u0}^2) and the observed count is large the weight will approach 1 and the equation becomes:

$$\beta_{0j} = w_j\beta_J^* + (1 - w_j)\beta_0 = 1 * \beta_J^* + (1 - 1)\beta_0 = \beta_J^*$$

The random-effects estimate will be exactly the same as the fixed effects estimate and no shrinkage will take place. However, if there are small differences between cells and a cell has a small absolute count, the weight approaches 0, and equation becomes:

$$\beta_{0j} = w_j\beta_J^* + (1 - w_j)\beta_0 = 0 * \beta_J^* + (1 - 0)\beta_0 = \beta_0$$

[4] The weight is a form of interclass correlation coefficient for each cell that measures the amount of true variability (the level 2 variance) in the underling rates relative to the total observed variability. In the measurement literature, the reliability w_j is often symbolised by ρ_{yy} to convey the internal dependency of a measured y variable.

[5] The between cell variance at level 2 summarizes the differences between cells, but usefully it is not the variance of the shrunken differentials, but the variance of the raw differentials. Consequently it is not the estimated between group variance of the *sample*, but the estimated between-group variance in the *population*

The specific cell estimate will be shrunk back towards the mean estimate of all cells which due to the standardization used here will be the value zero. Consequently, the exponeniated random-effects estimate will be shrunk back to the value 1 which represents no difference between the observed and expected value, the overall Australian rate. The weight represents the proportion of the information that is being obtained locally from each cell as compared to information derived nationally from all cells.

Another way of looking at this is in terms of pooling of information (Jones and Spiegelhalter, 2011, 155-156). When the between-cell variance is equal to zero, that is identical rates for all type of people, the weight is zero so that relative risk is set to the overall mean. This is known as complete pooling and every cell gets the same value. As the level 2 variance increases the weight approaches 1 and there is no pooling of information between cells and the estimate is set to the raw rate. Each cell is maximally different from all others and in effect the between-cell variance is constrained to infinity in the fixed-effects approach. Between these two extremes, there is partial pooling where the degree of pooling is determined by data – it is the relative size of the between cell variance to the stochastic or measurement error variance that is driving the degree of shrinkage. Thus the random effects approach is a data-driven adaptive procedure which handles the uncertainty that is inherent in working with sparse data (Gelman, 2014).

Cells with small counts are likely to appear to be extreme by chance and although the fixed effects are unbiased they are troubled due to imprecision. The precision-weighted estimate is biased towards the overall mean but they are 'optimal' in having a smaller squared error between the estimate ($\hat{\beta}_{0j}$) and the true value (β_{0j}) averaged across many cells; that is the mean square error will be smaller. The multilevel estimates will minimize the following function (Jones and Bullen, 1994):

$$MSE = \sum(\hat{\beta}_{0j} - \beta_{0j})^2 = Var(\hat{\beta}_{0j}) + Bias(\hat{\beta}_{0j})^2$$

A small amount of bias is being traded for a large reduction in the measurement error variance, the shrunken estimates are useful because they are more precise.[6] For cells with enough data, we are more concerned about bias and want to shrink less, while for cells with less data, we are more concerned about variance and want to shrink more. The adaptive procedure does it automatically providing we get a good estimate of the between cell variance.

Another important advantage of the shrinkage approach is in relation to multiple comparisons. As demonstrated by Gelman *et al* (2012) it is much more efficient to shift estimates towards each other rather than try to inflate the confidence intervals in such procedures using a Bonferroni correction to control the overall error rate. Thus shrinkage automatically makes for more appropriately conservative comparisons while at the same time

[6] For a more general discussion of these advantageous properties see the classic papers of James and Stein (1961), and Lindley and Smith, (1972). Their benefits are extolled in Kendal's (1959) 'song', and in the expository paper of Bradley and Morris, (1977) which studies baseball averages and disease distributions.

not reducing the power to detect true differences. The final advantage is dealing with zero counts. With raw rates if the numerator of a cell is zero, then the associated rate can only be zero. But a zero based on a denominator of 1 and a denominator of a 100 mean quite different things – for the former it is uncertain whether the rate is really zero ; for the latter we can be quite confident. The random effects estimate shrink more towards the overall rate for the former than for the latter.

Figure 1 shows the relationship between the raw log rate (or equivalently the estimate derived from the saturated fixed effects model including all 360 parameters) and the Bayes-estimated log rate. Most of the rates fall on the diagonal, indicating that the multi-level precision-weighted estimates are the same as the raw rates. There are a number of cells however which experience substantial shrinkage to the overall mean of zero. These are characterised by being rather extreme in their raw form and having a weight below 0.9 so that more than ten per cent of their 'information' is borrowed (to use Tukey's felicitous phrase) from the overall national rate. Such cells are characterised by low counts; indeed the median observed count for those cells with a weight below 0.9 is only 27. The appropriateness of the Normality assumption was evaluated by a Normal probability plot and there was no evidence of marked outliers or skewness in the allowed–to-vary log differentials.

3.4 Model estimation for random effects: Empirical Bayesian (EB) and Full Bayesian (FB) procedures

We have so far discussed the properties of the estimates but not how to estimate the models. An important distinction is between EB and FB. Detailed technical comparison of EB and FB for the Poisson model is given by Bernardinelli and Montomoli (1992); here we briefly convey the underlying concepts and consider the details and choices needed from practical application.

Bayesian modelling is all about three distributions. The prior distribution is about subjective belief – what you think is the distribution of support for a parameter; the likelihood is the degree of support for different values of a parameter based on the data and assumptions; the posterior combines the prior and the likelihood and gives the evidential support for different values of a parameter. In the random-effects model, the allowed-to-vary differentials are assumed to come from a distribution with a mean and a variance. In the Bayesian formulation there is another layer whereby the mean and the variance are additionally assumed to come from a hyperprior distribution that in turn have hyperprior parameters. In EB these hyper parameters are estimated directly from the data but this is un-Bayesian as it assigns a point estimate and does not allow for inherent uncertainty; EB produces not the true posterior but its approximation. FB in contrast assigns hyperprior distributions to these hyper parameters so that in a full Bayesian model everything is a distribution. EB estimates the hyper parameters from the marginal distribution of the observations whereas FB takes account of the full multivariate distribution – so called full error propagation. The practical importance is that FB modelling allows the calculation of standard errors and confidence intervals without having to rely on asymptotic Normality assumptions that are unlikely to hold in applications with a relatively small number of cells.

Taking the model as specified above we can add in the specification of the hyper priors.

$$O_{ij} \sim Poisson(\pi_{ij});\ O_{ij} \sim \pi_{ij} + e_{ij}$$
$$\pi_{ij} = e^{(\beta_0 + u_j)}$$
$$Log_e(\pi_{ij}) = Log_e(E_{ij}) + \beta_0 + u_j$$
$$u_j \sim N(0, \sigma_u^2);\ Var(O_{ij}|\pi_{ij}) = \pi_{ij}$$
$$p(\beta_0) \propto 1; \qquad p(\frac{1}{\sigma_u^2}) \sim Gamma(0.001, 0.001)$$

The final line gives the hyperprior assumptions that we have used to derive the results of Figure 1 which are chosen so as to impose as little information as possible on the data. The probability prior for β_0 (overall all cell mean) is a uniform distribution in which any value is equally likely; an alternative would be to put a tight prior around 0 (that is 1 on the raw scale) to constrain the overall rate to the standardization we have used (Bell and Jones, 2014). It makes little difference in this application as this estimate is based on all the data and is a weighted average of the cell estimates, weighted to emphasize the reliable estimates.. The prior for the inverse of the between-cell variation ($\frac{1}{\sigma_u^2}$, known as the precision) is assumed to be a Gamma distribution in which the shape and scale parameters are small values close to zero. The Gamma distribution is a flexible one and can accommodate marked positive skewness which is appropriate for a variance parameter that cannot go negative. It has been found that FB estimation for these models is quite robust to the specification of the hyperprior for Gamma (Bernardinelli *et al* 1995) and we found here that the replacement of the Gamma distribution by a uniform prior in a sensitivity analysis made very little difference.

While both approaches have no closed form solutions, FB estimation has proved more challenging than EB. There are now various procedures for estimated EB-based parameters that are likelihood based and iterative algorithms have been produced which converge to point estimates. In recent years complex iterative sampling schemes – so- called Markov Chain Monte Carlo procedures – have been developed to generate the full posterior distribution required for FB estimates. These MCMC approaches allow a building block approach to estimation whereby complex problems are decomposed into lots of small ones. MCMC procedures are a way of evaluating the full joint posterior distribution of all parameter estimates by simulating a new value for each parameter in turn from its marginal distribution assuming that the current values for the other parameters are the true values (Jones and Subramanian, 2014).

Leyland and Davies (2005) compare a number of these different procedures for estimating the basic model. They note that it has been argued that EB in using estimates of the hyper parameters does not take into account their uncertainty resulting in potential under-estimation of true rates and too much shrinkage towards the mean. Moreover, the Full Bayes also gives the distribution of support for each estimate for each cell allowing credible intervals to be calculated. However, their review of the comparisons that have been made suggests little real difference and we found in the present analysis that there is virtually no difference between

EB (enabled through penalized quasi likelihood with the IGLS algorithm) and Full Bayes enabled through MCMC (using a combination of Metropolitan Hastings sampling of the fixed and random effects and Gibbs sampling of the higher-level variance). Much depends on the number of cells and the degree to which the random effects approximates a Normal distribution and we recommend using both approaches to appreciate the sensitivity of results to assumptions. Modern high-speed computers mean that highly-computational intensive FB takes just a few minutes even when 100,000 cycles of simulations are used. From our experience of this and other datasets we would not however recommend the used of the EB procedure known as marginal quasi-likelihood for Poisson models as this results in (unlike the Binomial case, Rodriguez and Goldman 1995) considerable over-estimation of the between cell -variance in comparison to both PQL and Full Bayes estimates. Although the MQL procedure is computationally quick and less prone to convergence problems it is too biased for routine use.

The FB procedure has a number of important by products. It is possible to monitor the chain of estimates for each of the cell-based random effects so that it possible to calculate asymmetrically distributed credible intervals for these estimates – although it made negligible difference here. It is also possible to calculate functions of the cell differences to monitor differences for particular cells and to produce a rank of the differentials from the Australian average and have credible intervals of the ranks. Another important by-product of the MCMC estimation is the Deviance Information Criterion (Spiegelhalter *et al* 2002) which has become a popular tool for choosing between models in terms of their predictive capacity. The DIC is a badness of fit measure penalized for model complexity. It is the sum of the posterior mean deviance (minus twice the log likelihood) representing the degree of fit, plus the effective degrees of freedom (pD), reflecting model complexity. The latter, which is the main interest here, is well defined in classical models as the count of the number of parameters in the model. However in the FB model the shrinkage imposed by hyperprior distribution effectively restricts this value. Indeed for approximately Normal likelihoods it can be shown (Best *et al*, 2005) that pD is the ratio of the information in the likelihood to the total information in the posterior distribution (that is the likelihood plus the prior).[7] When pD is close to the number of cells there is little shrinkage and the hyperprior does not greatly influence the results and there is little borrowing of strength. But large differences between estimated and nominal degree of freedom implies that the prior is providing a lot of information with considerable smoothing and structuring of the results. In the present analysis, the saturated fixed effects model has 360 parameters, one for each cell, while the pD is estimated to be 345. Consequently and as shown by Figure 1 there is very modest shrinkage with 96 percent (345/360) coming from the data and only 4% from the prior. However this small percentage is valuable in smoothing extreme rates of some stringly affected cells..

[7] The estimate of pD is given by the difference between the average deviance and the deviance at the expected value of the unknown parameters.

All of the analysis was carried out with the MLwiN software which has a range of EB (quasi-likelihood) and FB (MCMC) procedures (Jones and Subramanian; 2014; Rasbash et al, 2009; Browne, 2012). We followed the good practice recommendations of Draper (2008) to ensure that the MCMC chain has been run sufficiently long to characterise the posterior distribution. Specifically we initially used quasi-likelihood PQL estimates to produce reasonable starting points for the simulation; discarded a burn-in of 500 simulations to potentially get away from the PQL estimates, and then run a further 100,000 monitoring simulations. We found it beneficial to use hierarchical centering to obtain less correlated chains (Browne, 2012) and the resultant monitoring estimates had the information equivalent of 82k and 64k independent draws for the overall mean and level-2 variance respectively; plenty of information to evaluate the distribution. The between cell variance on the log scale was 0.107 and the 95% credible intervals are 0.092 and 0.125. These values are simply defined as the mean and lowest and highest 2.5% of the simulations for the posterior distribution; their symmetry around the mean suggests that the Gamma posterior distribution in fact approximates a Normal distribution.

4. The results

The output from the modelling process is a shrunken estimate for each of the 360 cells and its associated 95% confidence intervals. They are used here in two ways. First, a tabular analysis establishes whether the modelled rates are significantly larger than or less than 1.0: in the former case, this requires that the confidence interval around the modelled rate has a lower limit greater than 1.0; in the latter case, that the modelled rate has an upper limit less than 1.0. The third use establishes whether there were significant differences between ancestry groups in their logged observed/expected ratios within each occupational, generational and age group.

The approach taken in presenting and interpreting these results is illustrated in Table 5, which refers to those who were aged 20-29, in the first generation, and unemployed. For the first set of interpretations, seven of the eight modelled rates exceed 1.0 and have confidence intervals whose lower limits are also greater than 1.0; the only exception is for Ireland, which has a modelled rate of 0.987 but a confidence interval with its upper limit exceeding 1.0. Thus all of the ancestry groups except Ireland had a significantly higher proportion of their number of first-generation individuals aged 20-29 who were unemployed than was the case across all age groups and generations. Unemployment was concentrated among the young, first-generation settlers.

4.1 Differences by age and generation

For each age group and generation, Table 6 gives the number of modelled rates significantly above and below 1.0, across the eight ancestry groups, by occupational group. Some very clear patterns emerge. In the professional occupational category, for example, in each generation the majority of those aged under 30 have modelled rates significantly below 1.0 (the maximum possible number of values in each cell is 8): whatever their generation, fewer young people than expected are in the highest status occupational class. In the second and third generations, on the other hand, those aged over 30 are significantly more likely than expected to be in that occupational status group. At the other end of the status scale, those

aged 20-29 are significantly more likely than expected to be unemployed – and people in that age group are also more likely to be employed in skilled blue-collar or technical occupations.

The rows at the foot of the table summarise the variations between the generations and age groups. For the former, in each generation there is a possible maximum of 24 modelled ratios that are significantly different from 1.0 and in only one case – for the third generation in clerical occupations – is that number smaller than one-half of the maximum; most of the modelled rates for each generation for each occupational class are significantly different from the expected according to the null model, therefore, indicating substantial generational differences in occupational structures. This is the case also for each of the age groups, although in two cases the number of significantly-different modelled rates is only just 12; occupational structures also differ significantly by age.

Overall, those in the youngest age group are significantly concentrated in the clerical and technical/trade occupations and among the unemployed, and are significantly under-represented in the professional/managerial class and in the semi- and unskilled occupations. The middle-aged and older settlers, on the other hand, are significantly under-represented among the unemployed and over-represented in the highest status occupations, and the same pattern applies across the generations – the first generation are most likely to be unemployed, the later two generations to be in the highest status occupations. Across the eight ancestry groups, the maximum possible number of modelled ratios significantly different from the expected is 45. All had 25 or more, with only two – India and the former Yugoslavia – having less than 30; none had more than 38. In general, therefore, in a clear majority of cases the occupational structure across all ancestry groups differed significantly by age and generation.

4.2 Differences between ancestry groups

But were there differences between the ancestry groups as well? For the second set of interpretations, the eight ancestry groups are arranged in Table 5 according to their modelled ratio (observed/expected number of unemployed). If an ancestry group has a significantly lower modelled rate than that immediately above it in the table, this is shown in bold: a significant difference occurs where the confidence intervals for the two groups do not overlap. Thus the Lebanese have a significantly lower, precision-weighted, modelled unemployment rate than the Chinese; the range between the confidence intervals for the latter – 3.691:4.005 – does not overlap that for the former – 2.718:3.347. Similarly, there is a significant difference in the modelled rates (3.010 and 2.066 respectively) for Lebanese and Indians, but none between the Italians, Greeks and Yugoslavs. Finally, the modelled rate of 1.271 for UK immigrants is significantly lower than that for Yugoslavs, and that for immigrants of Irish descent significantly lower again than that for those from the UK. Overall, therefore, those from English-speaking backgrounds have significantly fewer of their first-generation, 20-29-year-olds unemployed than expected, whereas those from Asian origins (China, Lebanon and India) have significantly more: those from European backgrounds have neither significantly more nor significantly fewer unemployed than expected for that generation and age group.

Table 7 illustrates the use of this procedure for the unemployment occupational group. For each generation and age group, it lists the eight ancestry groups according to their modelled rates, with those that are significantly different from the ancestry immediately above it identified in bold. One clear conclusion from this table is that significant differences between ancestry groups in their modelled unemployment rates occur in a minority of cases only: of

the 63 paired differences (i.e. seven in each of the table's nine segments) only 18 are statistically significant. Nine of these apply to the first generation, seven to the second, and only two to the third; seven apply to those aged 20-29, eight to those aged 30-49, and three to those aged 50-69. In general, therefore, significant differences in unemployment rates between countries are much more likely to be found among the more recent and younger arrivals than among their older, longer-established contemporaries.

Looking at the ordering of the ancestry groups in Table 7 indicates also that, in general, the significant differences are across the four waves of immigrant groups – UK/Ireland; Greece/Italy; China/India; and Lebanon/Yugoslavia. Thus among first- and second-generation arrivals, in all three age groups, those from Lebanon, China and India have the highest modelled rates and those from the UK and Ireland the lowest. This clear sequence is absent from the third-generation orderings, however – where in any case there are virtually no significant differences.

Similar analyses have been undertaken for each of the other four occupational groups, but the detailed tables are not reproduced here; they are summarised in Table 8 which shows the number of significant differences between modelled rates for each occupational category, generation and age group. A number of conclusions stand out. First, there are more significant differences in the lower status occupations and in unemployment levels than in the two white-collar categories; secondly, there are many more significant differences in the first than the other two generations (45 of the 69, compared to 16 and eight in the other two respectively); and thirdly, there are more significant differences in the two younger age groups (20 and 31 for 20-29 and 30-49 year-olds respectively) than for the older (18). As with the discussion of unemployment rates alone, therefore, this table shows that there are many more differences in the occupational composition of the various ancestry groups in their earlier generations of settlers, and the younger age groups within each of those, than among the older and longer-settled residents. Our null hypothesis of no difference is clearly confirmed for the latter, therefore, but not for the former.

5. Conclusions

This paper has introduced a method for analysing large contingency tables in which the cell sizes differ substantially. Based on a procedure developed for the analysis of disease patterns it derives precision-weighted estimates of the ratio of observed to expected values for each of the contingency table's cells, and Bayesian-derived confidence intervals for each of those estimates (derived through the deployment of a random effects multi-level model). Once stabilized rates have been achieved graphical and tabular analysis can be deployed to examine the patterns.

Use of this novel procedure has been illustrated using a large contingency table showing the occupational structure of eight of the largest ancestry groups in contemporary Australia, by age and generation. This has very largely sustained the null hypothesis that although there are significant differences between age groups, between generations, and between age groups within generations, taking those differences into account there are few significant variations (within each age, by generation, by occupation segment of the table) between the eight ancestry groups – and where there is, it is concentrated among the younger and recent settlers. To the extent that immigrants experience disadvantage in the operation of the Australian labour market, therefore, such disadvantage is not greater for some immigrant groups than others.

The modelling framework introduced here has a wide range of potential applications where researchers are evaluating differentials within large and complex tables. The modelling smooths the estimates towards overall rate when the local cell information is unreliable but preserves patterns with high statistical significance. Moreover there is no need to adjust the values for multiple comparisons and the procedure is readily implemented using existing software.

References

Bell, A. and Jones, K. (2014) Explaining fixed effects: random effect modelling of time series, cross-sectional and panel data, *Political Science and Research Methods*, available online at doi 10.1017/psrm.2014.7.

Bell, A and Jones, K (2014) Bayesian informative priors with Yang and Land's hierarchical age–period–cohort model, *Quality and Quantity*, in press, DOI 10.1007/s11135-013-9985-3

Bernardinelli ,L. Clayton, D. and Montomoli, C. (1995) Bayesian estimates of disease maps: how important are priors? *Statistics in Medicine*, 14: 2411–2431

Bernardinelli L. and Montomoli, C. (1992) Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk, Statistics in Medicine 11: 983-1007.

Best, N., Richardson, S., and Thomson, A. (2005) A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research* 14: 35–59.

Borjas, G. J. (1992) Ethnic capital and intergenerational mobility, *The Quarterly Journal of Economics*, 107: 123-150.

Boyd, M. and Grieco, E. (1998) Triumphant transitions: Socioeconomic achievement of the second generation in Canada, *International Migration Review*, 32: 853-876.

Breslow NE. and Day NE. (1975) Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data , *Journal of Chronic Diseases*, 28: 289-303.

Breslow NE. and Day NE. (1987) *Statistical methods in cancer research, volume II: The design and analysis of cohort studies*, International Agency for Research on Cancer, Lyon.

Browne, W. J. (2012) *MCMC Estimation in MLwiN, v2.25*. Bristol: University of Bristol, Centre for Multilevel Modelling, available at http://www.bristol.ac.uk/cmm/software/mlwin/download/manuals.html

Browne, W. J., Subramanian, S. V., Jones, K. and Goldstein, H. (2005) Variance partitioning in multilevel logistic models that exhibit over-dispersion, *Journal of the Royal Statistical Society, Series A*, 168: 599-614.

Clayton, D. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardized relative risk for use in disease mapping, *Biometrics*, 43: 671-681.

Draper, David (2008) Bayesian multilevel analysis and MCMC, in de Leeuw, J and Meijer. E (eds.) *Handbook of Multilevel Analysis*. Nrew York: Springer, 77–139.

Efron, B. and Morris, C. (1977) Stein's paradox in statistics. *Scientific American*, 237: 119–127.

Forrest, J. Hermes, K., Johnston, R. and Poulsen, M. (2013) The housing resettlement of refugee immigrants to Australia, *Journal of Refugee Studies*, 20:187-206.

Forrest, J., Poulsen, M. and Johnston, R. (2006) A 'multicultural model' of the spatial assimilation of ethnic minority groups in Australia's major immigrant-receiving cities, *Urban Geography*, 27: 451-463.

Gelman, A (2014) How Bayesian analysis cracked the red-state, blue-state problem, *Statistical Science*, 29(1) 26-35.

Gelman, A., Hill, J. and Yajima, M.,(2012) Why we (usually) don't have to worry about multiple comparisons, *Journal of Research in Educational Effectiveness* 5: 189–211.

Goldstein, H. (2011) *Multilevel Statistical Models* (4th edition). Chichester: John Wiley.

Gorard, S**.** (2013) *Research Design: Robust approaches for the Social Sciences*, London: Sage.

Hawthorne, L. (2005) "Picking winners": the recent transformation of Australia's skilled migration policy, *International Migration Review*, 39: 663-696.

Ho, C. (2013) From social justice to social cohesion: a history of Australian multicultural policy, in Jakubowitz, A. and Ho. C. (eds) *For Those Who've Come Across the Seas: Australian Multicultural Theory, Policy and Practice*, North Melbourne: Australian Scholarly Publishing, 31-44.

James, W. and Stein, C (1961), Estimation with quadratic loss, *Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, 1: 361–379

Jones H E., and Spiegelhalter, D J. (2011) The identification of 'unusual' health-care providers from a hierarchical model. *The American Statistician* 65(3): 154-163.

Jones, K. and Bullen, N. (1994) Contextual models of urban house prices: A comparison of fixed- and random-coefficient models developed by expansion. *Economic Geography* 70: 252-272.

Jones, K and Kirby, A (1980) The use of chi-square maps in the analysis of census data, *Geoforum*, 11: 409-417.

Jones, K. and Subramanian, S. V. (2014) *Developing Multilevel Models for Analysing Contexuality, Heterogeneity and Change*. Bristol: University of Bristol, Centre for Multilevel Modelling – available at http://www.bristol.ac.uk/cmm/software/ mlwin/mlwin-resources.html

Kendall, M. G. (1959) Hiawatha designs: an experiment, *The American Statistician* 13: 23-24.

Jupp, J. (ed.) (2001) *The Australian People: an Encyclopedia of the Nation, its Peoples and their Origins*. Oakleigh: Cambridge University Press.

Leckie, G., Pillinger, R., Jones, K. and Goldstein, H. (2012) Multilevel modelling of social segregation, *Journal of Educational and Behavioral Statistics*, 37: 3-30.

Leyland A. H. and Davies, C. A. (2005) Empirical Bayes methods for disease mapping *Statistical Methods in Medical Research* 14:17–34.

Lindley, D. and Smith, A. (1972) Bayes estimates for the linear model, *Journal of the Royal Statistical Society Series B*, 34: 1-41.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*. London: Chapman and Hall

McCulloch CE and Neuhaus JM. (2011) Misspecifying the shape of a random effects distribution: why getting it wrong may not matter, *Statistical Science*, 26,:388–402.

Owen, D. and Jones, K. (2014) Geographical inequalities in mortality: a model-based approach to analysing fine-grained differences over time: England and Wales, 2002-2012, in preparation

Papageorgiou, G. and Gosh, M. (2012) Estimation of small area event rates and of the associated standard errors, *Journal of Statistical Planning and Inference*, 142: 2009-2016.

Portes, A. and Zhou, M. (1993) The new second generation: segmented assimilation and its variants, *Annals, American Academy of Political and Social Science*, 530: 74-96.

Rasbash, J., Charlton, C., Browne, W. J., Healy, M. and Cameron, B. (2009) *MLwiN Version 2.1*. Bristol: University of Bristol, Centre for Multilevel Modelling.

Rasbash, J., Charlton, C., Jones, K. and Pillinger, R. (2012) *Manual Supplement to MLwiN v.2.26*. Bristol: University of Bristol, Centre for Multilevel Modelling.

Rodriguez, G., and Goldman, N. (1995) An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, A*, 158:73-90.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* 64: 583-640.

Steenburgh, T. J., Ainslie, A. and Engebretson, P. H. (2003) Massively categorical variables: revealing the information in zip codes, *Marketing Science*, 22: 40-57.

Subramanian, S. V., Duncan, C. and Jones, K. (2001) Multilevel perspectives on modelling census data. *Environment and Planning A*, 33: 399-417.

Sweetman, A. and van Ours, J. C. (2014) Immigration: what about the Children and Grandchildren? Bonn: Institute for the Study of Labour, IZ A Discussion Paper 7919.

Talbot, D., Duchesne, T., Brisson, J., Vandal, N. (2011) Variance estimation and confidence intervals for the standardized mortality ratio with application to the assessment of a cancer screening program, *Statistics in Medicine*, 30: 3024-3037.

Table 1. The generational and age structures of the ten ancestry groups in Australia – percentages of those in the workforce.

	Generation			Age Group		
	1st	2nd	3rd	20-29	30-49	50-69
UK	16	8	76	19	46	35
Ireland	8	4	88	20	45	35
Greece	19	72	9	13	63	24
Italy	18	57	25	19	52	29
FYugoslavia Christian	51	47	2	18	56	26
FYugoslavia Muslim	86	13	1	33	51	16
China	86	11	3	37	43	20
India	95	4	1	39	52	9
Lebanon Christian	48	48	4	22	50	28
Lebanon Muslim	52	47	1	38	51	11
TOTAL	24	13	63	21	47	32

Table 2. Generational differences in occupational structures: all ancestry groups (percentages in brackets)

	Generation			
Occupation	1st	2nd	3rd	TOTAL
Managerial/Professional	394,969	204,614	1,018,094	1,617,677
	(37.3)	(35.6)	(35.6)	(36.2)
Clerical/Sales	303,662	196,421	913,677	1,413,760
	(28.6)	(34.3)	(32.2)	(31.7)
Technicians/skilled trades	145,760	80,897	378,995	605,652
	(13.7)	(14.1)	(13.4)	(13.6)
Semi- and Unskilled	154,300	67,697	404,777	626,774
	(14.6)	(11.8)	(14.3)	(14.0)
Unemployed	61,463	23,852	117,138	202,453
	(5.8)	(4.2)	(4.1)	(4.5)
TOTAL	1,060,154	573,481	2,832,681	4,466,316

Table 3. Age-group differences in occupational structures: all ancestry groups (percentages in brackets)

	Age Group			
Occupation	20-29	30-49	50-69	TOTAL
Managerial/Professional	252,333	823,687	541,657	1,617,677
	(27.2)	(39.0)	(37.9)	(36.2)
Clerical/Sales	330,951	639,833	442,976	1,413,760
	(35.8)	(30.3)	(31.0)	(31.7)
Technicians/skilled trades	153,823	285,283	166,546	605,652
	(16.6)	(13.5)	(11.7)	(13.6)
Semi- and Unskilled	119,835	279,648	227,921	626,774
	(13.0)	(13.2)	(16.0)	(14.0)
Unemployed	68,033	85,199	49,221	202,453
	(7.3)	(4.0)	(3.4)	(4.5)
TOTAL	924,975	2,113,650	1,427,691	4,466,316

Table 4. Ancestry differences in occupational structures (percentages in brackets)

	Occupation					
Ancestry	PM	CS	T/ST	SU	Un	TOTAL
UK	1,085,110	970,161	415,078	425,000	123,730	3,019,079
	(35.9)	(32.1)	(13.7)	(14.1)	(4.1)	
Ireland	203,260	157,232	63,997	61,916	20,622	507,027
	(40.1)	(31.0)	(12.6)	(12.2)	(4.1)	
Greece	40,817	36,119	14,472	14,059	4,608	110,075
	(37.1)	(32.8)	(13.1)	(12.8)	(4.2)	
Italy	91,571	92,970	43,910	35,697	9,051	272,199
	(33.50	(34.0)	(16.1)	(13.1)	(3.3)	
FYugoslavia Christian	24,520	27,329	15,226	20,157	3,322	90,554
	(27.0)	(30.2)	(16.8)	(22.2)	(3.7)	
FYugoslavia Muslim	813	1,106	869	1,113	291	4,192
	(19.4)	(26.4)	(20.7)	(26.6)	(6.9)	
China	88,444	64,790	25,745	29,160	21,625	229,764
	(38.5)	(28.7)	(11.2)	(12.7)	(9.4)	
India	66,254	47,516	16,846	31,615	14,814	177,045
	(37.4)	(26.8)	(9.5)	(17.9)	(8.4)	
Lebanon Christian	12,617	11,855	6,003	4,622	2,205	37,302
	(33.8)	(31.8)	(16.1)	(12.4)	(5.9)	
Lebanon Muslim	4,271	4,682	3,506	3,435	2,185	18,079
	(23.6)	(25.6)	(19.4)	(19.0)	(12.1)	.
TOTAL	1,617,677	1,413,760	605,662	626,774	202,453	4,466,316
	(36.2)	(31.7)	(13.6)	(14.0)	(4.5)	

Key to occupations: PM – Professional and Managerial; CS – Clerical and Sales; T/ST – Technical/skilled trades; SU – Semi- and Unskilled; Un – Unemployed.

Table 5. The modelling framework – an example: first-generation immigrants, aged 20-29 who were unemployed, by ancestry (a significant difference in the modelled rate between an ancestry group and that above it in the table is shown in bold)

	O	E	MRate	LoCI	HiCI
China	10,656	2,830	3.843	3.691	4.005
Lebanon	399	132	**3.010**	**2.718**	**3.347**
India	5,991	2,957	**2.066**	**1.980**	**2.162**
Italy	142	70	1.981	1.679	2.321
Greece	49	22	1.962	1.461	2.571
Yugoslavia	532	307	1.750	1.595	1.925
UK	2,171	1,741	**1.271**	**1.200**	**1.353**
Ireland	506	523	**0.987**	**0.900**	**1.086**

O – Observed Number; E – Expected Number; MRate – Modelled Rate (O/E); LoCI – Lower Confidence Interval; HiCI – Higher Confidence Interval.

Table 6. The number of modelled rates that were either significantly greater than 1.0 (>1) or significantly smaller than 1.0 (<1), by generation (g), age group and occupation

		Professional		Clerical		Technicians		S/Unskilled		Unemployed	
G	Age	>1	<1	>1	<1	>1	<1	>1	<1	>1	<1
1	20-29	0	7	4	2	6	2	3	3	7	0
	30-49	4	4	0	6	5	2	4	2	5	3
	50-69	3	5	1	6	5	1	6	1	3	4
2	20-29	2	6	8	0	6	2	0	8	7	1
	30-49	5	0	4	2	4	3	0	6	2	5
	50-69	5	0	5	0	0	7	1	2	0	5
3	20-29	0	6	7	0	6	2	0	6	5	0
	30-49	6	1	0	0	1	2	1	4	2	3
	50-69	5	0	0	0	0	6	3	3	0	3
Sum - Generations											
1		7	16	5	14	16	5	13	7	15	7
2		12	6	17	2	10	12	1	16	9	11
3		11	7	7	0	7	10	4	13	7	6
Sum – Age Groups											
	20-29	2	19	19	2	18	6	3	17	19	1
	30-49	15	5	4	8	10	7	5	8	9	11
	50-69	13	5	6	6	5	14	10	6	3	11

Table 7. The ordering and statistically-significant differences between modelled unemployment rates by generation and age group (modelled rates significantly different from those for the ancestry group in the preceding row are shown in bold)

Age group	20-29					
Generation	1st		2nd		3rd	
	China	3.84	Lebanon	2.24	China	1.68
	Lebanon	**3.01**	India	2.15	UK	1.48
	India	**2.07**	China	**1.66**	Ireland	1.43
	Italy	1.98	UK	**1.46**	Greece	1.33
	Greece	1.96	Ireland	1.33	India	1.28
	Yugoslavia	1.75	Greece	1.31	Lebanon	1.24
	UK	1.27	Italy	**0.98**	Italy	1.17
	Ireland	**0.99**	Yugoslavia	0.83	Yugoslavia	**0.89**

Age group	30-49					
Generation	1st		2nd		3rd	
	Lebanon	2.05	India	1.62	India	1.79
	India	**1.76**	Lebanon	**1.15**	China	1.15
	China	**1.46**	China	0.94	UK	**0.86**
	Greece	1.26	UK	0.92	Greece	0.84
	Yugoslavia	1.07	Ireland	0.89	Ireland	0.83
	Italy	**0.89**	Greece	0.87	Lebanon	0.81
	UK	**0.76**	Italy	**0.62**	Italy	0.77
	Ireland	0.67	Yugoslavia	**0.51**	Yugoslavia	0.73

Age group	50-69					
Generation	1st		2nd		3rd	
	India	1.75	India	1.51	India	1.32
	Lebanon	1.59	Lebanon	0.91	China	0.91
	China	1.51	China	0.85	Greece	0.85
	Yugoslavia	**0.95**	Ireland	0.82	Ireland	0.78
	Greece	0.88	UK	0.80	Lebanon	0.75
	UK	0.86	Greece	0.76	Yugoslavia	0.70
	Ireland	0.81	Italy	**0.59**	Italy	0.69
	Italy	**0.66**	Yugoslavia	0.48	UK	0.69

Table 8. A summary of the number of significant differences between modelled rates for ancestry groups within each generation and age group, by occupational class

Generation	1st			2nd			3rd			Σ
Age group	1	2	3	1	2	3	1	2	3	
Managerial/Professional	0	4	4	1	1	0	0	1	1	12
Clerical/Sales	2	3	2	0	1	0	0	0	0	8
Technicians/skilled trades	4	4	3	1	0	0	1	0	0	13
Semi- and Unskilled	2	4	4	2	4	0	0	1	2	19
Unemployed	3	4	2	3	3	0	1	1	0	17
TOTAL	11	19	15	7	9	0	2	3	3	69

Key to age groups: 1 – 20-29; 2 – 30-49; 3 – 50-69.

Figure 1. Scatterplot of raw against full Bayesian shrunken log rates