Bull. Aust. Math. Soc. **80** (2009), 173–176
doi:10.1017/S0004972709000483

MODELLING REGIONAL MIGRATION

ANGELA PEZIC

(Received 18 March 2009)

2000 *Mathematics subject classification*: primary 91D25; secondary 62-07, 91B72, 91D20.

Keywords and phrases: internal migration, regional Victoria, two-stage migration model, model specification, zero-inflated Poisson.

In many parts of regional Australia, internal migration is an important driver of population change. While international migration is a principal factor in population growth in major coastal cities, regional Australia relies more heavily on natural growth and internal migration for its demographic future.

Being a selective process, migration can have important effects on the age structure, sex ratios, literacy rates and other demographic factors. Migration has a substantial influence on the social and economic structure of an area. Thus, understanding the patterns associated with internal migration is important for regional Australia. Herein lies the motivation for this study. Policy makers, especially in local government in regional parts of Australia, should appreciate the complexity of internal migration patterns and the underlying factors.

Earlier quantitative models for internal migration in Australia have tended to have a national focus; these contributed to a broad view of internal migration and therefore have limited use for policy makers at a regional or local level. Thus, there is a need for models that describe internal migration from a regional perspective.

The aim of this thesis is to build regression models that describe internal migration patterns for two cities in Victoria – Bendigo and Warrnambool. Bendigo is a large city in central Victoria, and Warrnambool is a smaller city on the coast. The principal source of data is a special tabulation obtained from the Australian Census conducted in 2001. Internal migration was analysed between local government areas for the period from 1996 to 2001.

The task of developing such models poses statistical problems. How do we define geographic regions that are appropriate for the model and the sources of data? How

Thesis submitted to La Trobe University, July 2008. Degree approved, November 2008. Supervisors: Dr Graeme Byrne and Professor Terry Mills.

do we deal with very small migration flows (even zero flows) between regions? How should the model be specified in light of the previous questions? An overview of how this was done is given.

The literature review examines the many factors that influence internal migration, and past migration models. Some of the factors include age, sex, ethnicity, education, income, labour force status, house prices and distance. The mathematical bases for a range of models that are relevant to this study are considered, such as ordinary least squares regression, spatially autoregressive models, Poisson regression and similar models. Assumptions and limitations of the models are discussed with a view to choosing models which may be suitable for modelling internal migration in regional Victoria.

A descriptive analysis of regional migration patterns in Australia is presented from several perspectives. First, interstate migration is considered, with the majority of internal migration occurring between the three most populous eastern states of Queensland, New South Wales and Victoria. Then migration between statistical divisions, by age and labour force status, is examined. Migration within Victoria by age and sex is also examined, and a descriptive analysis of migration to and from Bendigo and Warrnambool is presented.

Some issues that have arisen in migration modelling are model specification and the roles of age, sex and spatial location. The issues are assessed to determine the need for separate migration models by age, sex and metropolitan area.

It is often argued that the decision to migrate is made in two stages: first, one decides to move; then, one chooses a destination [6]. The approach taken in this thesis reflects this point of view and is based on the production constrained gravity model put forward by Wilson [8], which can also be derived as a special case of the Alonso framework [1, 2].

To model the initial decision to move (Stage 1), age-specific regression models relating out-migration rates to determinants drawn from the demographic, socio-economic, labour market, housing sector, environmental and public policy categories were developed for Victorian local government areas (LGAs). This 'separate regressions' approach was used to avoid the intrinsic collinearity problems often associated with observational studies involving both qualitative and quantitative factors [5]. Before formulating and fitting these models, a general linear model was used to assess the extent of interaction between qualitative factors such as age and sex and the covariates. This resulted in considerable simplification of the initial model involving all main and interaction effects, implying maximal age specific models of the form

$$y_i = \mu + \beta_s S + \beta_c C + \sum_{k=1}^{m} \beta_k x_{ik} + \sum_{k=1}^{m} \beta_{ck} C x_{ik} + \varepsilon_i \tag{1}$$

where y_i is the response variable, μ is the intercept, S is a dummy variable for sex, and C is a dummy variable indicating whether the LGA is within a metropolitan area or not, x_{ik} are the covariates, and m is the number of covariates.

https://doi.org/10.1017/S0004972709000483 Published online by Cambridge University Press

Variable selection was carried out using bootstrap assisted stepwise regression [3], with subsequent calibration being based on those variables with consistent effect direction selected in at least 60% of 1000 bootstrap samples. This approach yielded models of moderate size (8 to 12 explanatory variables) and quite high explanatory power ($R^2 > 70\%$). Residual analysis indicated no significant residual spatial correlation or problems with other standard regression assumptions, and multicollinearity was found not to be an issue.

The starting point for a destination selection model (Stage 2) is the Poisson or, to allow for possible overdispersion, negative binomial distribution. In either case the number of out-migrants from origin i to destination j is connected to the predictors via the mean function

$$\ln \lambda_{ij} = \mu_i + \beta_{is} S + \beta_{ic} C + \sum_{k=1}^{m} \beta_{ik} x_{jk} + \sum_{k=1}^{m} \beta_{ick} C x_{jk} + \gamma \ln d_{ij}. \tag{2}$$

The above age and origin specific model is similar to the Stage 1 model (1), except the predictors now refer to destination properties and the distance term is included.

Variable selection for the Stage 2 models was carried out using stepwise regression with bootstrap validation of each effect for both Poisson and negative binomial models. The resulting variable sets were then used to fit zero-inflated and hurdle versions [4, Ch. 4] of each. The six resulting models were then compared both graphically using post-fit cumulative distribution function plots and using the Vuong test [7]. In all cases the zero-inflated Poisson model provided superior fit, and neither the pure Poisson nor the negative binomial models could adequately cope with the large excess of zero migrant flows in the data. Graphical analysis of residuals for each age group revealed no obvious violations of the standard assumptions.

Although the Stage 1 and Stage 2 models were specified and calibrated without specific reference to each other, they are linked when it comes to forecasting via the framework on which they are based. In a forecasting setting the Stage 1 model is used to predict the number of out-migrants for a given origin which, in turn is used as input to the Stage 2 model to predict the allocation of migrants to each destination.

This thesis is a contribution to the study of models of internal migration in the context of regional Victoria. Furthermore, it addresses a number of statistical problems related to variable selection, model specification and verification of basic assumptions.

References

[1] W. Alonso, *National Interregional Demographic Accounts: A Prototype*, Monograph 017 (Institute of Urban and Regional Development, University of California, Berkeley, 1973) [online]. Available: http://iurd.berkeley.edu/PDF/MG17.pdf [Accessed 19 June 2008].

[2] ———, 'A theory of movement, in: *Human Settlement Systems* (ed. N. M. Hansen) (Ballinger, Cambridge, MA, 1978), pp. 197–211.

[3] P. C. Austin and J. V. Tu, 'Bootstrap methods for developing predictive models', *Amer. Statist.* **58**(2) (2004), 131–137.

[4] A. C. Cameron and P. K. Trivedi, *Regression Analysis of Count Data* (Cambridge University Press, Cambridge, 1998).

https://doi.org/10.1017/S0004972709000483 Published online by Cambridge University Press

[5] R. D. Green and J. P. Doll, ‘Dummy variables and seasonality – A curio’, *Amer. Statist.* **28**(2) (1974), 60–62.

[6] J. Stillwell, ‘Inter-regional migration modelling: a review and assessment’, in: *Proceedings of the 45th Congress of the European Regional Science Association* (ed. P. Rietveld) (European Regional Science Association, Amsterdam, 2005), p. 770 [online]. Available: http://www.ersa.org/ersaconfs/ersa05/papers/770.pdf [Accessed 25 June 2008].

[7] Q. H. Vuong, ‘Likelihood ratio tests for model selection and non-nested hypotheses’, *Econometrica* **57**(2) (1989), 307–333.

[8] A. G. Wilson, ‘A family of spatial interaction models and associated developments’, *Environment and Planning A* **3** (1971), 1–32.

ANGELA PEZIC, Murdoch Childrens Research Institute, Flemington Rd, Parkville, Vic. 3052, Australia
e-mail: angela.pezic@mcri.edu.au

https://doi.org/10.1017/S0004972709000483 Published online by Cambridge University Press