Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 275–296, http://dx.doi.org/10.2478/JOS-2020-0015

Analysing Sensitive Data from Dynamically-Generated Overlapping Contingency Tables

Joshua J. Bon[1], *Bernard Baffour*[2], *Melanie Spallek*[3], *and Michele Haynes*[3]

Contingency tables provide a convenient format to publish summary data from confidential survey and administrative records that capture a wide range of social and economic information. By their nature, contingency tables enable aggregation of potentially sensitive data, limiting disclosure of identifying information. Furthermore, censoring or perturbation can be used to desensitise low cell counts when they arise. However, access to detailed cross-classified tables for research is often restricted by data custodians when too many censored or perturbed cells are required to preserve privacy. In this article, we describe a framework for selecting and combining log-linear models when accessible data is restricted to overlapping marginal contingency tables. The approach is demonstrated through application to housing transition data from the Australian Census Longitudinal Data set provided by the Australian Bureau of Statistics.

Key words: Count data; log-linear model; marginal model; privacy restriction.

1. Introduction

Governments, statistical agencies and data custodians are increasingly using contingency or frequency tables to make data available to the public on a wide range of topics including health, demography, education, and the economy. Tabular data can provide insights on associations among variables and are underpinned by the long-standing statistical framework of log-linear models (see Birch 1963; Bishop et al. 1975; Agresti, 1981; Cameron and Trivedi 1998; Nelder and Wedderburn 1972; Agresti 2002; Bergsma et al. 2009, among others). Importantly, contingency table data can provide statistical outputs whilst preserving the privacy and anonymity of the individuals from which the data are derived.

Making data publicly available is difficult while maintaining the legal and ethical requirements to protect individuals' privacy. Recently, new software and resources have

[1] Queensland University of Technology, School of Mathematical Sciences, GPO Box 2434, Brisbane, Queensland, 4001, Australia. Email: joshuajbon@gmail.com

[2] Australian National University, School of Demography, 9 Fellows Road, Acton, ACT 2601, Australia. Email: bernard.baffour@anu.edu.au

[3] Australian Catholic University, Institute for Learning Sciences and Teacher Education, 229 Elizabeth St, Brisbane, Queensland, 4000, Australia. Emails: Melanie.Spallek@acu.edu.au and Michele.Haynes@acu.edu.au

Acknowledgments: This research was part funded by the Australian Research Council Centre of Excellence for Children and Families over the Life Course (CE140100027). This article uses data from the 2006–2011 Australian Census Longitudinal Data set (ACLD) made available by the Australian Bureau of Statistics (ABS). The authors would like to thank the ABS for their encouragement and helpful comments on this research. The findings and views reported in this article are those of the authors and should not be attributed to the ABS. This work was also supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia.

© Statistics Sweden

enabled organisations to provide safe online access to sensitive data by generating contingency table summaries dynamically from user queries, for example TableBuilder used by the Australian Bureau of Statistics, ABS (ABS 2012). The derived tables are only released after balancing the utility and confidentiality risk (Chipperfield et al. 2016). These *dynamically-generated contingency tables* are a powerful resource for applied researchers to utilise for discovering patterns and associations in the data whilst preserving privacy. In particular, dynamically-generated tables have found favour among national statistical agencies, including the United States Census Bureau, the United Kingdom's Office for National Statistics, Statistics Netherlands, and the Australian Bureau of Statistics (Duncan et al. 2011; Chipperfield et al. 2016), and are often used to release census data.

To mitigate privacy risks, data custodians can employ a number of statistical techniques to control disclosure. For example, they may limit the number of variables allowed to be reported simultaneously, or place limits on the frequency of small cell sizes. Query restrictions, such as these, reduce disclosure risk for sensitive information, but do come at a cost to statistical analysis (Domingo-Ferrer and Mateo-Sanz 1999). Specifically, these restrictions may prohibit the user from accessing all the variables of interest in a single contingency table which is problematic for robust analysis. Practically, users can mitigate these restrictions by requesting a set of separate but overlapping contingency tables to analyse individually. In this article we focus on overlapping tables, specifically a set of contingency tables where the pairwise intersection of variables in any two tables is a common nonempty subset of the available variables. Section 2 provides an illustrative example.

To address privacy and disclosure concerns in the analysis of unit record administrative data, Lee et al. (2017) have recently proposed a modelling framework that computes sufficient statistics from separate data sources that may include subsets of "similar structure" (e.g., subsetting by natural spatial groupings such as state) from a single big database, and potentially subsets from other databases that are relevant to different levels of a hierarchical model. The sufficient statistics are computed by the data custodian, but are combined by the researcher for construction of the log-likelihood to obtain model estimates that approximate those that would be estimated from the full data set. It is further proposed that this modelling framework could be incorporated into data extraction tools provided by data custodians, such as TableBuilder. However, until this has been achieved, researchers will need to rely on analysis of aggregated data from overlapping contingency tables for many applications. In this article, we outline an approach for model estimation by combining output from separate contingency tables.

In the framework of log-linear models we show that, after an appropriate adjustment, model selection can be used to compare overlapping contingency tables, thereby computing relative importance of the explanatory variables. In addition, we re-purpose an existing technique to combine models for the overlapping tables to form the appropriate higher order model allowable by the restricted data.

The article is structured as follows. In Section 2, an illustrative example of a relevant scenario requiring access to and analysis of confidential data is explored. Section 3 provides an overview on (marginal) log-linear models, and describes the methods we use to compare and combine the models in detail. Section 4 applies the methodology to Australian housing tenure transition data from the Australian Longitudinal Census Dataset, and a summary of the method and conclusions is discussed in Section 5.

2. Illustrative Example

Scenario: A regional subset of a population census classifies each person by four sensitive categorical variables. The data set is held securely by a data custodian who has chosen to release the data online using dynamically generated contingency tables. However, the custodian has deemed that releasing the full contingency table (the *super-table*) poses a privacy risk due to the small regional population size (in reality this assessment can be done in real-time based on some measured sparsity of the table requested, see Chipperfield et al. (2016) for example). As such, they will only allow tables with up to two variables (the *marginal tables*) to be released. Under this restriction, there are two analyses that may be of interest – but currently unavailable to researchers. The first involves investigating which two variables (of the four) best explain the count data observed. Meanwhile, the second builds a model that encompasses all four marginal (overlapping) tables.

The above scenario is simplified, but in essence demonstrates the problem this article addresses. The relationship between super- and marginal tables is illustrated in Figure 1, where an inaccessible contingency table (the super-table S) is marginalised into three unique tables (the marginal tables M_1, M_2, and M_3) each with fewer variables than the super-table. In this example, the super-table has four categorical variables, C_1, C_2, C_3, and C_4. All of which can each take one of two values in this example. For simplicity, we label these values with integers 1 and 2. The marginal tables each contain two of the variables from the super-table, always the overlapping variable C_1 and one remaining variable from $\{C_2, C_3, C_4\}$. The count data are aggregated according to the marginalisation of the variables excluded in each marginal table.

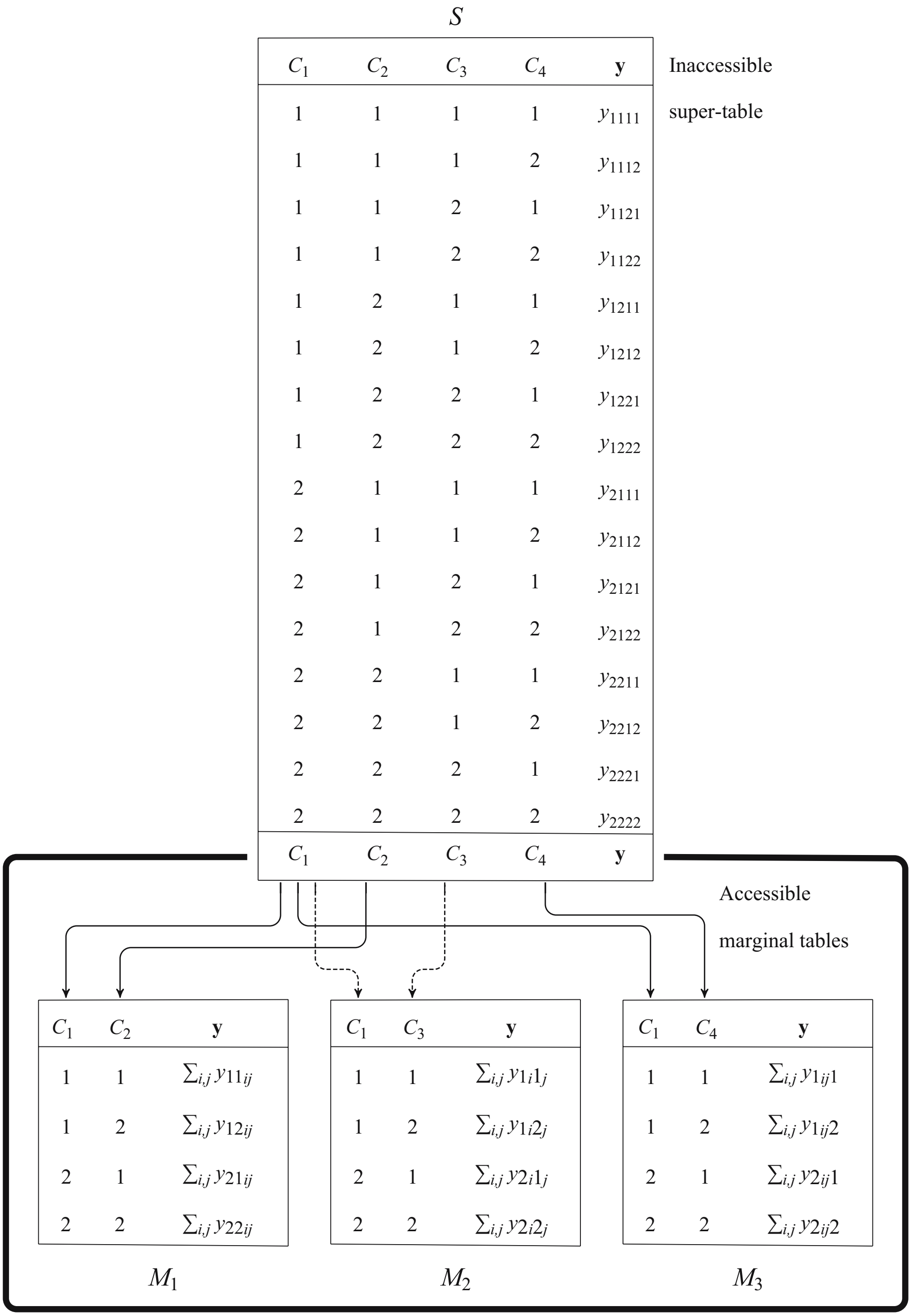
The issue here is that the log-linear models are not directly comparable when they are estimated from the marginal tables. Specifically, straight-forward comparison requires that the cell probabilities in the super-table are estimated under the constraints imposed by each marginal model (Bergsma et al. 2009). As the super-table is inaccessible, neither the cell probabilities nor the estimated model parameters can be compared without some adjustment. After addressing this first issue, we will discuss how to perform joint inference on the marginal tables.

3. Methods for Overlapping Marginal Log-Linear Models

3.1. Background

Contingency table cell counts can be used to fit log-linear models formulated under the generalised linear modelling (GLM) framework (Nelder and Wedderburn 1972) with a log link function and Poisson distributed counts. Contingency tables can also be modelled with a multinomial distribution. These are also referred to as log-linear models as the Poisson and multinomial regressions have equivalent maximum likelihood point estimates under mild assumptions (Lang 1996).

The sufficient statistics of marginal log-linear models are the maximum likelihood estimates of the expected frequencies under the corresponding marginal contingency table. This follows from Birch's theorem (Birch 1963), which implies that the maximum likelihood estimates match the marginal distributions and also ensures that the associations and interactions satisfy the model-implied patterns. In other words, there is a unique set of



C_1	C_2	C_3	C_4	**y**
1	1	1	1	y_{1111}
1	1	1	2	y_{1112}
1	1	2	1	y_{1121}
1	1	2	2	y_{1122}
1	2	1	1	y_{1211}
1	2	1	2	y_{1212}
1	2	2	1	y_{1221}
1	2	2	2	y_{1222}
2	1	1	1	y_{2111}
2	1	1	2	y_{2112}
2	1	2	1	y_{2121}
2	1	2	2	y_{2122}
2	2	1	1	y_{2211}
2	2	1	2	y_{2212}
2	2	2	1	y_{2221}
2	2	2	2	y_{2222}
C_1	C_2	C_3	C_4	**y**

C_1	C_2	**y**
1	1	$\sum_{i,j} y_{11ij}$
1	2	$\sum_{i,j} y_{12ij}$
2	1	$\sum_{i,j} y_{21ij}$
2	2	$\sum_{i,j} y_{22ij}$

C_1	C_3	**y**
1	1	$\sum_{i,j} y_{1i1j}$
1	2	$\sum_{i,j} y_{1i2j}$
2	1	$\sum_{i,j} y_{2i1j}$
2	2	$\sum_{i,j} y_{2i2j}$

C_1	C_4	**y**
1	1	$\sum_{i,j} y_{1ij1}$
1	2	$\sum_{i,j} y_{1ij2}$
2	1	$\sum_{i,j} y_{2ij1}$
2	2	$\sum_{i,j} y_{2ij2}$

Fig. 1. Illustration of accessible marginal tables nested in an inaccessible super contingency table. Each marginal table contains C_1, the overlapping variable, and one of the remaining variables from the super-table $\{C_2, C_3, C_4\}$.

fitted values that both satisfy the marginal model and match the data in the sufficient statistics, and this unique solution is the maximum likelihood estimate. Regression coefficients from log-linear regression models can be equivalently specified as associations, expected counts or cell probabilities.

Following notation from Lang (1996), a probability vector $\mathbf{p}$ containing probabilities from a contingency table can be specified by the log-linear model

$$\log(\mathrm{E}(\mathbf{p})) = \boldsymbol{\xi} = \mathbf{X}\boldsymbol{\beta} \tag{1}$$

where the cell probabilities are related to the cell counts $\boldsymbol{\mu}$ by $\mathbf{p} = n^{-1}\boldsymbol{\mu}$, and n is the sum of all counts. The vector $\boldsymbol{\xi}$ contains the expected probabilities on the log-scale, and $X\boldsymbol{\beta}$ codes the associations between cells as in generalised linear modelling with a Poisson distribution.

Marginal log-linear models have been studied extensively (see Bergsma and Rudas 2002; Bergsma et al. 2009, and references therein). These can be used to fit models where some associations (or equivalently cell probabilities) among the table cells are restricted or removed. Marginal models can be estimated with restrictions specified by Lagrangian multipliers added to the log-likelihood of the full model. Hence, the full data set is used in the fitting procedure. This is appealing since it ensures that when several marginal models are estimated, they are comparable because only the constraints have changed (the likelihood is still based on the same data set). While this study considers marginal models, the issue addressed here differs with regard to availability of the data for analysis. Specifically, the joint (or full) contingency table is not accessible, hence the constrained estimation approach to marginal models is not possible. We consider the situation where several marginal, overlapping contingency tables are accessible instead.

To clearly distinguish the general contingency tables we reference within this article, we refer to the inaccessible table which would encompass all of the marginal tables as the *super-table*. In particular, we are interested in fitting and comparing decomposable graphical log-linear models on this super-table.

Decomposable graphical models have several advantages over their non-decomposable counterparts. First, the maximum likelihood estimates can be found explicitly. Second, closed form solutions exist for the sufficient statistics. Third, a necessary condition for decomposability is that the models are hierarchical; the absence of an interaction forces all related higher-order interactions to be excluded, which aids in interpretability. Finally, an attractive feature of decomposable graphical models is that they can be interpreted in terms of their patterns of conditional independencies, which can also be displayed graphically.

A decomposition method for fitting hierarchical log-linear models with large contingency tables was proposed by Dahinden et al. (2010). In essence, associated subsets of variables are identified from a super-table, after which cell probabilities for each subset (or marginal table) are estimated using log-linear models. The results are combined using the decomposability property of graphical models (Lauritzen 1996). Sparsity can be considered using Lasso or model selection on the sub-models. In our application, with access restricted to marginal tables, it is the decomposability property that can be used to combine the results from several marginal tables. This approach is described in Subsection 3.3.

3.2. Comparing Models from Overlapping Tables

When the super-table is inaccessible, the constrained formulation of marginal models is not possible to implement. As such, the marginal models must be estimated from the different marginal data sets available, as illustrated in Section 2.

The approach addressed in this article is the converse of the situation in Allison (1980), where the results demonstrated the equivalence in estimated probabilities between collapsed and uncollapsed data sets. Their intention was to fit marginal models on contingency tables while avoiding collapsing the table itself. We, on the other hand, would like to estimate the same probabilities (or equivalently frequencies) for the full table, using only the collapsed data set.

Specifically, Allison (1980) demonstrates how the estimated cell frequencies from the full table and the collapsed table are equivalent when the frequencies from the former are also collapsed. The two frequency vectors share the same association structure (model equation) that must be collapsible for the given data. Collapsibility, as discussed in Bishop et al. (1975), is the key to ensuring these frequencies are equivalent (after adjusting for multiplicity) – it says that collapsing over one set of variables will not affect the parameters in a second set, if the two sets are independent. In our case, and in Allison (1980), the collapsed variables are not included in the model, and are therefore independent of the variables that are included.

In our analysis we fit several Poisson GLMs with overlapping explanatory categories with their respective collapsed or marginal data. We adjust the log-likelihood of each marginal model (after estimation) so that it is as if each model had been estimated using the super-table data, which contains all categorical variables used in every model. The association structure of each marginal model does not change. We must make this adjustment so that model selection techniques can be appropriately applied. As mentioned, the adjustment relies on the equivalence between probabilities from the model estimated with marginal data to the same model fitted using the super-table (Bishop et al. 1975; Allison 1980). The linear model coefficients (describing the associations) will be equal under both scenarios, except for the intercept terms which differ depending on the number of rows in each model matrix.

Below, we derive the exact adjustment needed to compare overlapping marginal models using likelihood-based metrics such as AIC (Akaike 1974). This adjustment is implicitly linked to sufficiency in marginal log-linear models and the constrained formulation of marginal log-linear models (Bergsma et al. 2009), but the authors have not been able to locate a previous derivation in the literature. Note that in the following derivation we describe two estimated vectors of probabilities that share a single association structure, the *model equation*. The first model is hypothetically estimated using the super-table as data, since this data is unavailable in our application, while the second model uses marginal data sufficient to estimate the model equation of interest.

In order to prove our result, we start by establishing a connection between two log-linear models estimating the model equation (identical design matrix, $\mathbf{X}$, and coefficient vector, $\boldsymbol{\beta}$). The models differ only in the data used to fit each of them – but both data sets are sufficient to estimate the given association structure. The first model uses the vector of counts $\mathbf{y} = \left[y_1\, y_2 \cdots y_m\right]^\top$, of length m (counts from the super-table), while the second uses a collapsed vector of counts $\mathbf{y}^\kappa = \left[y_1^\kappa\, y_2^\kappa \cdots y_{m^\kappa}^\kappa\right]^\top$, of length m^κ. This technique is then applied to the set of marginal, overlapping data sets, so that correct model comparisons can be made.

The following assumptions are required in order to equate the probabilities estimated from the marginal data to those estimated using the super-table:

1. The association structures, or model equations, to be estimated for each data set of counts ($\mathbf{y}$ and $\mathbf{y}^\kappa$) are identical.
2. The marginal data ($\mathbf{y}^\kappa$) is sufficient to estimate the model equation.
3. The counts (or cells) from the super-table, $\mathbf{y}$, have been collapsed to $\mathbf{y}^\kappa$ using $\mathbf{M}$, as described in Equation (2).
4. The variables that are collapsed are irrelevant under the given model equation.

Of the above assumptions the fourth is the strongest, although it is one that we have to make under any modelling strategy when only the marginal contingency table is available.

Let $\mathbf{M}$ be a matrix that collapses a vector of counts, $\mathbf{y}$, from the super-table to the observed counts in the marginal table, $\mathbf{y}^\kappa$. Specifically, these two vectors are related by

$$\mathbf{y}^\kappa = \mathbf{M}\mathbf{y}. \tag{2}$$

The matrix $\mathbf{M}$ is a $m^\kappa \times m$ matrix containing only zeros and ones. Every column of $\mathbf{M}$ contains only one unit element, while every row contains $r = m/m^\kappa$ (an integer) unit entries. The matrix $\mathbf{M}$ is a type of incidence matrix that sums the counts in $\mathbf{y}$ to the counts in $\mathbf{y}^\kappa$ and describes the marginalisation of the model. The following identity is useful:

$$y_i^\kappa = \sum_{j=1}^{m} M_{ij} y_j = \sum_{j:M_{ij}=1} y_j \tag{3}$$

where the first equality holds by definition, and the last equality holds since each element of M is either one or zero.

Using the illustrative scenario in Section 2 as an example, the matrix $\mathbf{M}$ that collapses the counts in super-table S to the marginal table M_1 is

$$\mathbf{M} = \begin{bmatrix} \mathbf{1}_4^\top & & & \\ & \mathbf{1}_4^\top & & \\ & & \mathbf{1}_4^\top & \\ & & & \mathbf{1}_4^\top \end{bmatrix} \text{ where } \mathbf{1}_4^\top = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$$

with zeros filling the blank cells.

The estimated cell probabilities from the model with marginal data, say $\hat{\mathbf{p}}^\kappa$, are related to the cell probabilities, $\hat{\mathbf{p}}$, in the constrained marginal formulation by

$$\hat{\mathbf{p}}^\kappa = \mathbf{M}\hat{\mathbf{p}} \tag{4}$$

under Assumptions 1–4. In other words, the estimated probabilities of the models with collapsed data and uncollapsed data (super-table) are equal, up to a multiplicative constant that accounts for the difference in the number of cells for each data set. This is only true because the association structure being fitted, as in Allison (1980), does not change across the marginal (collapsed) data and super-table (uncollapsed data).

Another property of the estimated probabilities, $\hat{\mathbf{p}}$, is that its elements repeat according to the pattern of zeros and ones in $\mathbf{M}$, as such

$$\hat{p}_s = \hat{p}_t \text{ if } M_{is} = M_{it} = 1, \text{ for some } i \in \{1, 2, \ldots, m^{\kappa}\}. \tag{5}$$

That is, cells in the super-table that share the same association structure under the marginal model (the marginal model mean equation) will have the same estimated probability (Allison 1980).

The identity for the counts in Equation (3), also holds for the probabilities, that is

$$\hat{p}_i^{\kappa} = \sum_{j:M_{ij}=1} \hat{p}_j \tag{6}$$

which can be combined with Equation (5) to give

$$\hat{p}_i^{\kappa} = r\hat{p}_j \text{ if } M_{ij} = 1, \tag{7}$$

since each $\hat{p}_j$ in the sum of Equation (6) for a given i are equal, and there are r probabilities being summed in each row of $\mathbf{M}$. To explain intuitively, as per Equation (5), using the super-table to estimate the marginal model (instead of a sufficient collapsed data set) results in dividing each probability evenly across r cells of the super-table since the same mean equation is repeated r times. It is also true that

$$\sum_{j:M_{ij}=1} 1 = r, \tag{8}$$

that is the row sums of M are equal to r, the multiplicity factor that arises from collapsing the cells and hence probabilities.

Using the estimated cell probabilities, the Poisson log-likelihood for the model with collapsed, or marginal data is

$$\log L(\hat{\mathbf{p}}^{\kappa} | \mathbf{y}^{\kappa}) = \sum_{i=1}^{m^{\kappa}} \left(y_i^{\kappa} \left(\log \hat{p}_i^{\kappa} + \log n \right) - n\hat{p}_i^{\kappa} - \log y_i^{\kappa}! \right) \tag{9}$$

for vector of observed count data, $\mathbf{y}^{\kappa}$ where $n = \sum_{i=1}^{m^{\kappa}} y_i^{\kappa}$. The log-likelihood for the model with marginal data can be rewritten using Equations (3), (7), and (8) in the following way

$$
\begin{aligned}
&\log L\left(\hat{\mathbf{p}}^\kappa|\mathbf{y}^\kappa\right)\\
&\quad = \sum_{i=1}^{m^\kappa}\left(\left(\sum_{j:M_{ij}=1} y_j\right)\left(\log\left(\hat{p}_i^\kappa\right) + \log n\right) - n\hat{p}_i^\kappa\right) + c\left(\mathbf{y}^\kappa\right)\\
&\quad = \left(\sum_{i=1}^{m^\kappa}\left(\sum_{j:M_{ij}=1} y_j\left(\log\left(\hat{p}_i^\kappa\right) + \log n\right)\right) - \frac{1}{r}\sum_{j:M_{ij}=1} n\hat{p}_i^\kappa\right) + c\left(\mathbf{y}^\kappa\right)\\
&\quad = \left(\sum_{i=1}^{m^\kappa}\sum_{j:M_{ij}=1}\left(y_j\left(\log\left(\hat{p}_i^\kappa\right) + \log n\right) - \frac{n}{r}\hat{p}_i^\kappa\right)\right) + c\left(\mathbf{y}^\kappa\right)\\
&\quad = \left(\sum_{i=1}^{m^\kappa}\sum_{j:M_{ij}=1}\left(y_j\left(\log\left(r\hat{p}_j\right) + \log n\right) - n\hat{p}_j\right)\right) + c\left(\mathbf{y}^\kappa\right)\\
&\quad = \sum_{j=1}^{m}\left(y_j\left(\log\hat{p}_j + \log r + \log n\right) - n\hat{p}_j\right) + c\left(\mathbf{y}^\kappa\right)\\
&\quad = \log L\left(\hat{\mathbf{p}}|\mathbf{y}\right) - c\left(\mathbf{y}\right) + n\log r + c\left(\mathbf{y}^\kappa\right).
\end{aligned}
$$

The integer n is the total of the counts, $n = \sum_{i=1}^{m} y_i = \sum_{j=1}^{m^\kappa} y_j^\kappa$, whose equality across holds approximately when perturbations have been added for further privacy. The constants are defined as $c\left(\mathbf{y}^\kappa\right) = -\sum_{i=1}^{m^\kappa}\log y_i^\kappa!$ and $c\left(\mathbf{y}\right) = -\sum_{j=1}^{m^\kappa}\log y_i!$. Thus an equivalence between the log-likelihood using the super-table, log $L\left(\hat{\mathbf{p}}|\mathbf{y}\right)$, and log-likelihood using the marginal data, log $L\left(\hat{\mathbf{p}}^\kappa|\mathbf{y}^\kappa\right)$, can be expressed as

$$\log L\left(\hat{\mathbf{p}}|\mathbf{y}\right) - c\left(\mathbf{y}\right) = \log L\left(\hat{\mathbf{p}}^\kappa|y^\kappa\right) - c\left(\mathbf{y}^\kappa\right) + n\left(\log m^\kappa - \log m\right). \tag{10}$$

The constant $c(\mathbf{y})$ cannot be calculated because the super-table is inaccessible. However, for quantities where a difference is of interest, such as information criteria, the $c(\mathbf{y})$ cancel out. The relative adjusted AIC for a marginal-data model with probability vector $\mathbf{p}^\kappa$ and size m^κ can be calculated as

$$\mathrm{aAIC}\left(\mathbf{p}^\kappa,\mathbf{y}^\kappa,k,m^\kappa\right) = 2k - 2\left(\log L\left(\hat{\mathbf{p}}^\kappa|\mathbf{y}^\kappa\right) - c\left(\mathbf{y}^\kappa\right) + \left(\log m^\kappa - \log m\right)\sum_{i=1}^{\kappa} y_i^\kappa\right). \tag{11}$$

The above aAIC is relative because it does not include the constant from the log-likelihood. The number of association parameters is k, and the constant m should be fixed for a given set of overlapping marginal tables. It is the product of the number of levels in the set of unique variables among all marginal tables (see Subsection 4.2).

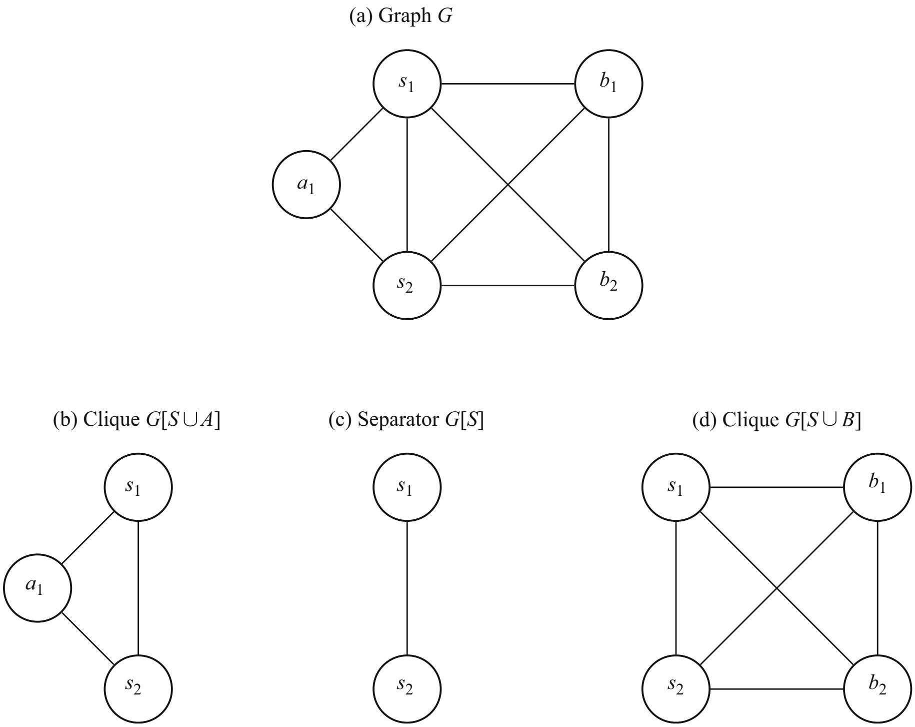
3.3. *Combining Models for Overlapping Tables*

We refer to the combination of marginal models as *stitching* and refer to the result as a *stitched* model. The stitching process takes the several overlapping log-linear models and generates the equivalent model if all parameters had been estimated jointly with a contingency table that would enable this. The method we describe has been re-purposed from Dahinden et al. (2010) who consider the case where the full data set is available.

The stitching of marginal models is possible when the set of marginal models to be combined together are a decomposition of a possible hierarchical model on the super-table. Decomposability can be described by considering the log-linear regression as a graphical model (Darroch et al. 1980) with graph $G = (V, E)$, having vertex set V, and edge set E. Define a subgraph of G induced by $W \subset V$ as $G[W] = (W, \{(u, v) \in E : u, v \in W\})$, effectively the graph remaining from G after removing all vertices absent from W (and all hanging edges). A partition of the vertex set, V, into $\{A, S, B\}$ is a decomposition if $G[S]$ separates $G[A]$ from $G[B]$, and $G[S]$ is a complete graph. A vertex subset $S \subset V$ is a (vertex) separator for A and B if its removal from G separates A and B into disconnected components. We refer to the vertex set S as the separator.

The decomposability of a graph is defined recursively; a graph is decomposable if it is complete or if there exists a decomposition $\{A, S, B\}$ such that the subgraphs $G[A \cup S]$ and $G[S \cup B]$ are decomposable. We refer the reader to Leimer (1993) for further discussion. An example decomposable graph is shown in Figure 2 and a comprehensive guide can be found in Darroch et al. (1980).

If a graph is decomposable then a relationship exists between the full graph and its complete subgraphs: the separators and cliques (vertex set W is a clique if $G[W]$ is a complete graph and, for our purposes, not a separator) (Frydenberg and Lauritzen 1989).


(a) Graph G

(b) Clique $G[S \cup A]$

(c) Separator $G[S]$

(d) Clique $G[S \cup B]$

Fig. 2. (a) Example of decomposable graph, G, with separator $S = \{s_1, s_2\}$. After the removal of S from the graph, the subgraphs with nodes $A = \{a_1\}$ and $B = \{b_1, b_2\}$ are disconnected. (b) Clique graph $G[S \cup A]$. (c) Separator graph $G[S]$. (d) Clique graph $G[S \cup B]$. Notice that $G[S \cup A]$ and $G[S \cup B]$ are complete, and hence this graph satisfies the decomposability definition. Moreover, if $G[S \cup B]$ were not complete, it would need to be decomposable to satisfy the recursive definition of decomposability.

The relationship is the special structure afforded the graph — conditioning on the separator vertices emits a conditional decomposition of the graph. In the case of statistical models where each vertex has associated parameter(s), separators act as the only intermediaries between cliques, and fixing their values results in independence between the remaining cliques (Frydenberg 1990). Let the set of cliques be $\mathcal{C}$, and set of separators be $\mathcal{S}$ and note that these sets can be constructed using the recursive definition of decomposability above.

Following Dahinden et al. (2010) we change notation slightly to accomodate the graph theory used in this section. Let $p(i)$ be the probability of belonging to particular categories of a set of variables denoted by i, from a decomposable graph (log-linear model). Let $p(i_C)$ and $p(i_S)$ denote the probability of the same categories but from the clique C and separator S respectively. In terms of log-linear modelling, $p(i)$ is the estimated probability of a particular observation from the super-table, whereas $p(i_C)$ and $p(i_S)$ are calculated from specific marginal models (derived from the overlapping tables we have access to). The relationship in logarithmic terms is

$$\log p(i) = \sum_{C \in \mathcal{C}} \log p(i_C) - \sum_{S \in \mathcal{S}} v(S) \log p(i_S) \qquad (12)$$

where $v(S)$ is the index of separator S, describing the number of times S acts as a separator (Lauritzen 1996). Using the relation in Equation (12), the estimated marginal models can be stitched or combined together. The resulting probabilities are from the equivalent joint model on the inaccessible super-table. Equation (12) accounts for the multiplicity of the separator in the estimates from the cliques in the decomposed graph. For example, the separator S appears in both cliques shown in Figure 2. In this case, the index of the separator is $v(S) = 2 - 1 = 1$ (see (Lauritzen (1996) for further details).

In the case of a log-linear model with a non-decomposable graphical counterpart, a minimal triangulation can be used in order to form a decomposable graph [see Rose et al. 1976; Olesen and Madsen 2002, for example]. Under the overlapping structure we consider, the graph generated by stitching saturated marginal models together is already decomposable, so no triangulation is needed. However, for stitching non-saturated marginal models together we suggest beginning with the triangulation equivalent to the saturated models, then using thinning (removing edges added during triangulation) to construct a minimal triangulation (Jones and Didelez 2017). In some circumstances, edge removal (removing existing model edges) may also be necessary to guarantee a model that is both decomposable and graphical – in order to ensure that the necessary sufficient statistics are available from the marginal models.

Our analysis in Section 4 stitches the saturated models from each of the marginal tables together to form a joint model across all available combinations of variables. The estimated probabilities from the resulting model can be used to calculate standard model summaries, such as estimated association coefficients, prevalence ratios, and information criteria.

4. Analysing Housing Transitions from the ACLD

Purchasing a home in Australia is a significant stage in an individual's life course and the prevalence of home ownership is an important indicator of a country's economic performance. Understanding the drivers of transitions into home ownership is therefore of considerable social and economic interest and is often the subject of life course research (see, for example Spallek et al. 2014). In Australia, a rich source of data on housing tenure transitions is the Australian Census of Population and Housing ("Census"), conducted by the ABS. We investigate home ownership transitions and their associations with demographic factors using a derivative of the 2006 and 2011 Censuses, the Australian Census Longitudinal Dataset (ACLD), (Chipperfield et al. 2017).

4.1. Data

The 2006–2011 ACLD contains information from a five-percent random sample of the Australian population selected from the 2006 Census and then subsequently linked to the 2011 Census. The final linked data set (the ACLD) consists of 800,759 records (ABS 2013). The differences between the original sample of the 2006 Census and the final linked sample are attributable to either deaths and overseas departures that occurred between the 2006 and 2011 censuses, or due to unsuccessful linkages because of inconsistent or missing information. The ACLD may be accessed with the TableBuilder software product, an online table creator that allows users to build contingency tables from ABS data without accessing unit records (Chipperfield et al. 2016; ABS 2012). TableBuilder is subject to both query restrictions and perturbations to ensure anonymity of the individuals from the underlying data.

The ideal approach to investigating home ownership transitions using the ACLD would be to create a super-table including housing tenure transitions cross-tabulated with all other variables of interest. However, due to query restrictions, requests to TableBuilder with more than 14 variables and 44 categories exceed the cell limit allowed for contingency tables. Therefore, a new strategy is needed to develop a model with the required variables. We created what we refer to as a base contingency table including housing tenure transitions between 2006 and 2011, categorised by age, and gender. Transitions of each variable, previously shown to be associated with housing tenure transitions, were added to the base contingency table to form a new, separate contingency table. This resulted in six contingency tables (CT1–CT6), where CT1-base contains age, gender and housing tenure transitions; CT2-children contains all CT1-base variables and children status transition; CT3-family contains all CT1-base variables and family status transition; CT4-labour contains all CT1-base variables and labour transition; CT5-marital contains all CT1-base variables and marital transition; and CT6-geography contains all CT1-base variables and geographical transition. To assess which of these variables was most strongly associated with housing tenure transitions, we applied a set of log-linear models to each of the contingency tables CT1–CT6 and used the AIC to select the best model (BM1–BM6) for each contingency table (CT1–CT6). The set of log-linear models applied to each of CT1–CT6 ranged from models with single main effects to saturated models which includes all interactions. After adjusting the AIC for each best model (BM1–BM6), as discussed in Section 3, we compared BM1–BM6 to identify which of the demographic variables has the strongest association with housing tenure transitions in conjunction with age group and sex.

The analytical sample in this example is restricted to non-Indigenous Australians aged between 20 and 60 years old who did not own a house outright in 2006. We do not consider those aged over 60 years old, because individuals in this age group experience transitions in home ownership related to different events, for example retirement. Our final sample consists of 260,595 individuals from the ACLD who have data linked between the two census time points.

Each of the six contingency tables (CT1–CT6) contain variables age, sex, housing tenure transition, and one additional transition variable. Housing tenure transitions are distinguished between renting (or other) and owning with a mortgage, coded as 1 and 2 respectively. Individuals owning a home outright are excluded from this analysis. For illustrative analysis we consider the core variables of interest to be sex (coded as 1 = 'Male'; 2 = 'Female') and age (coded as 1 = '21 to 30 years old'; 2 = '31 to 40 years old'; and 3 = '41 to 60 years old'), While the additional transition variables are; children status (coded as 1 = 'No children (0–4 years)'; 2 = 'One child (0–4 years)'; 3 = 'Two or more children (0–4 years)'; and 4 = 'Not applicable'), family status (coded as 1 = 'Couple family, no children'; 2 = 'Couple family, children under 15'; 3 = 'Couple family, no children under 15'; 4 = 'One parent family, children under 15'; 5 = 'One parent family, no children under 15'; 6 = 'Other family, or not applicable'), labour status (coded as 1 = 'Employed'; 2 = 'Unemployed'; 3 = 'Not in the labour force or other'), marital status (coded as 1 = 'Married'; 2 = 'Never married'; and 3 = 'Separated, divorced, or widowed'), and geography status (coded as 1 = 'Australia Major Cities'; 2 = 'Australia Regional'; and 3 = 'Remote or other'). Table 1 summarises the demographic variables of the analytical sample from the ACLD.

The analytical sample includes slightly more females than males (52.2% versus 47.8%), with the majority being aged between 41–60 years (40.7%). In 2006, 36.6% of individuals were renting, which decreased to 31.5% in 2011. The percentage of people owning their own home with a mortgage increased, as would be expected with the same ageing population, from 63.4% to 68.5% during these five years to 2011. In 2006 the majority of people had no children between the age 0–4 years (41.1%), defined their family status as being a couple family with children less than 15 years old (40.0%), were employed (80.0%), married (55.9%) and lived in an Australian major city (72.3%). With regard to these variables and related categories, there were no major differences in their distributions between 2006 and 2011 with the exception of the percentage of married individuals, which increased by 5.5%. While the distribution of individuals across the categories of most variables was stable, this masks the changes at the individual level.

4.2. Empirical Validation

Before conducting the main analysis, we validate the comparison method discussed in Subsection 3.2 by fitting the identical models on the six contingency tables (CT1–CT6) generated by TableBuilder and introduced in Subsection 4.1. The "validating" log-linear model regresses the counts from each table with combinations of the categories in Table 1 from sex (S), age (A), and housing tenure transition ($T_{2006} \times T_{2011}$). The notation $T_{2006} \times T_{2011}$ represents the interactions of the categories in housing tenure in 2006 and 2011, namely, rent to rent, rent to own, own to rent, own to own (ownership is always with a mortgage). In general, we will suppress multiplication symbol in $C_1 \times C_2$ and simply write the $C_1\,C_2$. The validating model can therefore be represented by the notation $S\,A\,T_{2006}\,T_{2011}$.

Table 1. Categories of variables taken from Australian census longitudinal data set with aggregated counts and percentages.

				Counts (percent)			
Variable	Variable name	Value	Value name	2006	(%)	2011	(%)
S	Sex	1	Male	124,622	(47.8)		
		2	Female	135,973	(52.2)		
A	Age bracket	1	21 to 30 years old	69,312	(26.6)		
		2	31 to 40 years old	85,237	(32.7)		
		3	41 to 60 years old	106,046	(40.7)		
T	Housing tenure status	1	Rent (or other)	95,435	(36.6)	82,205	(31.5)
		2	Owned (mortgage)	165,160	(63.4)	178,390	(68.5)
C	Children status	1	No children (0–4 years)	107,175	(41.1)	115,056	(44.2)
		2	One child (0–4 years)	39,469	(15.1)	36,039	(13.8)
		3	Two or more children (0–4 years)	16,587	(6.4)	16,363	(6.3)
		4	Not applicable	97,323	(37.3)	93,096	(35.7)
F	Family status	1	Couple family, no children	53,473	(20.5)	53,026	(20.3)
		2	Couple family, children under 15	104,238	(40.0)	103,209	(39.6)
		3	Couple family, no children under 15	32,218	(12.4)	37,941	(14.6)
		4	One parent family, children under 15	13,715	(5.3)	12,570	(4.8)
		5	One parent family, no children under 15	10,116	(3.9)	11,690	(4.5)
		6	Other family, or not applicable	46,780	(18.0)	42,104	(16.2)
L	Labour status	1	Employed	208,491	(80.0)	210,268	(80.7)
		2	Unemployed	8,630	(3.3)	7,529	(2.9)
		3	Not in the labour force, or other	43,466	(16.7)	42,790	(16.4)
M	Marital status	1	Married	145,543	(55.9)	159,959	(61.4)
		2	Never married	80,796	(31.0)	61,016	(23.4)
		3	Separated, divorced, or widowed	34,197	(13.1)	39,561	(15.2)
G	Geography status	1	Australia Major Cities	188,400	(72.3)	189,022	(72.5)
		2	Australia Regional	68,204	(26.2)	67,799	(26.0)
		3	Remote or other	3,964	(1.5)	3,747	(1.4)

Table 2. Adjusted AIC of marginal contingency tables on validating model: $S\ A\ T_{2006}\ T_{2011}$.

Table	unadjusted AIC	n. cells	Relative adjusted AIC	
			aAIC − min(aAIC)	$\frac{\text{aAIC}}{\text{min(aAIC)}} - 1$ %
CT1-Base	308	24	428	0.019
CT2-Children	450,743	384	140	0.006
CT3-Family	632,602	864	16	0.001
CT4-Labour	630,264	216	327	0.014
CT5-Marital	525,031	216	0	0.000
CT6-Geography	682,548	216	263	0.012

To use the adjusted log-likelihood (and hence the adjusted AIC) in Equation (10) we need to calculate the size of the super-table, m, which would have CT1–CT6 as marginal models. This can be calculated by taking all the unique variables in CT1–CT6, then finding the product of the number of levels for each variable. Alternatively, since CT1 is nested in each marginal model m can be calculated by $m = m_1^\kappa \prod_{i=2}^{6} \left(m_i^\kappa / m_1^\kappa\right)$ where m_i^κ is the number of cells in CTi.

Table 2 contains the unadjusted and adjusted AIC values for each table using the model $S\ A\ T_{2006}\ T_{2011}$. The unadjusted AIC values demonstrate that although each contingency table is derived from the same underlying information, and in aggregate will be almost identical, the AIC values from an identical model still differ. There are two reasons why the AIC values are different. Firstly, the number of cells differ across contingency tables (see "n. cells" in Table 2) and secondly, some counts have been perturbed. Since each contingency table is a non-identical data set the justification for the AIC no longer holds. As discussed in Section 3, the relative adjusted AICs can be used for model comparison instead. The relative adjusted AIC is shown as a value and as a proportion in columns 3 and 4 of Table 2. The relative value is given because the AIC from the super-table can only be evaluated up to a constant, as noted in Subsection 3.2. The proportional value of the adjusted AIC is shown to demonstrate the magnitude of the differences of the adjusted AIC values. The adjusted AIC results show that there is a small discrepancy in the adjusted AIC of less than 0.02%. This error is due to the perturbations that differ for every unique data set retrieved from TableBuilder (Chipperfield et al. 2016).

4.3. Comparing Marginal Models from the ACLD

To identify the best model (BM1–BM6) for each of the contingency tables CT1–CT6, we performed step-wise selection using the unadjusted AIC. The analysis was conducted in the statistical language R (R Core Team 2016). Following this, the adjusted AIC is calculated for each of the best models (BM1–BM6) so that a comparison is possible across BM1–BM6. Table 3 shows these relative adjusted AIC along with the total number of observations (m^κ), the number of coefficients in the model (k) and remaining degrees of freedom (df). Table 3 is ordered by the relative adjusted AIC. The model including the geographical transition was selected as the best model overall (of BM1–BM6), followed by the models including labour status and family status transitions. Unsurprisingly, the

Table 3. Adjusted AIC comparison of the best models for each marginal contingency table (CT1–CT6).

				Relative adjusted AIC	
Best model	m^{κ}	k	df	aAIC − min(aAIC)	Ranking
BM6-Geography	216	162	54	0	1
BM4-Labour	216	200	16	52,502	2
BM3-Family	864	684	180	54,644	3
BM5-Marital	216	178	38	157,406	4
BM2-Children	384	324	60	233,260	5
BM1-Base	24	22	2	680,874	6

base model which contains only age, sex and housing tenure transition yielded the highest relative adjusted AIC and hence was the worst performing model by the AIC.

Table 4 details the interactions that are present in the best model (BM1–BM6) for each of the six contingency tables (CT1–CT6). Each of BM1–BM6 contains the main effects and all 2nd degree interactions, but inclusion of higher degree interactions differed across models. For example BM1-base includes all 3rd but no 4th degree interactions, namely the model represented by $A\ T_{2006}\ T_{2011} + S\ T_{2006}\ T_{2011} + S\ A\ T_{2011} + S\ A\ T_{2006}$.

Table 4. Best models (BM1–BM6) for each contingency table (CT1–CT6) by step-wise AIC. All BMs have main effects and 2nd degree interactions.

Best model	3rd degree interactions	4th degree interactions	5th degree interactions
BM1-Base	All	None	–
BM2-Children	All	All, excluding: $S\ A\ T_{2011}\ C_{2006}$, $S\ A\ T_{2011}\ C_{2011}$	$S\ T_{2006}\ T_{2011}\ C_{2006}\ C_{2011}$, $S\ A\ T_{2006}\ C_{2006}\ C_{2011}$, $A\ T_{2006}\ T_{2011}\ C_{2006}\ C_{2011}$
BM3-Family	All	All, excluding: $S\ A\ T_{2011}\ F_{2006}$	$S\ T_{2006}\ T_{2011}\ F_{2006}\ F_{2011}$, $A\ T_{2006}\ T_{2011}\ F_{2006}\ F_{2011}$
BM4-Labour	All	All	$S\ A\ T_{2006}\ L_{2006}\ L_{2011}$, $S\ A\ T_{2011}\ L_{2006}\ L_{2011}$, $S\ T_{2006}\ T_{2011}\ L_{2006}\ L_{2011}$, $A\ T_{2006}\ T_{2011}\ L_{2006}\ L_{2011}$
BM5-Marital	All	All, excluding: $S\ A\ T_{2011}\ M_{2006}$ $S\ A\ T_{2006}\ T_{2011}$, $S\ T_{2006}\ T_{2011}\ M_{2011}$	$S\ A\ T_{2006}\ M_{2006}\ M_{2011}$, $A\ T_{2006}\ T_{2011}\ M_{2006}\ M_{2011}$
BM6-Geography	All, excluding: $S\ T_{2006}\ G_{2006}$, $S\ T_{2011}\ G_{2006}$	All, excluding: $S\ A\ T_{2006}\ G_{2006}$, $S\ A\ T_{2011}\ G_{2006}$, $S\ T_{2006}\ T_{2011}\ G_{2006}$, $S\ T_{2006}\ G_{2006}\ G_{2011}$, $S\ T_{2011}\ G_{2006}\ G_{2011}$	$S\ A\ T_{2006}\ T_{2011}\ G_{2011}$, $A\ T_{2006}\ T_{2011}\ G_{2006}\ G_{2011}$

BM6-geography was found to be the best model including two of a possible six 5th degree interactions, two-thirds of the possible 4th degree interactions, and almost all 3rd degree interactions as specified in Table 4. Note that all models BM1-BM6 except the base model (BM1-base) included a 5th degree interaction containing age, housing tenure transition, and the transition of their additional variable. This indicates that there is an important association with tenure transition across all the additional variables investigated.

Selected prevalence ratios from the BM6-geography are described in Table 5 to demonstrate some inferences that can be draw from this model. The prevalence ratios indicate that individuals remaining in the city (i.e., city to city transition) were most associated with transitioning from renting in 2006 to owning (with a mortgage) in 2011, across all age groups and sexes. The second strongest association was observed for individuals remaining in regional locations, which also persisted across all age groups and sexes. Females, aged 21 to 30 in 2006, and remaining in the city had the highest mean association with transition into home ownership. The prevalence ratio indicated that compared to females, aged 21 to 30, who remain in remote areas, females in the same age-group who remain in the city were approximately 108.54 times more likely to move from renting to owning. Comparing males to females in each age group and geographical transition, shows that the confidence intervals for the prevalence ratios overlap. This is to be expected since the two 5th degree interactions relevant to the best geography model (see Table 4) do not include an interaction between sex and the geographical locations in both years. Some of these results are not supported by the analysis once we combine the models in Subsection 4.4.

4.4. *Combining the ACLD Marginal Models*

Table 6 contains a selection of results from the stitched model; the model combined from five marginal models using the decomposability property discussed in Subsection 3.3. It shows the same values as Table 5, but for the stitched model which has the (factored) model equation

$$SAT_{2006}T_{2011}[C_{2006}C_{2011}+F_{2006}F_{2011}+L_{2006}L_{2011}+M_{2006}M_{2011}+G_{2006}G_{2011}]. \quad (13)$$

The stitched model has 1,800 parameters, which is small compared to the 10,077,696 cells of the contingency table that would usually be needed (inaccessible super-table). To calculate the 95% confidence intervals of the prevalence ratios we used a stratified (Pearson) residual bootstrap. In each iteration a new data set was randomly generated from sampling the residuals, and the model estimated using the stitching method. Some confidence intervals were unstable, and were omitted from the table. Bootstrapping techniques for stitched log-linear models with variable cell counts will be developed further in future research.

The mean prevalence ratios of the stitched model (Table 6) are very similar to the best geography model (Table 5) with baseline geography transition remote-to-remote. However, the 95% confidence intervals (that were stable) are generally wider than those from the best geography model. The unstable confidence intervals indicate little information is actually available for that estimate. The difference arises since the confidence intervals in the best geography model are calculated from likelihood profiling

*Table 5. Selected estimated prevalence ratio comparisons and cell counts from the BM6-Geography model. *Indicates baseline of comparison.*

	Prevalence ratios (95% CI) n = cell count					
	Rent to Own					
	21 to 30		31 to 40		41 to 60	
	Male	Female	Male	Female	Male	Female
City → City	87.48 (24.88, 306.00) $n = 4886$	108.54 (30.32, 387.08) $n = 5545$	68.31 (14.24, 320.29) $n = 3988$	65.70 (13.76, 306.24) $n = 4114$	44.51 (10.30, 188.90) $n = 2772$	47.64 (11.27, 198.60) $n = 2608$
City → Region	3.95 (1.14, 13.68) $n = 225$	4.78 (1.37, 16.59) $n = 240$	2.70 (0.43, 16.49) $n = 156$	2.21 (0.34, 13.80) $n = 140$	2.18 (0.50, 9.24) $n = 126$	2.70 (0.62, 11.53) $n = 158$
City → Remote	0.65 (0.46, 0.92) $n = 34$	0.71 (0.50, 1.01) $n = 39$	0.38 (0.26, 0.56) $n = 21$	0.36 (0.24, 0.53) $n = 24$	0.36 (0.25, 0.53) $n = 22$	0.52 (0.35, 0.74) $n = 29$
Region → City	5.50 (1.61, 18.72) $n = 333$	8.25 (2.42, 28.14) $n = 395$	4.73 (1.13, 19.70) $n = 266$	4.43 (1.09, 17.92) $n = 287$	3.35 (0.85, 13.00) $n = 191$	3.44 (0.86, 13.64) $n = 206$
Region → Region	22.16 (6.67, 73.52) $n = 1242$	29.28 (8.96, 95.81) $n = 1491$	19.61 (3.58, 106.16) $n = 1146$	18.72 (3.43, 101.47) $n = 1171$	16.45 (4.29, 62.53) $n = 1026$	18.00 (4.49, 71.55) $n = 984$
Region → Remote	0.75 $n = 41$ (0.54, 1.03)	1.04 $n = 54$ (0.76, 1.43)	0.71 $n = 37$ (0.52, 0.97)	0.98 $n = 66$ (0.73, 1.31)	0.62 $n = 37$ (0.45, 0.86)	0.57 $n = 33$ (0.41, 0.80)
Remote → City	0.33 (0.21, 0.50) $n = 20$	0.32 (0.21, 0.49) $n = 15$	0.16 (0.09, 0.26) $n = 11$	0.14 (0.08, 0.24) $n = 7$	0.19 (0.11, 0.30) $n = 14$	0.22 (0.13, 0.36) $n = 10$
Remote → Region	0.33 (0.22, 0.50) $n = 18$	0.42 (0.28, 0.63) $n = 22$	0.11 (0.05, 0.20) $n = 4$	0.06 (0.03, 0.11) $n = 6$	0.22 (0.13, 0.35) $n = 16$	0.17 (0.10, 0.28) $n = 7$
Remote → Remote	1* $n = 60$	1* $n = 47$	1* $n = 64$	1* $n = 57$	1* $n = 64$	1* $n = 53$

Table 6. Selected estimated prevalence ratio comparisons from combined log-linear model (stitched) related to geography component of the model.

	Prevalence ratios (95% CI)					
			Rent to own			
	21 to 30		31 to 40		41 to 60	
	Male	Female	Male	Female	Male	Female
City → City	81.43†	117.98 (0.00, 2339.23)	62.31†	72.18 (22.76, 671.63)	43.31†	49.21†
City → Region	3.75†	5.11 (0.92, 55.05)	2.44†	2.46 (0.65, 23.00)	1.97†	2.98†
City → Remote	0.57†	0.83 (0.00, 10.20)	0.33†	0.42 (0.13, 3.75)	0.34†	0.55†
Region → City	5.55†	8.4 (2.24, 94.31)	4.16†	5.04 (1.87,44.90)	2.98†	3.89†
Region → Region	20.7†	31.72 (6.11,383.34)	17.91†	20.54 (1.38, 268.53)	16.03†	18.57†
Region → Remote	0.68†	1.15 (0.11, 13.24)	0.58†	1.16 (0.13, 12.02)	0.58†	0.62†
Remote → City	0.33†	0.32 (0.08, 3.00)	0.17†	0.12 (0.01, 1.07)	0.22†	0.19†
Remote → Region	0.30†	0.47 (0.15, 4.00)	0.06†	0.11 (0.00, 0.94)	0.25†	0.13†
Remote → Remote	1*	1*	1*	1*	1*	1*

*Indicates baseline of comparison.
†Indicates that confidence intervals were unstable.

rather than bootstrapping, and come from a model which is more likely to overfit the data given that the relative number of parameters versus observations is high (162 vs 216), even after AIC model selection. Even if bootstrapping were undertaken on the best geography model, the low degrees of freedom will dictate smaller residual sizes and hence smaller confidence intervals. As such, the stitched model with bootstrapping is a more robust analysis given the data available to us. It shows that there is actually inconclusive evidence for many of the categories we made inference about in Subsection 4.3.

There are several inferences from Table 6 that can be made. Females in the 31–40 age bracket have high odds of becoming homeowners, especially those staying in the city, staying in regional areas, or moving from regional to city areas. Females in the 21–30 age group that are likely to become homeowners are those staying in regional areas and moving from regional to city areas. The mean prevalence ratio for females aged 21–30 staying in the city was the highest estimate in Table 6 but had a large confidence interval.

5. Conclusion

Decomposition and combination of large log-linear models has been used in work by Dahinden et al. (2010). We adapt this approach to the scenario where contingency table output is restricted from table builders to a set of overlapping marginal tables. We also discuss how to compare these separate marginal models appropriately, but find that in our example the stitched model is more robust.

Table 6 is one of many outputs that can be derived from the stitched model in our example using housing tenure transitions. There are the other variables, and other housing transition categories to consider. Different baselines can also be chosen, which emphasises certain comparisons. We have presented Table 6 since it best relates to our research question about how Australians move from renting to owning. Defining a research question is very important in this analysis (as always) because it determined which tables to request from TableBuilder, which results to extract from our stitched model, and how to display these results.

This article contributes to the toolbox of applied statisticians and researchers to make better use of tabular data where access is subject to query restrictions. It is particularly useful to national statistical agencies (and users of their data sets) who are required to preserve privacy and implement disclosure controls. Future research may address whether similar methods can be used for over- or under-dispersed count data.

6. References

ABS. 2012. TableBuilder user manual. Technical report, Australia Bureau of Statistics, Canberra, ACT (cat.no 2065.0). Available at: http://www.abs.gov.au/tablebuilder (accessed October 2016).

ABS. 2013. *Australian Census Longitudinal Dataset: Methodology and quality assessment – 2080.5 – 2006-11*. Technical report, Australia Bureau of Statistics, Canberra, ACT. Available at: https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/2080.5Main+Features12006-2016 (accessed October 2016).

Agresti, A. 1981. "Measures of nominal-ordinal association." *Journal of the American Statistical Association* 76(375): 524–529. DOI: https://doi.org/10.1080/01621459.1981.10477679.

Agresti, A. 2002. *Categorical Data Analysis*. Springer, second edition.

Akaike, H. 1974. "A new look at the statistical model identification." *IEEE Transactions on Automatic Control* 19(6): 716–723. DOI: https://doi.org/10.1109/TAC.1974.1100705.

Allison, P.D. 1980. "Analyzing collapsed contingency tables without actually collapsing." *American Sociological Review* 45(1): 123–130. DOI: https://doi.org/10.2307/2095247.

Bergsma, W., M.A. Croon, and J.A. Hagenaars. 2009. *Marginal models: For dependent, clustered, and longitudinal categorical data*. Springer Science & Business Media.

Bergsma, W.P. and T. Rudas. 2002. "Marginal models for categorical data." *Annals of Statistics* 30(1): 140–159. DOI: https://doi.org/10.1214/aos/1015362188.

Birch, M. 1963. "Maximum likelihood in three-way contingency tables." *Journal of the Royal Statistical Society. Series B (Methodological)* 25: 220–233. Available at: https://www.jstor.org/stable/2984562 (accessed November 2017).

Bishop, Y., S. Fienberg, and P. Holland. 1975. *Discrete multivariate analysis: Theory and practice*. The MIT Press, Cambridge, Massachusetts.

Cameron, A. and P. Trivedi. 1998. *Regression analysis of count data*. Cambridge University Press.

Chipperfield, J., D. Gow, and B. Loong. 2016. "The Australian Bureau of Statistics and releasing frequency tables via a remote server." *Statistical Journal of the IAOS* 32(1): 53–64. DOI: https://doi.org/10.3233/SJI-160969.

Chipperfield, J., J. Brown, and N. Watson. 2017. “The Australian Census Longitudinal Dataset: using record linkage to create a longitudinal sample from a series of cross-sections.” *Australian and New Zealand Journal of Statistics* 59(1): 1–16. DOI: https://doi.org/10.1111/anzs.12177.

Dahinden, C., M. Kalisch, and P. Bühlmann. 2010. “Decomposition and model selection for large contingency tables.” *Biometrical Journal* 52(2): 233–252. DOI: https://doi.org/10.1002/bimj.200900083.

Darroch, J.N., S.L. Lauritzen, and T.P. Speed. 1980. “Markov fields and log-linear interaction models for contingency tables.” *The Annals of Statistics* 8(3): 522–539. DOI: https://doi.org/10.1214/aos/1176345006.

Domingo-Ferrer, J. and J. Mateo-Sanz. 1999. “Resampling for statistical confidentiality in contingency tables.” *Computers & Mathematics with Applications* 38(11–12): 13–32. DOI: https://doi.org/10.1016/S0898-1221(99)00281-3.

Duncan, G., M. Elliot, and J.-J. Salazar-González. 2011. *Statistical Confidentiality: Principles and Practice*. Statistics for Social and Behavioral Sciences. Springer, New York, NY, second edition.

Frydenberg, M. 1990. “Marginalization and collapsibility in graphical interaction models.” *The Annals of Statistics* 8(2): 790–805. DOI: https://doi.org/10.1214/aos/1176347626.

Frydenberg, M. and S.L. Lauritzen. 1989. Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika* 76(3): 539–555. DOI: https://doi.org/10.2307/2336119.

Jones, E. and V. Didelez. 2017. “Thinning a triangulation of a Bayesian network or undirected graph to create a minimal triangulation.” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 25(3): 349–366. DOI: https://doi.org/10.1142/S0218488517500143.

Lang, J.B. 1996. “On the comparison of multinomial and Poisson log-linear models.” *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1): 253–266. Available at: https://www.jstor.org/stable/2346177 (accessed October 2017).

Lauritzen, S.L. 1996. *Graphical models*, volume 17. Clarendon Press.

Lee, J.Y., J.J. Brown, and L.M. Ryan. 2017. “Sufficiency revisited: Rethinking statistical algorithms in the big data era.” *The American Statistician* 71(3): 202–208. DOI: https://doi.org/10.1080/00031305.2016.1255659.

Leimer, H.-G. 1993. “Optimal decomposition by clique separators.” *Discrete mathematics* 113(1–3): 99–123. DOI: https://doi.org/10.1016/0012-365X(93)90510-Z.

Nelder, J. and R. Wedderburn. 1972. “Generalized linear models.” *Journal of the Royal Statistical Society. Series A (General)* 135(3): 370–384. DOI: https://doi.org/10.2307/2344614.

Olesen, K.G. and A.L. Madsen. 2002. “Maximal prime subgraph decomposition of Bayesian networks.” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 32(1): 21–31. DOI: https://doi.org/10.1109/3477.979956.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 2016. Available at: https://www.R-project.org/ (accessed November 2018).

Rose, D.J., R.E. Tarjan, and G.S. Lueker. 1976. “Algorithmic aspects of vertex elimination on graphs.” *SIAM Journal on computing* 5(2): 266–283. DOI: https://doi.org/10.1137/0205021.

Spallek, M., M. Haynes, and A. Jones. 2014. “Holistic housing pathways for Australian families through the childbearing years.” *Longitudinal and Life Course Studies* 5(2): 205–226. DOI: https://doi.org/10.14301/llcs.v5i2.276.

Received June 2019
Revised October 2019
Accepted February 2020

© 2020. This work is published under http://creativecommons.org/licenses/by-nc-nd/3.0 (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.