UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

# Working Paper M08/05

## Methodology

# Modeling International Student Migrant Tables

Guy J. Abel

## Abstract

This paper demonstrates the use of spatial interaction models for international student migrant tables using a negative binomial regression in order to account for overdispersion. The Expectation-Maximization (EM) algorithm is used in fitting these models to account for missing cells, which are a common occurrence in international population mobility tables. Data for the five largest sending and receiving nations of international student migrants between 1998 and 2005 are used. The results of fitting a quasi-independent model, main effects models with multiple covariates and interaction models are compared with respect to the Akaike Information Criterion in order to establish the most parsimonious model. By using the EM algorithm to determine parameters in these models provides imputations for cell values previously unknown.

# Modeling International Student Migrant Tables

Guy J. Abel
Division of Social Statistics
University of Southampton

July 2008

**Abstract**

This paper demonstrates the use of spatial interaction models for international student migrant tables using a negative binomial regression in order to account for overdispersion. The Expectation-Maximization (EM) algorithm is used in fitting these models to account for missing cells, which are a common occurrence in international population mobility tables. Data for the five largest sending and receiving nations of international student migrants between 1998 and 2005 are used. The results of fitting a quasi-independent, main effects with multiple covariates and interaction models are compared with respect to the Akaike Information Criterion in order to establish the most parsimonious model. By using the EM algorithm to determine parameters, imputations for cell values previously unknown are obtained.

## 1 Introduction

The application of gravity or spatial interaction models to international population mobility tables is almost non-existent. This is due to a lack of reliable data on movements between multiple countries as no single agency exists to manage data collection. The unreliability of migration data can be divided by two characteristics: inconsistent data and missing data. Data inconsistencies can be caused by many factors such as a lack of comparability in definitions, timings and effectiveness of collection methods between national statistical collection agencies; see Nowok et al. (2006). Missing data occurs when individual nations fail to provide enough depth in data collection to detect the origin or destination of migrants entering or departing their country leaving rows, columns or cells in a mobility table missing.

For international student migrant data the problems of inconsistent and missing data can be overcome to enable the modeling of a consistent and complete international migrant table. Inconsistencies between international student migrant data are minor; see Tremblay (2002). Unlike estimates of national level population flows, definitions of a student migrants may vary only slightly between countries, timings are irrelevant as data is

collected on current stocks rather than flows and data collection is relatively easy due to the formal process involved in enrolment at a foreign higher education institution. Missing international student mobility data, however, like national level mobility data still occurs, especially into countries that have traditionally received fewer student migrants. In this paper, models traditionally used in the analysis of internal mobility are applied at the international level to student migrant tables. These models are fitted using the Expectation-Maximization (EM) algorithm to account for the missing reported cells values, where previous fitting methods for gravity and spatial interaction model are unable to do so. This new application of a popular statistical algorithm to fitting population mobility tables can overcome the problem of missing cell values often found in international migration data.

## 2 Models for Population Mobility Tables

Population mobility between multiple regions is commonly presented in a square table with off diagonal entries containing the number of people moving or residing from any given origin to any given destination. These are known as tables or matrices of migration flows or migrant stocks. Flowerdew (1991) outlines two main approaches to the analysis of these tables that are commonly used for internal mobility data: the gravity model and the spatial interaction model. The gravity model derives from Stewart (1948) and Olsson (1965) relying on statistical estimation of mobility levels given information on each origin, destination and measures of interactions between them. The spatial interaction model, associated with Wilson (1970) is based on mathematical algorithms to calibrate a constrained model to origin and destination totals. There are numerous formulations of spatial interaction models such as bi-proportional adjustment, information gain minimizing and entropy maximizing which include various constraints and interaction terms.

Poisson regression models have become a popular method for representing migration models as they relate gravity and many spatial interaction models in a single comparative framework. Willekens (1983) and Flowerdew (1991) showed that a Poisson regression model with either a row or column dummy covariate are equivalent to a origin or destination constrained spatial interaction model, and where both sets of covariates are present a doubly constrained spatial interaction are obtained. Such representations, with only categorical covariates present, are also known as log-linear regression models of Birch (1963). When row or column dummy covariates are not included a gravity model with an assumed Poisson distributed response are represented.

Poisson regression models are part of a range of statistical models known as generalized linear models of Nelder and Wedderburn (1972), which link together a number of models and techniques that relate a random response variable to a systematic linear predictor.

This statistical formulation of a mobility table has several important advantages over more traditional approaches. Willekens (1983) noted that log-linear regression models enhance the structural analysis of spatial interaction, have greater clarity and simplification of parameter estimation and open the opportunity to apply a wide range of statistical theory. Guy (1987) expand upon this final point for all Poisson regression models, noting the ability to provide standard diagnostics and better model specification. In addition non-specialist statistical software may be used to fit generalized linear models using efficient algorithms for obtaining maximum likelihood parameter estimates and with great flexibility for alternative functional forms to extend models beyond conventional size and distance variables and with alternative error specifications.

Flowerdew and Aitkin (1982) noted some drawbacks in implementing Poisson regression models to population mobility tables. Arguably, the most prominent of these were an inability to provide an adequate fit to data. Previous attempts to fit log-linear models, such as that of Flowerdew and Lovett (1988) and Flowerdew (1991), showed that the best fitted models had origin and destination (or table row and column) covariates. Despite adding further interaction-based covariate information which improved model fits, the remaining deviance of models were still deemed unsatisfactory. The lack of fit was diagnosed to the equivalence of the first and second moment of the Poisson distribution. The use of a single parameter distribution assumed each migrant moved from a given origin to a destination occurred independently, having controlled for explanatory factors. Congdon (1991) noted that at an aggregate level on which mobility tables tend to based, there may exist numerous factors that operate at lower levels. Without the ability to disaggregate data by such factors, such as personal characteristics, Poisson regression models may fit poorly.

One solution to this problem was to fit an ordinary least squares linear regression to the logarithm of migrant counts. Flowerdew and Aitkin (1982) noted this approach had a number of problems when fitted to migration count data. Difficulties included the introduction of the logarithmic scale which consequently biased an estimate of the mean when the antilogarithm was taken. This may have resulted in wrongly signed or insignificant coefficients included in a model. In addition, a log normal assumption for a count response had a theoretical dissatisfaction of modeling a discrete valued process by a continuous distribution and also presupposes a common variance for mobility table data where there is often a wide variation (heteroscadicity) in cell values. Davies and Guy (1987) suggested three alternative solutions for when a Poisson assumption in mobility tables was violated: a parametric approach of a negative binomial regression model, a quasi-likelihood approach of introducing a new parameter for the mean-variance ratio and a pseudo likelihood approach of estimating a variance-covariance matrix of parameter estimates given a misspecified model. In this paper, the former of these three is further

explored due to its ease of implementation when missing cell values are present.

### 2.1 Negative Binomial Regression Model

The negative binomial distribution is a two-parameter family that allows the mean and variance to be fitted separately, as opposed to a Poisson regression model. Consider a response variable $Y$ and a set of explanatory variables of $X$ of dimension $p$. A Poisson regression models would stipulate the distribution of $Y$ given $X$ is Poisson with mean equal to $\mu = \exp(\beta X)$ where $\beta$ is a vector of $p$ regression parameters. This may be abbreviated to $Y \sim Po(\log \mu)$ or in a generalized linear model formulation as $Y \sim Po(g(\mu))$ where $g(\mu) = \log \mu$ is the canonical (log) link function that links the random and systematic components of the models. Lawless (1987) showed that in order to allow for extra Poisson variation we may employ a negative binomial regression models considered as

$$\Pr(Y = y \,|X) = \frac{\Gamma\left(y + a^{-1}\right)}{y!\Gamma\left(a^{-1}\right)} \left(\frac{ag(\mu)}{1 + ag(\mu)}\right)^{y} \left(\frac{1}{1 + ag(\mu)}\right)^{a^{-1}}, y = 0, 1, \ldots, \qquad (1)$$

where $a \geq 0$ and is often referred to as the index or dispersion parameter. The mean and variance of Y are

$$E(Y \,|X) = g(\mu) \text{ and } Var(Y \,|X) = g(\mu) + ag(\mu)^2. \qquad (2)$$

For such a model this may be written as $Y \sim NB(g(\mu), a)$, where the log-link function used in a Poisson regression model can be employed. When the dispersion is zero the Poisson model is obtained. Agresti (2002) noted that a negative binomial model may be fitted in a similar manor as Poisson regression models when the dispersion parameter is known. This can commonly be done by implementing a Iteratively Reweighted Least Squares (IRLS) procedure which McCullagh and Nelder (1983) proved to converge to the maximum likelihood solutions for parameter estimates. When the dispersion parameter is not known, three possible methods exist to obtain maximum likelihood parameter estimates: a Newton-Raphason routine for fitting all parameters simultaneously, the evaluation of the profile likelihood for various fixed $a$ and an alternation strategy of 1) using IRLS to solve mean parameter estimates $\beta$ for fixed $a$ and using 2) using Newton-Raphason to estimate $a$ from fixed $\beta$, until convergence.

### 2.2 The Expectation-Maximization (EM) Algorithm

The EM algorithm is an iterative algorithm for maximum likelihood estimation in incomplete data problems. Used in multiple statistical settings the EM algorithm is a prominent tool in estimation when there are missing data on random variables (such as the number of migrants between two countries) whose realizations would otherwise be observed. Developed by Dempster et al. (1977) the motivating idea behind the EM algorithm is rather

than perform a complex estimation we may augment the missing parts of a data set with temporary, complete data to allow the estimation of model parameters to proceed in a cycle of simple estimation steps. Each cycle of the EM algorithm consists of two steps.

1. If we let $\theta^r$ denote the current guess of the parameters $\theta = (\beta, a)$ at iteration $r$, $\mathbf{y}$ be the vector of known data containing $n$ counts, and $\mathbf{z}$ denote the missing data to be augmented, the E-step (expectation step) finds the expected augmented log-likelihood $Q(\theta)$ if $\theta^r$ were $\theta$. This can be expressed as

$$\begin{aligned} Q(\theta|\theta^r) &= E(l(\theta|\mathbf{y},\mathbf{z})|\mathbf{y},\theta^r \\ &= \int_{\mathbf{z}} l(\theta|\mathbf{y},\mathbf{z}) f(\mathbf{z}|\mathbf{y},\theta^r)\, dz \end{aligned} \tag{3}$$

   where $l(\theta|\mathbf{y},\mathbf{z})$ is the log likelihood of $\theta$ given the augmented data

2. The M-step (maximization step) determines $\theta^r$ by maximizing the expected augmented log-likelihood

The algorithm is iterated until $||\theta^{r+1} - \theta^r||$ or $||Q(\theta^{r+1}|\theta^r) - Q(\theta^r|\theta^r)||$ is sufficiently small, and hence a maximum of the augmented log-likelihood is reached.

When fitting a regression model for mobility tables the M-step can be easily implemented in standard statistical software by performing a fit to the current complete data at each iteration. Little and Rubin (2002) noted the EM algorithm in many cases is conceptually and computationally easy to construct. In addition the algorithm can be shown to coverage reliably to a local maximum or saddle point of the observed likelihood but may do so with a slow rate when there is a large fraction of missing data.

## 3 Student Migrant Data

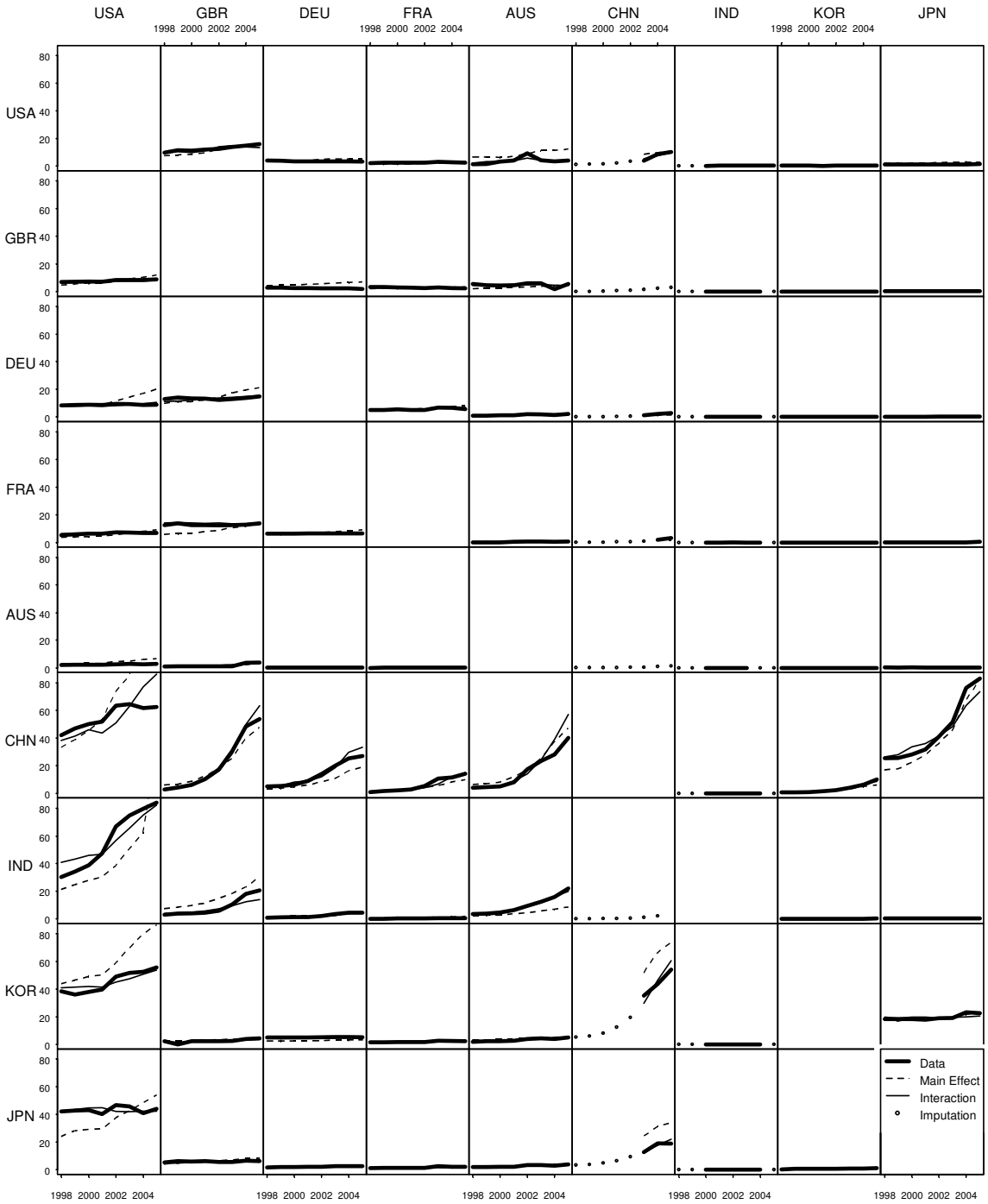
A negative binomial regression model was fitted using the EM algorithm to international student migrant data between 1998 and 2005 for the number of foreign students from nine nations (USA, Great Britain, Germany, France, Australia, China, India, Korea and Japan) in each of these countries. These nations were chosen as they often rank as the five highest receiving countries (USA, Great Britain, Germany, France, Australia) by volume of foreign students in higher education, and the five largest sending nations of foreign students (China, India, Korea, Japan and Germany) throughout the time period. Data were obtained from United Nations Educational, Scientific and Cultural Organization (UNESCO) and Organization for Economic Co-operation and Development (OECD) web sites. A combination of data was used as figures often matched and a single source alone would leave some columns (of destination reporting countries) completely empty. Such characteristics in the data would not allow the EM algorithm to fit a model that included destination-specific effects as no sufficient statistics required for parameter estimation

would be obtainable for countries that provided no information. Data from both these organizations included origin information of foreign students defined as either the student's citizenship or place of permanent residence. Definitions vary by reporting country which often pursue national practises rather than that of the organizations guidelines. Tremblay (2002) noted that such inconstancies are believed to have only a minor effect and data may serve as an adequate picture of student migrant populations. The use of citizenship or residence definitions does not register a mobility measure (such as migration flow), which might be obtainable from origin information collected by previous place of residence or education; see Kelo et al. (2006). However, the tracking of migratory trends may still be performed by analyzing changes over a number of time periods of foreign student stocks information obtained by the student's citizenship or place of permanent residence.

Of the 576 cells (made from a $9 \times 9$ non-diagonal mobility table over 8 time periods), 487 had observed values from at least one of the identified data sources. In 271 cases data from both organizations were available for which 191 reported the same value. Differences occurred as OECD data were based on a UNESCO-OECD-Eurostat coordinated collection system on education statistics, whilst UNESCO publishes data from its UIS annual data collection. Hence reporting partners such as Ministries of Education or National Statistical Offices may have differed. Differences between data were all minor with a few exceptions where a reporting problem seemed apparent, for example only 191 Korean students were reported to study in the USA in 1999 by UNESCO compared to 36,085 in the OECD data. As the OECD had less missing values for the selected nations and no apparent reporting problems, it was treated as the preferred source for all cells where UNESCO data was also available. For one nation, China, no information in either data sources was given. Additional data was sought, again to enable a destination-specific covariate to be estimated by the EM algorithm. Reported levels of foreign students were collected from the China Scholarship Council web site, the main organization responsible for foreign student data in that country. This reduced the number of cells without observation from 89 to 73. One final adjustment was made to the number of foreign students from China in the USA from 2003 to 2005 which included Taiwanese students unlike past numbers. In these cases OECD and UNESCO data were replaced by levels reported on the Institute of International Education web site whose data had a separated Chinese and Taiwanese students in all years, and from 1998 to 2002 had reported number of Chinese students matching that of both international organizations. Checks for the combining of Chinese and Taiwanese foreign student numbers in all other nations' local data sources were made but no large errors were found.

Plots of the data over time for each cell of the migrant table are shown by the thick solid line in Figure 1 where origins are shown on the vertical axis and destinations on the horizontal axis. The order of nations are arranged by predominantly receiving countries

Figure 1: Data and Model Fits of Foreign Students (000's) from each Origin-Destination Combination, 1998-2005.



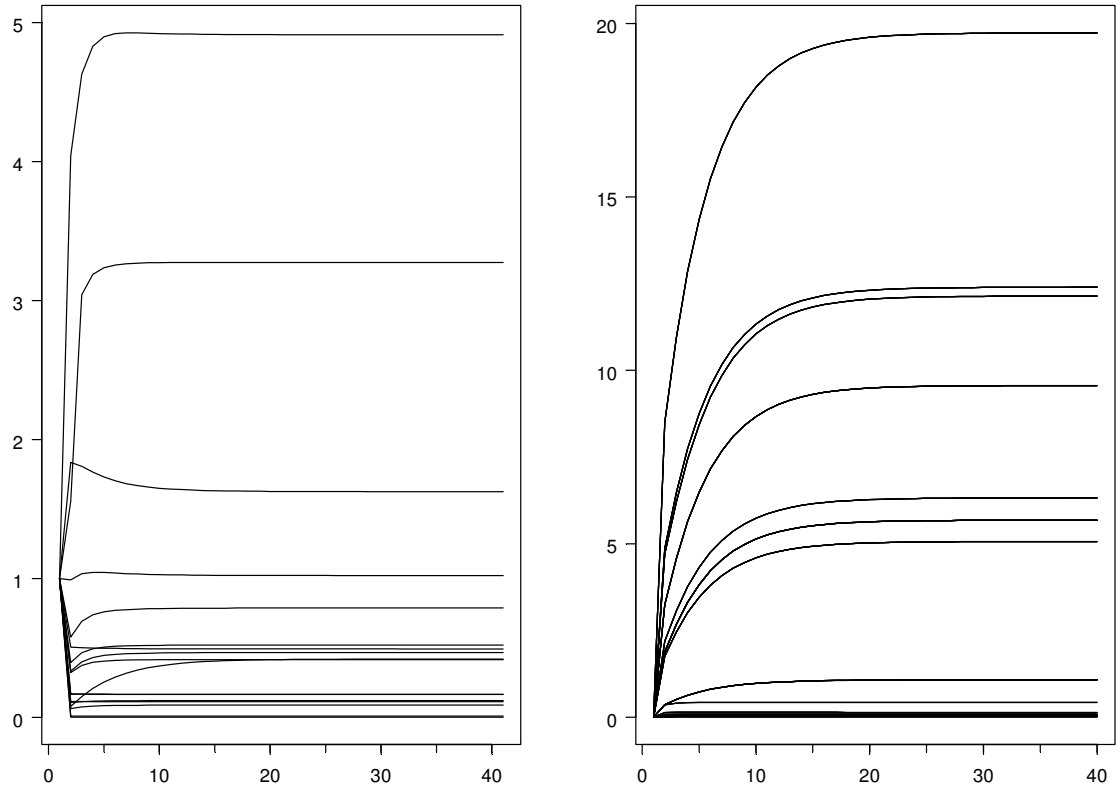
followed by predominantly sending nations. When ordered as such, the lowest number of foreign students will tend to be in the upper right quarter of the migrant table, moving from traditional receiving countries to traditional sending countries. The highest counts are in the bottom left moving from traditional sending to receiving countries. Intermediate values are found in the top left and bottom left right quarters. Higher values in these quarters occur for movements between countries in the same region, such as the European countries and East Asia. The majority of movements remain constant or increase over the time period. Selected moves out of China showed increasing trends, notably to Great Britain, Australia and Germany. The number of Indian students in the USA, Australia and Great Britain also showed an increasing pattern. Notable over time was a deceleration of the number of students moving into the USA from India, China, and Korea in the later part of the series. This trend may be attributed to the changes in immigration policy of the USA after the attacks of September 11th 2001, motivating students to remain at home or go to other countries for higher education; see Altbach (2004). This event may also be responsible for some of the post-2001 increases in the number of Chinese students entering the British and Japanese higher education systems.

# 4 Modeling

A function was programmed in S-Plus 6.1 to fit negative binomial regression models using the EM algorithm. This employed the `glm.nb` function in the MASS library, Venables and Ripley (1999) in the M-step to determine the parameter estimated at each cycle, given the augmented data. As the dispersion parameter was unknown, the `glm.nb` function estimates parameters by an alternating strategy similar to that described previously.

Initially, a spatial interaction model equivalent to quasi-independent model was fitted to determine an overall push and pull effects of each nation. As previously mentioned, such models give superior fits than that of gravity models but at the cost of aliasing out additional origin and destination effects. All origin and destination effects are able to be estimated using the EM algorithm as all rows and column have at least a single observation. The exponentiated origin parameter values estimated from this model measured the level of attraction of foreign students in the selected system in comparison to the USA which was used as a reference category. Values varied from 4.912 and 1.624 for China and Korea to 0.088 and 0.417 for Australia and France respectively (with respect to unity for the USA). Exponentiated destination parameter values (where the USA was again the reference category) varied from 0.492 and 0.419 for Great Britain and China to 0.003 and 0.009 for India and Korea respectively. The dispersion parameter was estimated to be 1.1863 from the `glm.nb` function, equivalent to $\frac{1}{1.1863}$ of $a$ in (1), with standard error 0.063. A z-test provided strong evidence that $a > 0$. Hence, suggesting a negative binomial model

Figure 2: Exponentiated Covariate Parameter Estimates, $\theta$ (left) and Missing Data Values, $\mathbf{z}$ in 000's (right) of Quasi-Independent Fit




was more appropriate than a Poisson. Figure 2 shows a trace of the iterative estimates from the EM algorithm of this model (right plot), alongside the imputed values for 73 missing cell values (left plot). A initial value of one was chosen for all parameter estimates whose values all met a convergence criteria of $||\theta^{r+1} - \theta^r|| < 0.0001$ after 41 iterations.

### 4.1 Additional Information

In order to provide more reasonable imputations the quasi-independent model was expanded upon. Eight covariates were chosen to reflect differing education systems, economic determinants, geographical and population factors that past literature suggested to influence the interaction between origin and destinations. Where possible, information across time was taken to help reflect trends in foreign student numbers seen in Figure 1. Two covariates on education systems: the total of enrollment in tertiary education of each country and origin-destination ratio on the quality of the university systems, were selected. Enrollment data was obtained from World Bank EdStats Database as a measure of the relevant population of interest, commonly used in multiple formulations of spatial interaction models; see Sen and Smith (1995). A measure of the quality of university systems

was calculated from the ratio of the number of institutions in top 500 of Shanghai Jiao Tong University Rankings between each origin and destination, regarded internationally as one of the most authoritative educational ranking indexes; see Healey (2007). Rankings and their implied reputation of a country's education system have appeared significant in studies into the motivations of studying abroad, such as those by Mazzarol and Soutar (2002) and Altbach (2004). Two covariates on the economic relationships: difference in Gross Domestic Product (GDP) and the logarithm of trade volume between each origin and destination, were selected. Data on GDP at Purchasing Power Parity were obtained from the World Bank EdStats Database. Previous empirical modeling research in student migration to and from single countries, such as that of Lee and Tan (1984), Agarwal and Winkler (1985), McMahon (1992) and Dreher and Poutvaara (2005), have all included GDP information in their models suggesting that higher numbers of foreign students in richer nations may be due to the pursuit of finding employment in their country of study after finishing their education and the lack of availability of education in less wealthy countries. McMahon (1992), Tremblay (2002) and Mazzarol and Soutar (2002) have suggested foreign student numbers are related to the connectedness of countries' economies with each other and to a world system. Data on the value of all commodities imported into each country for all origin nations, obtained from the United Nations Commodity Trade Statistics Database, was included to reflect this factor. Two measures of geographical links: distance and region covariate, were selected. Data on the logarithm of distance in kilometers involved in a movement between each nation's capital city was obtained from Gleditsch and Ward (2001). Distance, as with population, has appeared in multiple formulation of spatial interaction models fitted for internal migration data. A region covariate was set up to allow some consideration for the ease of moves between countries that share a boarder or that take place within a local system. This had three levels, between region moves, inter-East Asia moves and inter-European moves. The later was expected to be highest due to programs that encourage student mobility such as Erasmus; see Ruiz-Gelices et al. (2000). Finally two population measures: migrant stocks and language were chosen. The inclusion of a migrant stock covariate allowed control on the number of students whose family may have migrated to a particular country where they may have previously been educated to the secondary level and still classed as a foreign student in the UNESCO and OECD data. In addition, stock data also expresses a level of social links between each country, a factor studied by Mazzarol et al. (1996) in the destination choices of potential student migrants. A origin-destination migration stock tables was derived from Parsons et al. (2005) who had complied a global bilateral stock database based on the 2000 round of population censuses. A categorical covariate on language was constructed to reflect four different types of moves involved in a study abroad: moves to English speaking nations, moves between English speaking nations, moves away from English speaking nations to a

different language and all other moves. These separations were created to reflect Anglo-Saxon student's tendencies to move in lower numbers to non-English speaking countries as noted by Findlay et al. (2006) and the role of acquiring new languages, especially English for student migrants from non-English speaking nations, as noted by Tremblay (2002). An additional continuous covariate for time was also added to account for changes in the number of student migrants during the time period, and the correlation amongst repeated counts of the same stock over time.

### 4.2 Main Effects Model

In order to attain a better model fit, more realistic imputations and elaborate on which factors influenced patterns in the student migrant tables the Akaike Information Criterion (AIC) was used to select the most suitable variables to remain in a main effects model.

$$AIC = 2k - 2l(\theta|\mathbf{y}), \tag{4}$$

where $l(\theta|\mathbf{y})$ is the log likelihood of $\theta$ given the observed data. Comparisons of potential models were undertaken using the `stepAIC` function in the MASS library, Venables and Ripley (1999). The function operated by examining the inclusion of potential covariates by their contribution to the AIC of the model by performing a stepwise search in both directions, adding and dropping variables in the model. Included in a pre-condition in the scope of models to be searched were origin and destination covariates. All covariates were found to be effective in reducing the AIC, for which the final main effects model was 9,171 in comparison to 10,095 of the quasi independent model. Convergence when fitted with the EM algorithm was obtained after 62 iterations and the fitted values are shown by the broken line (where the imputations on previously missing student numbers are not displayed) in Figure 1.

Parameter estimates of origin and destination effects strayed from their values found in the quasi-independent as additional factors were controlled for. The exponentiated parameters effects for time (1.038), educational factors (3.368 for the logarithm of bilateral enrollment totals and 1.019 for quality based on the ratio of institution rankings), economic factors (1.189 for the logarithm of trade volume and 1.064 for the difference in GDP) and logarithm of migrant stocks (1.575) were all greater than unity implying higher levels of these covariates were associated with higher student migrant numbers, conditional upon the value of all other covariates. Problems occurred in fitting some levels of the language covariate. Parameters for the levels of moves away and to English speaking nations were linear combinations of origin and destination effects and hence unique estimates where unobtainable. This left a categorical covariate with only two levels, for which the exponentiated parameter estimate of moves between English speaking nations was 0.808, implying lower levels of moves in comparison to the reference category of moves

that incurred a change of languages. The regional categorical covariate found higher exponentiated effects on student migrants numbers between East Asian nations (5.235) and European nations (13.593) than the reference category of intra-regional moves (unity). The exponentiated effect on the logarithm of distance covariate was unexpectedly greater than unity (1.418), implying further distances where preferred. Such a result may be due to the difficulty of obtaining a single measure of physical distances between such large countries. For example the effect of a distance measure from East Asian nations to USA may have been over accentuated as the actual distance of many movements are to universities on Americas west coast but measured by distances between Asian capital cities and Washington DC. The dispersion parameter was estimated to be $\frac{1}{4.361}$ with standard error 0.262, noticeably smaller than the quasi-independent fit indicating evidence for a control on overdispersion in the main effects models.

### 4.3 Interaction Model

To gain a further superior fit the `stepAIC` function was run once more with an extended scope of models to consider all two-way interaction, with one exemption, the origin-destination interaction. This was not included as for some levels, such as an interaction between British students in China, no data existed and hence such a parameter could not be estimated using the EM algorithm. The fitting function selected eleven new interaction covariates, whilst dropping two main effects: distance and quality. Convergence when fitted with the EM algorithm was obtained after 660 iterations and the fitted values are shown in Figure 1 by the thin solid line where observed data existed, and by dots for imputations on previously missing data. For origin-destination combinations where no data existed, such as British and Australian students in China, imputations remain constantly small over time. When partial data existed in origin-destination combinations, the imputed values tend to follow the trend of model fits on the available data, for example growing over time for Koreans in China or remaining constant for Australians in India. The AIC of the interaction model was 8,198, a further reduction in comparison to the main effects model but with many more parameter estimates (from 28 to 84). The higher number of estimates, for which only a brief discussion is given in the remainder of this section, was due to the multiple interaction terms that include categorical covariates.

Of the eleven new interaction terms four (enrollment, stock, trade and region) were associated with origin and three (GDP, trade and time) with destination. These seven covariates, with the exception of origin-region interaction, were mixtures of a continuous and categorical covariates measures, providing 48 parameter estimates. Their inclusion indicated evidence for different effects by origin (or destination) of the continuous measure on the number of foreign students leaving (or arriving) varied by country conditional upon the value of all other covariates. Of the four remaining interactions (not involving origin

and destination), three were mixtures of continuous and categorical variables, namely region and stock, region and trade and language and time. These interactions indicated different effects for populations stock, trade and time on the number of student migrants varying by the given region or type of language change incurred in a origin-destination combination. The remaining interaction included in the model was between enrollment and foreign population stock. As the exponentiated estimated parameter was less than one this parameter indicated a decreasing influence of enrollment totals in tertiary education sectors at higher levels of population stocks for a given level of student migrants. The dispersion parameter was estimated to be $\frac{1}{28.583}$ with standard error 1.974, again noticeably smaller than previous model indicating further control on overdispersion in the interaction model.

# 5 Summary and Discussion

This paper demonstrates the use of negative binomial regression for international student migrant tables across time with missing cells. Such models allow the empirical modeling of the spatial interactions of student migrants between multiple countries. Previous empirical studies of international student migrants has tended to focus on single nations, using separate models for push and pull effects, such as Lee and Tan (1984), Agarwal and Winkler (1985), McMahon (1992) and Dreher and Poutvaara (2005). Such models may fail to fully capture the spatial interactions of migrant patterns. The use of the EM algorithm, as illustrated with the negative binomial regression model, could be be further expanded to model student migrant patterns with more nations and covariate information. In this paper a restriction to nine nations was used to enable an easier demonstration of methods. Consequently some large movements such as those from North African countries into France (which promoted France as a top five receiving nation) were excluded and hence may distort the results. Covariate information in international comparisons must be carefully selected to exclude data that may not be consistent when compared across nations. Previous student migration models have included covariates on educational fees, cost of living, excess demand in origin tertiary education systems and government spending on education as a percentage of GDP. For such measures, data may lack comparability across nations and is not always complete. In addition, covariate measures such as distance may not be adequate to fully capture their effect of moves between such diverse and large nations.

When missing data is present, the success of the EM algorithm to fit a spatial interaction model is dependent on some data being available from all reporting destination countries. In this study this was achieved by combining comparable international organization data and analyzing trends over multiple time periods. The spatial interaction

model was chosen in order to provide the best fit to the data. This however caused problems in estimating parameters which may be of interest. Fitting all levels of categorical covariates in spatial interaction models was not always possible, as demonstrated with the language covariates, where origin and destination effects aliased moves to and from English speaking nations. Row and column covariates in mobility tables also alias out any origin or destination-specific effects that a modeler may wish to study, and hence their inclusion may not always be desired. However, these effects could explored through interaction terms with origin and destinations covariates which, as seen in this paper, aid the fit of the model. Considering interaction terms also reduced the amount of dependent variables required, dropping distance and rankings from the model, which may be of benefit if the availability of consistent and accurate covariates is limited. An alternative approach for including categorical covariates could consider origin-destination combinations in a marginal models, which may be fitted using Generalized Estimating Equation of Zeger et al. (1988). Marginal models would enable the exclusion of origin and destination specific parameters, allowing more complex categorical covariates to be fitted. Such an approach was not taken in this paper due as imputation methods for missing data would become more complicated.

Better fits for the interaction model could be further achieved by considering further covariates or redefining existing ones. The time covariate was considered as continuous in this study for ease of interpretation, but it could have been considered as a categorical factor. This would allow time-specific effects to be estimated in the same manor as origin-specific and destination-specific resulting in a superior fitting model, but at the cost of more parameters. Interactions terms between these covariates would lead to a saturation of the model but effects may not always be estimated using the EM algorithm if no data exist in a given time period for a given reporting destination. It is useful to note that if interest lay in controlling for specific origin-destination combinations, such as Chinese students in the USA before and after 2001, a covariate could be built to include this term and induce a better model fit in that cell.

The negative binomial regression model proved an effective tool to deal with overdispersion of the data. The use of alternative error assumptions such as a Poisson would have lead to worse fitting models and non robust standard errors. The building of models relied upon comparisons of competing AIC calculated on the log-likelihood of the incomplete, observed data, rather than the complete data. As Cavanaugh and Shumway (1998) noted it is more desirable to fit a model based on the complete data for which models are originally postulated for and hence include information on the missing data. Criterions, such as the AIC-cd of Cavanaugh and Shumway (1998) and KIC-cd of Seghouane et al. (2005), allow the calculation of the separation between the fitted model for the complete data and the true or generating model. Both criterions require models to be fitted using

the Supplemented-EM algorithm of Meng and Rubin (1991) which requires further computations during the EM algorithm. Both criterions have demonstrated the tendency for the AIC based on only observed data to over fit data when model building.

Despite the common occurrence of missing data in international population mobility tables, the application of the EM algorithm is sparse. Willekens (1999) suggested the EM algorithm as a possible method to fit spatial interaction models to constrained margins. Raymer et al. (2007), in an expansion of his model found the EM algorithm in this situation to be equivalent to a conditional maximization. Imputations for missing cells in international tables have tended to focus on mathematical relationships of different data sets rather then a statistical solutions. Parsons et al. (2005) used an entropy measure between different migrant stock definitions, whilst Poulain (1999) used stock data to replace missing flow data and a constrained minimization technique to harmonize migration flow data. The EM algorithm allows a wide range techniques for the statistical modeling of mobility tables to be applied. By doing so, models are able to account for missing data and impute missing cell values based on statistical assumptions and covariate information based on migration theory. In this paper this was conducted on a set of combined data but could be applied to a data set with more missing cells at the cost of a slower convergence and perhaps more intrinsically non-identifiable parameters.

In conclusion, the modeling of international student migrant tables may be undertaken despite missing data. There exists a number of options for fitting and building models to account for overdispersion and missing data. In this paper, the application of negative binomial regression were compared using comparisons of model's AIC which proved an effective strategy to deal with overdispersion. Fitting such models with the EM algorithm provided a foundation for imputing the missing cell values, a common occurrence in international population mobility tables.

## 6 Acknowledgements

This work was undertaken with financial support of the Economic and Social Research Council. I am very grateful to Dr. John Aston (Institute of Statistical Science, Academia Sinica, Taiwan) and Professor Ji-Ping Lin (Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan) for their comments and suggestions throughout the work as well as Dr. James Raymer and Professor Peter W.F. Smith (Division of Social Statistics, University of Southampton, UK) and Corrado Giulietti (Division of Economics, University of Southampton, UK) for their feedback on earlier drafts.

## References


Agarwal, V. and D. Winkler (1985). Foreign Demand for United States Higher Education: A Study of Developing Countries in the Eastern Hemisphere. *Economic Development and Cultural Change 33*(3), 623–644.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley-Interscience.

Altbach, P. (2004). Higher Education Crosses Borders. *Change 36*(2), 18–24.

Birch, M. (1963). Maximum Likelihood in Three-Way Tables. *Journal of the Royal Statistical Society, Series B 25*, 220–233.

Cavanaugh, J. and R. Shumway (1998). An Akaike Information Criterion For Model Selection In The Presence Of Incomplete Data. *Journal of Statistical Planning and Inference 67*(1), 45–65.

Congdon, P. (1991). General linear modelling: Migration in london and south east england. In J. Stillwell and P. Congdon (Eds.), *Migration Models: Macro and Micro Approaches.*, Chapter 7, pp. 113–136. London, England: Belhaven Press.

Davies, R. and C. Guy (1987). The Statistical Modeling Of Flow Data When The Poisson Assumption Is Violated. *Geographical Analysis 19*(4), 300–314.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Dreher, A. and P. Poutvaara (2005). Student Flows and Migration: An Empirical Analysis. Technical report, Center for Economic Studies, Ifo Institute for Economic Research.

Findlay, A., R. King, A. Stam, and E. Ruiz-Gelices (2006). Ever Reluctant Europeans: The Changing Geographies of UK Students Studying and Working Abroad. *European Urban and Regional Studies 13*(4), 291.

Flowerdew, R. (1991). Poisson Regression Models Of Migration. In J. Stillwell and P. Congdon (Eds.), *Migration Models: Macro and Micro Approaches.*, Chapter 6, pp. 92–113. London, England: Belhaven Press.

Flowerdew, R. and M. Aitkin (1982). A Method Of Fitting The Gravity Model Based On The Poisson Distribution. *Journal of Regional Science 22*(2), 191–202.

Flowerdew, R. and A. Lovett (1988). Fitting Constrained Poisson Regression Models To Interurban Migration Flows. *Geographical Analysis 20*(4), 297–307.

Gleditsch, K. and M. Ward (2001). Measuring Space: A Minimum-Distance Database and Applications to International Studies. *Journal of Peace Research 38*(6), 739.

Guy, C. (1987). Recent Advances In Spatial Interaction Modelling: An Application To The Forecasting Of Shopping Travel. *Environment and Planning A 19*(2), 173–186.

Healey, N. (2007). Is Higher Education In Really Internationalising? *Higher Education*, 1–23.

Kelo, M., U. Teichler, and B. Wachter (2006). Toward Improved Data on Student Mobility in Europe: Findings and Concepts of the Eurodata Study. *Journal of Studies in International Education 10*(3), 194.

Lawless, J. (1987). Negative Binomial And Mixed Poisson Regression. *Canadian Journal of Statistics 15*(3), 209–225.

Lee, K. and J. Tan (1984). The International Flow of Third Level Lesser Developed Country Students to Developed Countries: Determinants and Implications. *Higher Education 13*(6), 687–707.

Little, R. and D. Rubin (2002). *Statistical Analysis With Missing Data*. Wiley.

Mazzarol, T., L. Savery, and S. Kemp (1996). *International Students who Choose Not to Study in Australia: An Examination of Taiwan and Indonesia*. AEIF Policy, Research and Analysis Section.

Mazzarol, T. and G. Soutar (2002). Push-Pull Factors Influencing International Student Destination Choice. *The International Journal of Educational Management 16*(2), 82–90.

McCullagh, P. and J. Nelder (1983). *Generalized Linear Models*. London, England: Chapman Hall.

McMahon, M. (1992). Higher Education in A World Market. *Higher Education 24*(4), 465–482.

Meng, X. and D. Rubin (1991). Using EM to Obtain Asymptotic Variance-Covariance Matrices: the SEM Algorithm. *Journal of the American Statistical Association 86*(416), 899–909.

Nelder, J. and R. Wedderburn (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370–384.

Nowok, B., D. Kupsizewska, and M. Poulain (2006). Statistics on international migration flows. In N. P. M. Poulain and A. Singleton (Eds.), *Towards the Harmonisation of*

*European Statistics on International Migration (THESIM)*, Chapter 8, pp. 203–233. Louvain-La-Neuve, Belguim: UCL–Presses Universitaires de Louvain.

Olsson, G. (1965). *Distance and Human Interaction: A Review and Bibliography*. Regional Science Research Institute.

Parsons, C., R. Skeldon, T. Walmsley, and L. Winters (2005). Quantifying the International Bilateral Movements of Migrants. *8th Annual Conference on Global Economic Analysis, Lübeck, Germany, June*, 9–11.

Poulain, M. (1999). International Migration Within Europe: Towards More Complete and Reliable Data. Joint ECE-Eurostat Work Session on Demographic Projections, Perugia, Italy, June.

Raymer, J., G. Abel, and P. Smith (2007). Combining census and registration data to estimate detailed elderly migration flows in England and Wales. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 170*(4), 891–908.

Ruiz-Gelices, E., R. King, and A. Favell (2000). International Student Migration in Europe and the Institutionalization of a'European Identity.

Seghouane, A., M. Bekara, and G. Fleury (2005). A criterion for model selection in the presence of incomplete data based on Kullback's symmetric divergence. *Signal Processing 85*(7), 1405–1417.

Sen, A. and T. Smith (1995). *Gravity models of spatial interaction behavior*. Springer-Verlag New York.

Stewart, J. (1948). Demographic Gravitation: Evidence and Applications. *Sociometry 11*(1/2), 31–58.

Tremblay, K. (2002). Student Mobility Between And Towards OECD Countries: A Comparative Analysis. *International Mobility of the Highly Skilled*.

Venables, W. and B. Ripley (1999). *Modern Applied Statistics with S-Plus*. Springer.

Willekens, F. (1983). Log-Linear Modelling Of Spatial Interaction. *Papers in Regional Science 52*(1), 187–205.

Willekens, F. (1999). Modeling Approaches to the Indirect Estimation of Migration Flows: From Entropy to EM. *Mathematical Population Studies 7*(3), 239–78.

Wilson, A. (1970). *Entropy In Urban And Regional Modelling*. Pion, London.

Zeger, S., K. Liang, and P. Albert (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics 44*(4), 1049–1060.