Combining Census and Registration Data to Analyse Ethnic Migration Patterns in England from 1991 to 2007

James Raymer, Peter W.F. Smith and Corrado Giulietti
Southampton Statistical Sciences Research Institute
University of Southampton

Forthcoming in Population, Space and Place

Revised 14 April 2009

ABSTRACT

In this paper, we develop a model that allows us to combine annual (incomplete) registration data with (auxiliary) census data. The result is a synthetic data base that can be used to analyse the evolution of specific migrant groups over time. For illustration, we model the evolution of ethnic interregional migration in England by age and sex from 1991 to 2007 by combining National Health Service registration data with 1991 and 2001 Census data. This annual time series of detailed migration flows are useful for both planning and for understanding ethnic population redistribution. Furthermore, changes over time can be related to regions exhibiting, for example, high unemployment, high costs of living or high immigrant concentrations.

Keywords: combining data, internal migration, ethnicity, log-linear models, England

1. INTRODUCTION

We develop a model that allows us to combine annual (incomplete) registration data with (auxiliary) decennial census data. The result is a synthetic data base that can be used to analyse the evolution of specific migrant groups over time and their relationships, for example, with areas of high unemployment, high costs of living or high immigrant concentrations. The advantages to having such a data set are numerous. Detailed estimates of migration flows are needed so that local governments have the means to improve their planning policies directed at supplying particular social services or at influencing levels of migration. This is important because migration is currently (and increasingly) the major factor contributing to population change at sub-national levels in many countries throughout the world, including England. Furthermore, our understanding of how or why populations change requires more detailed and updated information about migrants. Without these, the ability to predict, control or understand that change is limited.

In this paper, we model age- and sex-specific interregional migration patterns of four ethnic groups in England from 1991 to 2007 to illustrate the methodology for combining data. The study of ethnic migration is important for understanding the social networks and behaviours of ethnic populations (Finney and Simpson, 2008; Raymer and Giulietti, 2009). Many ethnic minority populations are disadvantaged because of their relatively low socioeconomic status caused by their recent arrival or by their living in ethnically segregated areas. Hence, the study of ethnic internal migration provides researchers and policy makers with indicators on how well they are integrating into society and responding to changing economic conditions. The study of ethnic migration also tells us how the population is redistributing itself across the country, allowing one to assess where areas of growth are (or will be) and whether this growth is ethnic-specific. Finally, as pointed out by Finney and Simpson (2008) and Stillwell *et al.* (2008), very little is known about the internal migration behaviours of different ethnic groups in the United Kingdom. Our study provides a framework for combining migration data in England to study ethnic migration over time, thereby increasing the evidence base for analyses.

1.1 Motivation

The reasons for internal migration are many. People move for employment, family reunion or amenity reasons. Reported statistics on these flows, on the other hand, are relatively confusing or nonexistent (Bell *et al.,* 2002). There are three main reasons. First, no consensus exists on what exactly is a 'migration'. Second, the event of migration is rarely measured directly. More often it is inferred by a comparison of places of residence at two points in time or as a change in residence recorded by a population registration system. Third, countries often use multiple data collection systems (e.g., population registers, censuses and surveys) to obtain information on migration. So how does one overcome these obstacles to obtain an overall and consistent picture of the migration patterns occurring, say, within a specific country? One possibility is to have a methodology for combining existing migration data that accounts for the various strengths offered by the different sources.

Inadequate or missing migration data makes analysing the time trends of, for example, Whites and ethnic minorities, young and elderly, first and second generation immigrants, skilled and unskilled, and employed and unemployed very difficult or incomplete. Detailed migration data are usually only available from censuses, which occur only every ten years and are published three to four years after the census date. General purpose surveys often collect migration data but, because of relatively small sample sizes, they are usually inadequate below the national or broad regional levels. Population registers may be used to track migration flows. These sources, however, often do not contain much demographic, socioeconomic or spatial detail. Also, because migration data are generally collected from sources that have other purposes, the questions underlying the patterns may not fit a particular research question of interest, e.g., measuring migrant status tells us little about migration frequency. There may also be situations in which the required data are available but cannot be considered reliable due to, for example, age misreporting. Missing data are usually caused by data suppression or by non-response.

In order to include migration data from different sources in a study, one has to first account for the differences in measurement (Bell *et al.,* 2002; Long and Boertlein, 1990; Morrison *et al.,* 2004; Rogers *et al.,* 2003; Rogerson, 1990; United Nations, 1992). For example, migration events, which can occur multiple times within a one year time period, are captured by population registration systems while changes in residential status (or transitions) from one point in time to another are captured by censuses (and surveys). These two data collection systems capture two different types of migration data, i.e., 'migrations' and 'migrants' (Rees and Willekens, 1986). Despite these conceptual differences, Boden *et al.* (1992) found high levels of correlation between the NHSCR and census data's in-migration, out-migration and net migration totals for England and Wales. More recently, Raymer *et al.* (2007), in analysing elderly internal migration, found that that the main differences between the 2000-2001 NHSCR flows and the 2001 Census flows were found in the levels of migration. The spatial patterns, on the other hand, were very similar after controlling for the levels. Knowing that the census and population health registers have similar underlying structures allows us to combine these two data sources to study the evolution of detailed migration patterns over time.

1.2 Background

In the United Kingdom, there have been many studies that have examined or modelled internal migration flows (e.g., Bates and Bracken, 1982, 1987; Bell and Rees, 2006; Champion, 1996; Dixon, 2003; Finney and Simpson, 2008; Fotheringham *et al.,* 2000b; Kalogirou, 2005; Stillwell, 1994). Other studies have examined the determinants of internal migration (e.g., Fotheringham *et al.,* 2004) and the description of social change caused by international migration (e.g., Dorling and Rees, 2003; Rees and Butt, 2004), including the linkages between immigration and internal migration (Hatton and Tani, 2005; Simpson and Finney, 2009; Stillwell and Duke-Williams, 2005). These studies, however, have not combined the various internal migration data sources available in the United Kingdom to study the evolutions of detailed migration patterns over time. Our research does. Moreover, we identify the important structures underlying the detailed migration patterns, simplifying the modelling process. The result is a set of detailed estimates that contains the known levels of recent migration and auxiliary information from, say, a most recent census or an extrapolation of auxiliary information from two or more censuses. These estimates are useful for understanding intercensal migration patterns and for regional or local planning.

In England, ethnic minority populations differ greatly from the White majority population in terms of regional growth rates and socio-demographic compositions (Dorling and Rees, 2003; McCulloch, 2007; Rees and Butt, 2004; Robinson, 1993). To identify how internal migration is contributing to these differences, we need an account of ethnic migrants and their characteristics, such as place of residence, current position in the life course, health, socioeconomic status and ethnicity (see, e.g., Faggian *et al.,* 2006; Fotheringham *et al.,* 2004; Finney and Simpson, 2008; Hussain and Stillwell, 2008; Raymer and Giulietti, 2009; Simpson and Finney, 2009; Stillwell *et* *al.,* 2008). A data set on ethnic internal migration by age, sex and over time is useful for detecting when and how certain ethnic groups in the population have become more spread out or more concentrated as a result of internal migration, which can then be compared with other studies that focus on immigration or population change as measured by the decennial censuses.

2. AVAILABLE DATA

In England, the most reliable internal migration data come from the decennial censuses and the NHSCR. Census information contain much of the detail needed for analyses, but are only collected every ten years and contain some problems of incomparability between censuses for certain variables (see Stillwell and Duke-Williams, 2007 for recent discussion). Migration data from the NHSCR are available annually but with minimal information on migrant behaviour (i.e., only origin, destination, age and sex are available) and with a tendency to miss important population groups, such as young adult males, who are known to be less inclined to register (Fotheringham *et al.,* 2004). However, the registration data constitute a good up-to-date source of internal migration as nearly all residents in England are patients of a general practitioner employed by the NHS, including those who may also have private healthcare provision. Furthermore, the average delay between moving house and registering with a new general practitioner is about one month (ONS Migration Statistics Unit, 2002).

For this study, we estimate the 1991 to 2007 annual migration flows between the nine Government Office Regions (GOR) and for sixteen five-year age groups (i.e., 0-4, 5-9, ..., 75+ years), two sexes and four ethnic groups. The nine regions consist of the North East, North West, Yorkshire and the Humberland, East Midlands, West Midlands, East of England, South East, South West and London.[1] The four ethnic groups are White, South Asian (i.e., Indian, Pakistani and Other South Asian), Black (i.e., African, Carribbean and Other Black) and Chinese & Other (including mixed ethnicity). These broad classifications of ethnicity are used for two practical reasons. First, they are useful to show how different ethnic groups exhibit different migration patterns. Second, these classifications were, more or less, consistent in both the 1991 and 2001 Censuses. Including more ethnic groups, such as the 13 ethnic groups in Finney and Simpson (2008), would be more difficult because of changes in the measurement and identification of ethnicity (Simpson and Akinwale, 2007; Stillwell and Duke-Williams, 2007). For example, a 'mixed ethnicity' classification was not included in the 1991 census. The same could be said for higher levels of geography. The local authority geography in England and Wales, for example, changed three times between 1991 and 2007 (Raymer and Giulietti, 2009). Since the primary aim of this paper is to illustrate the application of the combining data methodology, we focused on these more simple ethnic and regional groupings. The methodology described below can include much higher levels of disaggregation, but for analyses in England and Wales, it would require additional efforts to harmonise the Census and NHSCR data over time before combining them.

The sources of migration data used in this study are the 1991 and 2001 Censuses and annual published NHSCR tables from 1991 to 2007. The 1991 census tables were obtained from the Special Migration Statistics (SMS) dataset called 'SMSGAPS' available on the Centre for Interaction Data Estimation and Research (CIDER) website (http://cider.census.ac.uk/). The 2001 Census and annual NHSCR data were obtained from the Office for National Statistics (http://www.statistics.gov.uk/) by request.

[1] See http://www.statistics.gov.uk/geography/downloads/GB_GOR98_A4.pdf for a map of these regions.

3. A LOG-LINEAR MODEL FOR COMBINING DATA

3.1 Identifying Key Structures

In this paper, we denote cross-classified tables by letters. For example, OD is a two-way (origin by destination) table of migration flows, OAS is a three-way (origin by age by sex) table of migration flows and ODSE is a four-way (origin by destination by sex by ethnicity) table of migration flows.

Once the data were collected, the next step was to identify an overall model that could accurately predict the complete ODASE table of migration flows. This was undertaken by comparing various unsaturated log-linear model fits of two four-way migration flow tables, i.e., ODAS and ODSE, with the corresponding observed data, representing flows obtained from the 2001 census. The complete five-way table ODASE was not publicly available for disclosure reasons; however, we were able to determine that the missing age-ethnicity (i.e., AE) information was not required to accurately estimate the migration patterns by analysing the 2001 Samples of Anonymised Records (see below).

We use log-linear models to identify important structures in the migration flow tables. These models are widely used in the analysis of cross-classified data. The following sets out a brief explanation on how these models can be applied to identify key structures in migration flow tables. We use the 2001 ODAS Census table described above for illustration. For more detailed explanation of log-linear models, we refer the reader to Agresti (2002, 2007) and Fienberg (2007). Note, in this paper, when we refer to interactions, we specifically mean association terms in a log-linear model. Other migration researchers often refer to origin-destination migration flow data as 'interaction' data irrespective of whether there is a statistical association between origins and destinations.

A simple log-linear model to estimate the number of migrants, n_{ijxy}^{ODAS}, in the complete 2001 Census ODAS data set, from origin $i = 1, ..., 9$, destination $j = 1, ..., 9$, age group $x = 1, ..., 16$, and sex group $y = 1, 2$, is

$$\log \mu_{ijxy}^{ODAS} = \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \lambda_y^S, \tag{1}$$

where μ_{ijxy}^{ODAS} is the expected number of flows, λ is the constant parameter, and λ_i^O, λ_j^D, λ_x^A and λ_y^S are parameters describing the main effects of origin, destination, age and sex, respectively. This model assumes no interactions between origin, destination, age and sex. A model which includes interaction effects between origins and destinations, λ_{ij}^{OD}, for example, is

$$\log \mu_{ijxy}^{ODAS} = \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \lambda_y^S + \lambda_{ij}^{OD}. \tag{2}$$

Both Models (1) and (2) are 'unsaturated' log-linear models. A saturated model perfectly predicts the observed data and contains the same number of parameters as observations. For the ODAS table, this model is

$$\begin{aligned}\log \mu_{ijxy}^{ODAS} = {} & \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \lambda_y^S + \lambda_{ij}^{OD} + \lambda_{ix}^{OA} + \lambda_{iy}^{OS} + \lambda_{jx}^{DA} + \lambda_{jy}^{DS} + \lambda_{xy}^{AS} + \lambda_{ijx}^{ODA} + \lambda_{ijy}^{ODS} \\ & + \lambda_{ixy}^{OAS} + \lambda_{jxy}^{DAS} + \lambda_{ijxy}^{ODAS}.\end{aligned} \tag{3}$$

The key to the log-linear modelling strategy is to identify which of the above interaction terms are necessary for an accurate estimation of the migration flows. We do this by comparing various unsaturated models with the saturated one. All models are estimated using maximum likelihood under the assumption that the counts follow a Poisson distribution.

Unsaturated models can be compared with the saturated model to assess their goodness of fit. Traditionally, the likelihood ratio statistic (G^2) is compared with a chi-squared distribution with degrees of freedom equal to the residual degrees of freedom (df). However, this is not appropriate for tables with large cell counts, since most, if not all, interaction terms will be significant. Therefore, we divide G^2 by the residual degrees of freedom. This measure allows us to compare the models by controlling for their relative complexities, which is useful for identifying the best model in terms of overall fit and simplicity. The residual degrees of freedom represent the number of parameters 'not used' to predict the flows. To calculate the residual degrees of freedom, we simply subtract the number of parameters in the unsaturated model from the number of parameters in the saturated model.

The number of non-redundant parameters for a particular hierarchical log-linear model for the ODAS table can be calculated by summing the numbers of parameters in Table 1A corresponding to the terms in the model. The number of parameters to be estimated in Model (1) is 1 + 8 + 8 + 15 + 1 = 33, Model (2) is 33 + 55 = 88 and Model (3) is the sum of all the numbers (i.e., 2304). A hierarchical model implies that, if a particular interaction term is required in the model, then all its lower order terms must also be included. For example, since Model (2) contains λ_{ij}^{OD}, the terms λ, λ_i^O and λ_j^D are also required. Note, the migration tables analysed in this paper contain structural zeros on the diagonal elements of the OD partial tables (i.e., within region flows). The numbers of parameters in Table 1 have taken this into account.

---------- Table 1 about here ----------

In our analysis of the unsaturated log-linear models for the ODAS table (Table 2A), we find that the best models are Models 5 and 8. The two-way interactions between origin and sex (OS) and destination and sex (DS) and the three-way interactions between origin, destination and sex (ODS), origin, age and sex (OAS) and destination, age and sex (DAS) did not contribute substantially to the overall model fit. Out of the two best models, we prefer Model 5, which only includes the two-way interactions between origin and destination (OD), origin and age (OA), destination and age (DA) and age and sex (AS). The ODA term slightly improved the fit but at the expense of a large number of parameters. Also, the two-way interaction model produced estimates that were nearly indistinguishable from the observed values in the complete ODAS table. Our model preference also supports Raymer *et al.* (2006) and Raymer and Rogers (2007), who found that the three-way interaction term between origin, destination and age does not contribute much beyond the two-way interaction models, except in very specific origin-destination-specific flow cases, e.g., those with a very pronounced retirement peak. That is, most of the age patterns of origin-destination-specific migration are captured by the age patterns exhibited by the total in-migration and out-migration flows. For analyses with smaller geographic units (e.g., migration between counties or local authorities in England), we still believe that this assumption would hold, although there may be some exceptions.

---------- Table 2 about here ----------

For the analysis of the ODSE table (Table 2B), we find that the best models, penalised for complexity, is Model 8. The number of parameters in the models considered in this table can be calculated using Table 1B. Here, we do not rely on the simpler Model 4 because in the 2001 Census in England and Wales, Whites and ethnic minorities exhibited very different origin-destination-specific patterns of migration (Finney and Simpson, 2008; Hussain and Stillwell, 2008; Raymer and Giulietti, 2009). For example, Finney and Simpson (2008:81) found that "Minority ethnic groups moved less far than White groups even when people of similar characteristics are compared." Without the ODE term, these spatial differences would be ignored.

Finally, as mentioned at the beginning of this section, the complete five-way table ODASE was not publicly available from the 2001 Census for disclosure reasons. For our model, it is important to know whether the missing age-ethnicity (i.e., AE) interaction is required. To address this, we fitted log-linear models to 2001 Samples of Anonymised Records (SAR) data (available at http://www.ccsr.ac.uk/sars/). The SAR is a 3% sample of individuals in the 2001 Census, containing most of the information collected but with some variables coarsened (e.g., geography) to reduce disclosure risk. For the particular interest of this study, it contains age, sex and ethnic information on approximately 62 thousand interregional migrants in England. The analysis of age-specific ethnic migration from the SAR (at the national level) confirmed that the overall model does not require an interaction terms between age and ethnicity. This result is not particularly surprising given all the research on age profiles of migration and their regularities over time and across space (Rogers and Castro, 1981). Tobler (1995:335) even goes so far as saying that these regularities "surely warrant designation as a migration 'law'". Note that regularities in age profiles do not necessarily imply similar rates of age-specific migration for different population groups. For example, Finney and Simpson (2008) found that the rates of age-specific migration differed by ethnicity. However, their Figure 1 suggests that a model for rates would not require an age-by-ethnic interaction term since the shapes of the curves, except possibly for the last age group for Chinese migration, are very similar.

The above analyses provide us with some direction on how to proceed with the combining of migration flow data. First, we do not need to include the complete data to produce accurate results. In fact, this model has the advantage of producing smoother estimates, particularly across age groups. Second, to produce good results, we only need the OD, OA, DA and AS and ODE tables. The NHSCR provides the first four on an annual basis. The ODE table, on the other hand, is only available from censuses on a decennial basis.

3.2 Model Specification

Our objective for this project is to estimate migration flows for an ODASE table for each year from 1991 to 2007. The basic idea is to supplement information from the NHSCR with more detailed information from the censuses. The log-linear model with offset developed by Raymer *et al.* (2007) is used as a starting point (see below). Note, for the model developed in this paper, the diagonals of the OD partial tables are excluded.

Log-linear models can be considered a type of spatial interaction model, commonly used to model origin-destination-specific migration flow data (for overviews, see Fotheringham *et al.,* 2000a:211-235; Stillwell and Congdon, 1991; Stillwell, 2009). Applications of log-linear models to model migration flows, including the use of offsets, can be found in Willekens (1982, 1983, 1999). Raymer *et al.* (2007) extended Willekens's (1999) spatial interaction model for two-way tables to include a third variable of interest not available in the incomplete migration data. For example, an origin by destination by ethnicity table, with counts n_{ijz}^{ODE} can be modelled by using the following log-linear with offset form of the spatial interaction model:

$$\log \mu_{ijz}^{ODE} = \lambda + \lambda_i^O + \lambda_j^D + \log m_{ijz}^{ODE}, \tag{4}$$

where μ_{ijz}^{ODE} is the expected flows from origin *i* to destination *j* for level *z* of the third variable. The λ_i^O and λ_j^D parameters represent background factors related to the characteristics of the origin and destination, respectively. The log of m_{ijz}^{ODE} is the offset, a factor representing the auxiliary information on migration flows (see Knudsen, 1992 for other spatial analysis applications using offsets). This is additional data relating to migration between the same origins and destinations as in the incomplete data but is not a parameter in the model. Note, there are no parameters corresponding to the dimension indexed by z. Here, we rely on the auxiliary data to provide the missing margin and association structures not contained in the incomplete data.

If information on two-way or higher associations exists in the incomplete data, the model can be extended to include this. Furthermore, we may not wish to impose the higher-way interactions from the auxiliary data. For example, as discussed in Section 3.1, we wish to use the OD, OA, DA and AS tables from the NHSCR data and impose the three-way associations from the ODE census table. This is achieved by using the following log-linear model for n_{ijxyz}^{ODASE}, the counts in the five-way ODASE table:

$$\log \mu_{ijxyz}^{ODASE} = \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \lambda_y^S + \lambda_{ij}^{OD} + \lambda_{ix}^{OA} + \lambda_{jx}^{DA} + \lambda_{xy}^{AS} + \log m_{ijz}^{ODE} . \quad (5)$$

Should a different model for the flows be thought appropriate, then Model (5) can be modified by adding or removing interaction parameters, or by changing the offset term, provided the pertinent information is available in the incomplete or auxiliary data, respectively.

Models (4) and (5) can be fitted by using maximum likelihood estimation. It is straightforward to derive and solve, using an iterative procedure, the likelihood equations for these models to obtain estimates of the λ-parameters and flows. Raymer *et al.* (2007) did this for Model (4). However, since our interest is primarily in the estimation of the flows, we just apply an iterative proportional fitting (IPF) algorithm to obtain the maximum likelihood estimates of the flows directly instead. Agresti (2002, Section 8.7.2) presents an example of the use of IPF to fit a log-linear model without an offset to a three-way table and Willekens (1982, 1983) provides examples with offsets for two- and three-way tables. For examples of applying IPF to combine survey and census data for small area population estimation, see Simpson and Tranmer (2005).

The initial values in Model (5) are given by the counts in the ODE table from the census: $\mu_{ijxyz}^{ODASE(0)} = m_{ijz}^{ODE}$ for all x and y. They are then successively multiplied by adjustment factors so that the marginal tables match the counts in the NHSCR OD table, then the NHSCR OA table, then the NHSCR DA table and finally the NHSCR AS table. This is repeated until the marginal tables of estimated flows simultaneously match all of the counts contained in the four NHSCR tables. Furthermore, the resulting table has the same OE, DE and ODE association structures as the census table.

The algorithm to fit Model (5) requires consistency in the marginal distributions of the incomplete data, namely of the OD, OA, DA and AS tables. Ideally, these would have come from a single four-way table. However, when we extracted the one-way margins from the publicly available OD, OAS and DAS tables provided by ONS, they did not match because the OAS and DAS tables included migration to and from Wales, Scotland and Northern Ireland. Furthermore, the OD table included within region flows. To make these tables consistent, we used the following procedure. We started with the OD table and removed the diagonal elements and the rows and columns corresponding to areas in Wales, Scotland and Northern Ireland. We then scaled the AS table so that its total matched that of the OD table, with the assumption that the age and sex proportions of migration for United Kingdom are the same as those for England. From here, we used iterative proportional fitting to force the OA and DA tables to match the O, D and A margins from of the OD and AS tables. Hence, all four tables required for modelling had the same totals and one-way margins as required.

Raymer *et al.* (2007) assumed the three-way auxiliary interaction structure remained constant over time. We, on the other hand, allow this structure to vary over time from 1991 to 2007. We do this by geometrically interpolating the counts from 1992 to 2000 and by geometrically extrapolating from 2002 to 2007. The 1991 and 2001 census values are used as benchmarks. Model (5) is then run for each year with these auxiliary structures used as offsets.

Once the models were run, we then checked the results for their reasonableness. In doing so, we identified an important problem with the NHSCR data relating to the age structure of migration by sex (i.e., the AS table). Here, it was found that females had higher levels of migration (52.3 percent on average) than males (47.7 percent on average), with the gap between the two sexes slightly widening over time. The corresponding patterns obtained from the 1991 and 2001 censuses, however, showed a different pattern with males representing 50.8 percent in both years. The reason for this difference has primarily to do with males being less likely to register with the NHS register, particularly in their young adult years (see Fotheringham *et al.,* 2004:1637-1640 for discussion). Note, this was not an issue in Raymer *et al.* (2007) because they only examined migration patterns of elderly persons, a group less likely to be missed in a health service population register. The differences between males and females by age are shown in Figure 1 for the years 1991 and 2007. The corresponding patterns reported by the 1991 and 2001 censuses (not shown) are very different in that the age-sex patterns are nearly identical, except in the last age group of 75+ years, where females have higher levels of migration (associated with their higher population numbers in these years).

---------- Figure 1 about here ---------

As illustrated in Figure 1, nearly all the differences in the age patterns of male and female migration as reported in the NHS data occur in the 15-19 year, 20-24 year and 25-29 year age groups. (Note, Fotheringham *et al.* (2004) also found differences across spatial units. Our analysis at the regional level did not find any substantial differences.) To correct for the differences in the age-sex patterns, there are two options. The first is to impose the interactions contained in the census AS tables instead of the NHS tables. Here, Model (5) can be rewritten as follows:

$$\log \mu_{ijxyz}^{ODASE} = \lambda + \lambda_i^O + \lambda_j^D + \lambda_x^A + \lambda_{ij}^{OD} + \lambda_{ix}^{OA} + \lambda_{jx}^{DA} + \log\left(m_{ijz}^{ODE} m_{xy}^{AS}\right). \qquad (6)$$

This model maintains all of the above associations but with the age-sex structure from the censuses. The problem with this model, however, is that it does not correct for the undercounting of males. The overall levels of migration would remain the same, which means that the levels of female migration would have to be lowered to make the age-sex differences correspond with the census patterns. We, on the other hand, assume that females are counted accurately in the NHSCR data.

The second option is to weight the estimates from Model (5) to account for the age-sex differences. The weights represent ratios of female to male migration for the 15-19, 20-24 and 25-29 age groups, marginalising over origin, destination and ethnicity. This approach maintains all of the associations implied by Model (5). The weights applied to the male migrants in the three age groups are set out in Table 3, along with the resulting adjustment ratios for all males (12 to 15 percent increase) and males plus females (6 to 7 percent increase).

--------- Table 3 about here ----------

4. PATTERNS OF ETHNIC MIGRATION

In this section, the estimated interregional migration flows by age, sex and ethnicity are presented. These flows represent the re-weighted estimates from Model (5) discussed in the previous section. First, we describe the patterns over time and across space and then by age and sex.

4.1 Over Time

The overall levels of interregional migration in England increased slightly from just less than 900 thousand to around one million persons per year between 1991 and 2007. The vast majority of these flows were comprised of Whites, representing about 94 percent in 1991, 90 percent in 2001, and 85 percent in 2007 (based on geometric extrapolation). The increasing levels of South Asian, Black and Chinese & Other migration are clearly visible in Figure 2. Here, we see that the flows of all three groups increased substantially over time, from around 48 thousand in 1991 to around 156 thousand in 2007. The relative shares of ethnic minority migration, however, remained pretty much the same over time with South Asians representing around 45 percent, Blacks around 22 percent and Chinese & Other around 33 percent.

---------- Figure 2 about here ---------

4.2 Spatial Patterns

Two examples of origin-destination-specific flows are set out to illustrate the differences between White, South Asian, Black and Chinese & Other migration. These represent migration flows from London (Figure 3) and from the South East (Figure 4), the two largest sources of interregional migration (note, the y-axis scales are different for White migrants). For migration from London (Figure 3), the top two destinations for all ethnic groups are the South East and East of England, for which the levels have been increasing steadily over time. Interestingly, Black migrants have relatively the same migration levels going to both regions, whereas for the other ethnic groups, the South East is the preferred destination. Larger differences in the migration patterns appear when the third choice of destination is considered. For Whites, the South West comes third in terms of destination choice, whereas it is West Midlands for South

Asians and Blacks. There is not much difference in the remaining destination choices for the Chinese & Other ethnic group. For migration from South East (Figure 4), the top destination for all three ethnic minority groups is London. For White migration, the patterns are more spread out and relatively level over time. Here, the top three destinations are London, South West and East of England.

---------- Figures 3 and 4 about here ---------

Finally, we see that the relative positions of ethnic migration (Figure 2) and their destination choices from London (Figure 3) and South East (Figure 4) have remained fairly stable over time. The only noticeable crossovers in these patterns are found in the Black flows from London to the Southeast and East of England regions (Figure 3) and in the White flows from Southeast to the Southwest and London regions (Figure 4). This suggests that the spatial patterns of ethnic migration, controlling for the overall increases in the levels, have not changed substantially between 1991 and 2007. In fact, when we examined the 1991 and 2001 census data, we found that the only major differences in the proportions from and to all regions were a relative increase in the shares of Blacks migrating from London (33 percent to 47 percent) and fairly large decrease in the shares of Blacks and Chinese & Other migrating to London (30 percent to 25 percent and 28 percent to 24 percent, respectively).

4.3 By Age and Sex

Next, consider the estimated age- and sex-specific interregional migration flows for each ethnic group. For illustration purposes, we first compare the 1991 differences of South Asian and Black migration between the West Midlands and London, the South West and London, and the South East and London (Figure 5). Second, in Figure 6, we compare the age-specific predictions for a specific sex and ethnic group, i.e., female South Asians, over time for the same flows as in Figure 5. These two examples provide some insights into the levels of detail available in the synthetic database estimated by the combining data model.

---------- Figures 5 and 6 about here ---------

In Figure 5, we see that the adjustment factors have resulted in very similar age patterns for males and females. The only major difference exists in the last age group, where females are known to contain a much larger share of the population. Also, by design, the age patterns of all ethnic groups have the same origin-destination-specific shapes. Finally, the figures show:

(1) differences in age-specific levels, with the greatest differences occurring in the young adult age groups (see, e.g., Figures 5A and 5B);

(2) regularities are maintained, even for very small flows (Figures 5C and 5D); and

(3) different shapes for different flows, for example, narrow labour force peaks in Figure 5A versus wider labour force peaks in Figure 5E.

In Figure 6, the levels of migration by age for female South Asians between West Midland and London, South West and London, and South East and London are illustrated for 1991, 1999 and 2007. For all flows, the levels of migration have increased over time. In Figures 6A and 6B, we see that the shape of the labour force peak has changed over time by attracting relatively more 15-19 year olds, whereas the shape remained relatively constant in Figures 6E and 6F.

5. CONCLUSION

Population and migration analysts require detailed and up-to-date information to inform policy and planning. This information is often not readily available. To overcome this limitation, we have proposed a methodology for combining incomplete registration data with auxiliary census data to study detailed migration patterns over time. The methodology is useful to migration researchers and population planners who are interested in making the best use of the data that are available to them, whether it comes from registrations, censuses or surveys. The results represent enhanced migration data that can be used to study the evolution of patterns over time, as inputs into projections, or to identify the numbers of specific migrants moving between areas for a particular year and for a specific measurement of migration (e.g., migration events).

For illustration, we have focused on the modelling of interregional ethnic migration in England by combining data obtained from the 1991-2007 National Health Service registers with data obtained from the 1991 and 2001 Censuses. Future research could expand the model to produce estimates more relevant to local policy planners. This would include adding more ethnic groups and higher levels of geography, as well as considering other migrant groups (e.g., flows by education or economic activity). The methodology could also be applied to other situations in the world, where countries have similar migration data situations in the United Kingdom. One particular example of how this methodology could be applied to improve migration data comes from United States, where the long-form of the census questionnaire will no longer be included and migration data will instead come from the American Community Survey, a source that is inadequate for capturing detailed migration patterns at high levels of geography (e.g., between states). Here, the American Community Survey data could be combined with the more reliable, but incomplete, interstate migration data from the Internal Revenue Service.

The analysis of ethnic migration in this paper has demonstrated the type of results that can be obtained from an estimated time series of ethnic interregional migration flows by age and sex. The result is a time series of accurate migration flows with ethnic characteristics. Future work should examine the error in the ethnic dimensions in these flows and in the smoothing of the NHSCR data. Also, the inclusion of a third data source that would capture, for example, recent changes in the migration patterns not captured by the most recent census. For England, this would include the changes in ethnic internal migration patterns likely to be resulting from the large numbers of White immigrants from Eastern Europe since the European Union expansion in 2004. The combining data framework presented in this paper could be extended to include a third data source, such as the Labour Force Survey which does include more recent information on the marginal totals of ethnic migration.

REFERENCES

Agresti A. 2002. *Categorical Data Analysis*, 2nd edition. Wiley: Hoboken.

Agresti A. 2007. *An Introduction to Categorical Data Analysis*, 2nd edition. Wiley: Hoboken.

Bates J, Bracken I. 1982. Estimation of migration profiles in England and Wales. *Environment and Planning A* **14** : 889-900.

Bates J, Bracken I. 1987. Migration age profiles for local authority areas in England, 1971-1981. *Environment and Planning A* **19** : 521-535.

Bell M, Blake M, Boyle P, Duke-Williams O, Rees PH, Stillwell J, Hugo G. 2002. Cross-national comparison of internal migration: issues and measures. *Journal of the Royal Statistical Society Series A* **165** : 435-464.

Bell M, Rees P. 2006. Comparing migration in Britain and Australia: harmonisation through use of age-time plans. *Environment and Planning A* **38** : 959-988.

Boden P, Stillwell J, Rees P. (1992) How good are the NHSCR data? In *Migration Processes & Patterns: Population Redistribution in the United Kingdom*, Stillwell J, Rees P, Boden P (eds.); Belhaven Press: London; 13-27.

Champion T. 1996. Population review: (3) migration to, from and within the United Kingdom. *Population Trends* **83** : 5-16.

Dixon S. 2003. Migration within Britain for job reasons. *Labor Market Trends* **April** : 191-201.

Dorling D, Rees P. 2003. A nation still dividing: the British census and social polarisation 1971-2001. *Environment and Planning A* **35** : 1287-1313.

Faggian A, McCann P, Sheppard S. 2006. An analysis of ethnic differences in UK graduate migration behaviour. *Annals of Regional Science* **40** : 461-471.

Fienberg SE. 2007. *The Analysis of Cross-Classified Categorical Data*, 2nd Edition. Springer: New York.

Finney N, Simpson L. 2008. Internal migration and ethnic groups: evidence for Britain from the 2001 Census. *Population, Space and Place* **14** : 63-83.

Fotheringham AS, Brunsdon C, Charlton M. 2000a. *Quantitative Geography: Perspectives on Spatial Data Analysis*. Sage: London.

Fotheringham AS, Champion T, Wymer C, Coombes M. 2000b. Measuring destination attractivity: a migration example. *International Journal of Population Geography* **6** : 391-421.

Fotheringham AS, Rees P, Champion T, Kalogirou S, Tremayne AR. 2004. The development of a migration model for England and Wales: overview and modelling out-migration. *Environment and Planning A* **36** : 1633-1672.

Hatton TJ, Tani M. 2005. Immigration and inter-regional mobility in the U.K., 1982-2000. *The Economic Journal* **115** : F342-F358.

Hussain S, Stillwell J. 2008. Internal migration of ethnic groups in England and Wales by age and district type. Working Paper 08/3, School of Geography, University of Leeds.

Kalogirou S. 2005. Examining and presenting trends of internal migration flows within England and Wales. *Population, Space and Place* **11** : 283-297.

Knudsen DC. 1992. Generalizing Poisson regression: including apriori information using the method of offsets. *Professional Geographer* **44** : 202-208.

Long JF, Boertlein CG. 1990. Comparing migration measures having different intervals. Current Population Reports No. 166, U.S. Census Bureau, Washington, DC.

McCulloch A. 2007. The changing structure of ethnic diversity and segregation in England, 1991-2001. *Environment and Planning A* **39** : 909-927.

Morrison PA, Bryan TM, Swanson DA. 2004. Internal migration and short-distance mobility. In *The Methods and Materials of Demography*, Siegel JS, Swanson DA (eds.); Elsevier Academic Press: San Diego; 493-521.

ONS Migration Statistics Unit. 2002. Using patient registers to estimate internal migration: customer guidance notes. Office for National Statistics. Available at: http://www.statistics.gov.uk/statbase/expodata/commentary/WebGuidanceNotes.htm.

Raymer J, Abel G, Smith PWF. 2007. Combining census and registration data to estimate detailed elderly migration flows in England and Wales. *Journal of the Royal Statistical Society A* **170** : 891-908.

Raymer J, Bonaguidi A, Valentini A. 2006. Describing and projecting the age and spatial structures of interregional migration in Italy. *Population, Space and Place* **12** : 371-388.

Raymer J, Giulietti C. 2009. Ethnic migration between area groups in England and Wales. *Area* 1-17. DOI:10.1111/j.1475-4762.2009.00884.x.

Raymer J, Rogers A. 2007. Using age and spatial flow structures in the indirect estimation of migration streams. *Demography* **44** : 199-223.

Rees P, Butt F. 2004. Ethnic change and diversity in England, 1981-2001. *Area* **36** : 174-186.

Rees P, Willekens F. 1986. Data and accounts. In *Migration and Settlement: A Multiregional Comparative Study*, Rogers A, Willekens FJ (eds.); D. Reidel: Dordrecht; 19-58.

Robinson V. 1993. Making waves? The contribution of ethnic minorities to local demography. In *Population Matters: The Local Dimension*, Champion T (ed.); Paul Chapman: London; 150-169.

Rogers A, Castro LJ. 1981. Model migration schedules. Research Report 81-30, International Institute for Applied Systems Analysis, Laxenburg, Austria. Available at: http://www.iiasa.ac.at/Admin/PUB/Documents/RR-81-030.pdf.

Rogers A, Raymer J, Newbold KB. 2003. Reconciling and translating migration data collected over time intervals of differing widths. *The Annals of Regional Science* **37** : 581-601.

Rogerson PA. 1990. Migration analysis using data with time intervals of differing widths. *Papers of the Regional Science Association* **68** : 97-106.

Simpson L, Akinwale B. 2007. Quantifying stability and change in ethnic group. *Journal of Official Statistics* **23** : 185-208.

Simpson L, Finney N. 2009. Spatial patterns of internal migration: evidence for ethnic groups in Britain. *Population, Space and Place* **15** : 37-56.

Simpson L, Tranmer M. 2005. Combining sample and census data in small area estimates: iterative proportional fitting with standard software. *Professional Geographer* **57** : 222-234.

Stillwell J. 1994. Monitoring intercensal migration in the United Kingdom. *Environment and Planning A* **26** : 1711-1730.

Stillwell J. 2009. Inter-regional migration modelling: a review. In *Migration and Human Capital: Regional and Global Perspectives*, Poot J, Waldorf B, van Wissen L (eds.); Edward Elgar: Cheltenham; Chapter 2.

Stillwell J, Congdon P, eds. 1991. *Migration Models: Macro and Micro Approaches*. Belhaven Press: London.

Stillwell J, Duke-Williams O. 2005. Ethnic population distribution, immigration and internal migration in Britain: What evidence of linkage at the district scale? Paper presented at the

British Society for Population Studies Annual Conference, University of Kent, Canterbury, 12-14 September.

Stillwell J, Duke-Williams O. 2007. Understanding the 2001 UK census migration and commuting data: the effect of small cell adjustment and problems of comparison with 1991. *Journal of the Royal Statistical Society Series A* **170** : 1-21.

Stillwell J, Hussain S, Norman P. 2008. The internal migration propensities and net migration patterns of ethnic groups in Britain. *Migration Letters* **5** : 135-150.

Tobler W. 1995. Migration: Ravenstein, Thornthwaite, and beyond. *Urban Geography* **16** : 327-343.

United Nations. 1992. Preparing migration data for subnational population projections. Department of International Economic and Social Affairs, New York.

Willekens F. 1982. Multidimensional population analysis with incomplete data. In *Multidimensional Mathematical Demography*, Land KC, Rogers A (eds.); Academic Press: New York; 43-111.

Willekens F. 1983. Log-linear modelling of spatial interaction. *Papers of the Regional Science Association* **52** : 187-205.

Willekens F. 1999. Modeling approaches to the indirect estimation of migration flows: from entropy to EM. *Mathematical Population Studies* **7** : 239-278.

Table 1. Number of non-redundant parameters for terms in a log-linear model: ODAS and ODSE tables with 9 origins and destinations, 16 age groups, 2 sexes and 4 ethnic groups

A. ODAS Table

Term	λ	λ_i^O	λ_j^D	λ_x^A	λ_y^S	λ_{ij}^{OD}	λ_{ix}^{OA}	λ_{iy}^{OS}	λ_{jx}^{DA}	λ_{jy}^{DS}	λ_{xy}^{AS}	λ_{ijx}^{ODA}	λ_{ijy}^{ODS}	λ_{ixy}^{OAS}	λ_{jxy}^{DAS}	λ_{ijxy}^{ODAS}
Number of Parameters	1	8	8	15	1	55	120	8	120	8	15	825	55	120	120	825

B. ODSE Table

Term	λ	λ_i^O	λ_j^D	λ_y^S	λ_z^E	λ_{ij}^{OD}	λ_{iy}^{OS}	λ_{iz}^{OE}	λ_{jy}^{DS}	λ_{jz}^{DE}	λ_{yz}^{SE}	λ_{ijy}^{ODS}	λ_{ijz}^{ODE}	λ_{iyz}^{OSE}	λ_{jyz}^{DSE}	λ_{ijyz}^{ODSE}
Number of Parameters	1	8	8	1	3	55	8	24	8	24	3	55	165	24	24	165

Table 2. Selected unsaturated log-linear model fits of ODAS and ODSE 2001 Census tables

	Model	G^2	Parameters Required	G^2 / residual df
A. ODAS Table				
0	ODAS	0	2,304	0
1	O, D, A, S	288,138	33	127
2	OD, A, S	97,009	88	44
3	OD, OA, DA, S	17,308	328	9
4	OD, OS, DS, A	96,765	104	44
5	OD,OA,DA,AS	12,287	343	6
6	OD, OA, OS, DA, DS, AS	11,932	359	6
7	ODA, S	9,458	1,153	8
8	ODA, AS	4,436	1,168	4
9	ODS, A	96,555	159	45
10	ODS, AS	91,534	174	43
11	ODA, ODS	9,003	1,224	8
B. ODSE Table				
0	ODSE	0	576	0
1	O, D, S, E	226,535	21	408
2	OD, S, E	36,584	76	73
3	OD, OS, DS, E	36,318	92	75
4	OD, OE, DE, S	3,834	124	8
5	OD, OS, OE, DS, DE, SE	3,468	143	8
6	ODS, E	36,122	147	84
7	ODS, SE	36,032	150	85
8	ODE, S	1,453	289	5
9	ODE, SE	1,363	292	5
10	ODS, ODE	991	360	5

Notes: (1) G^2 = likelihood ratio statistic; (2) residual degrees of freedom (df) = number of parameters in the saturated model less the number of parameters in the unsaturated model, which can be calculated using the numbers in Table 1.

Table 3. Adjustment ratios for NHSCR migration data, 1991-2007

	Age Group (Males)			All	Both
Year	15-19	20-24	25-29	Males	Sexes
1991	1.315	1.436	1.131	1.123	1.059
1992	1.287	1.411	1.128	1.117	1.056
1993	1.300	1.409	1.131	1.120	1.058
1994	1.294	1.398	1.136	1.116	1.056
1995	1.322	1.385	1.121	1.117	1.056
1996	1.334	1.403	1.130	1.122	1.058
1997	1.339	1.408	1.136	1.119	1.057
1998	1.348	1.441	1.156	1.125	1.059
1999	1.344	1.433	1.157	1.122	1.058
2000	1.343	1.454	1.176	1.127	1.060
2001	1.388	1.473	1.161	1.131	1.062
2002	1.393	1.482	1.177	1.130	1.062
2003	1.399	1.487	1.188	1.132	1.063
2004	1.400	1.514	1.217	1.137	1.065
2005	1.390	1.490	1.223	1.138	1.066
2006	1.399	1.523	1.255	1.145	1.069
2007	1.400	1.553	1.281	1.148	1.070

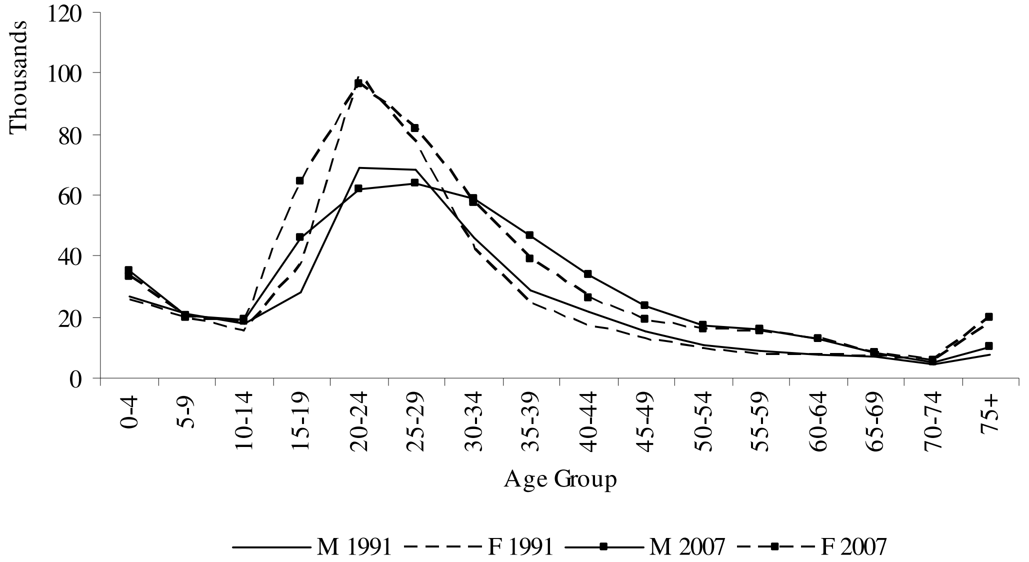

Figure 1. Age patterns of NHSCR interregional migration in England by sex, 1991 and 2007

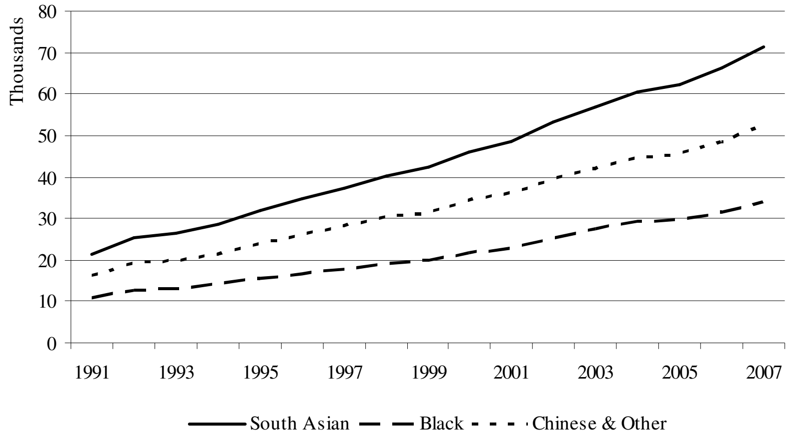

Figure 2. The levels of South Asian, Black and Chinese & Other interregional migration in England, 1991-2007

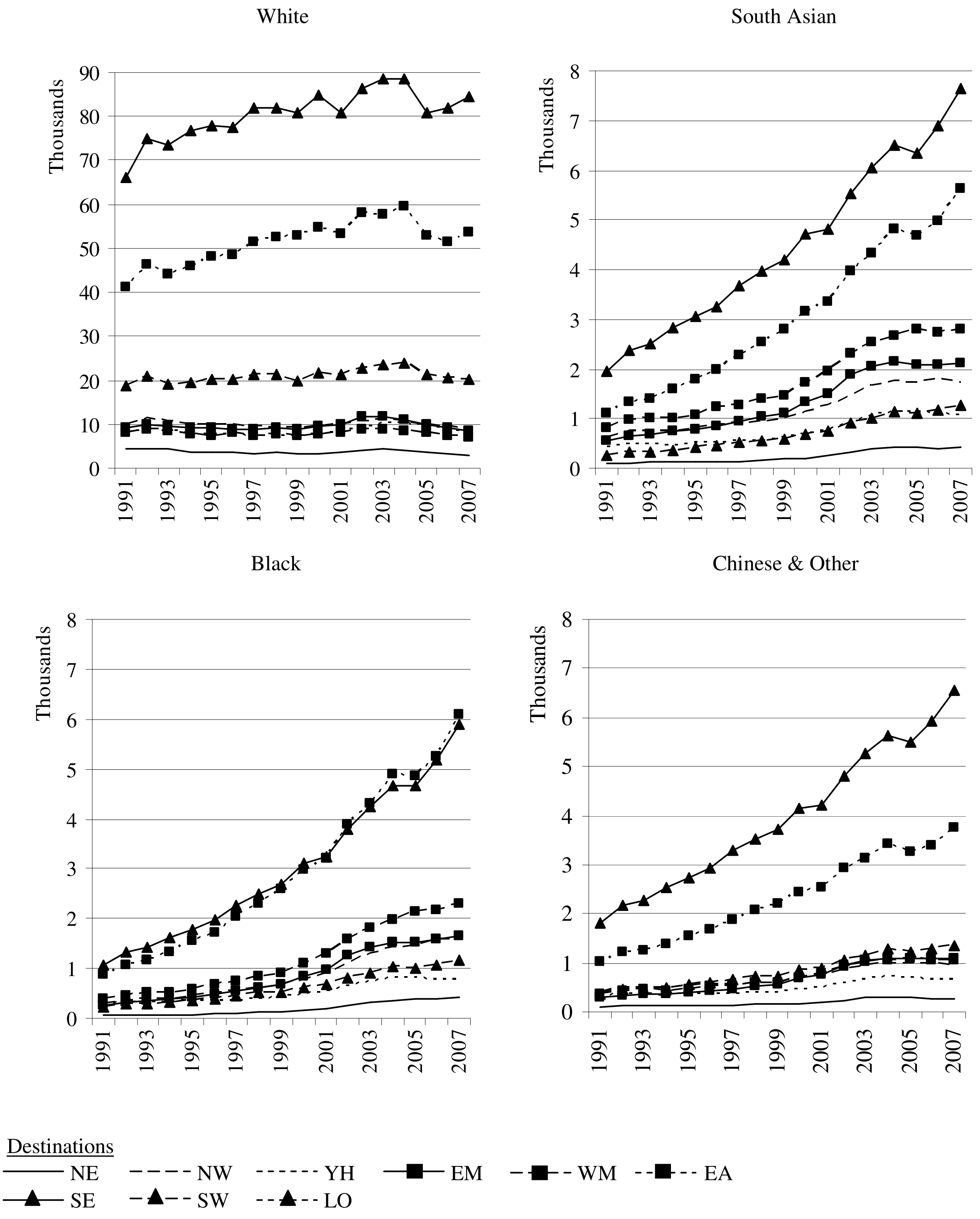

Figure 3. Interregional migration from London by ethnicity, 1991-2007

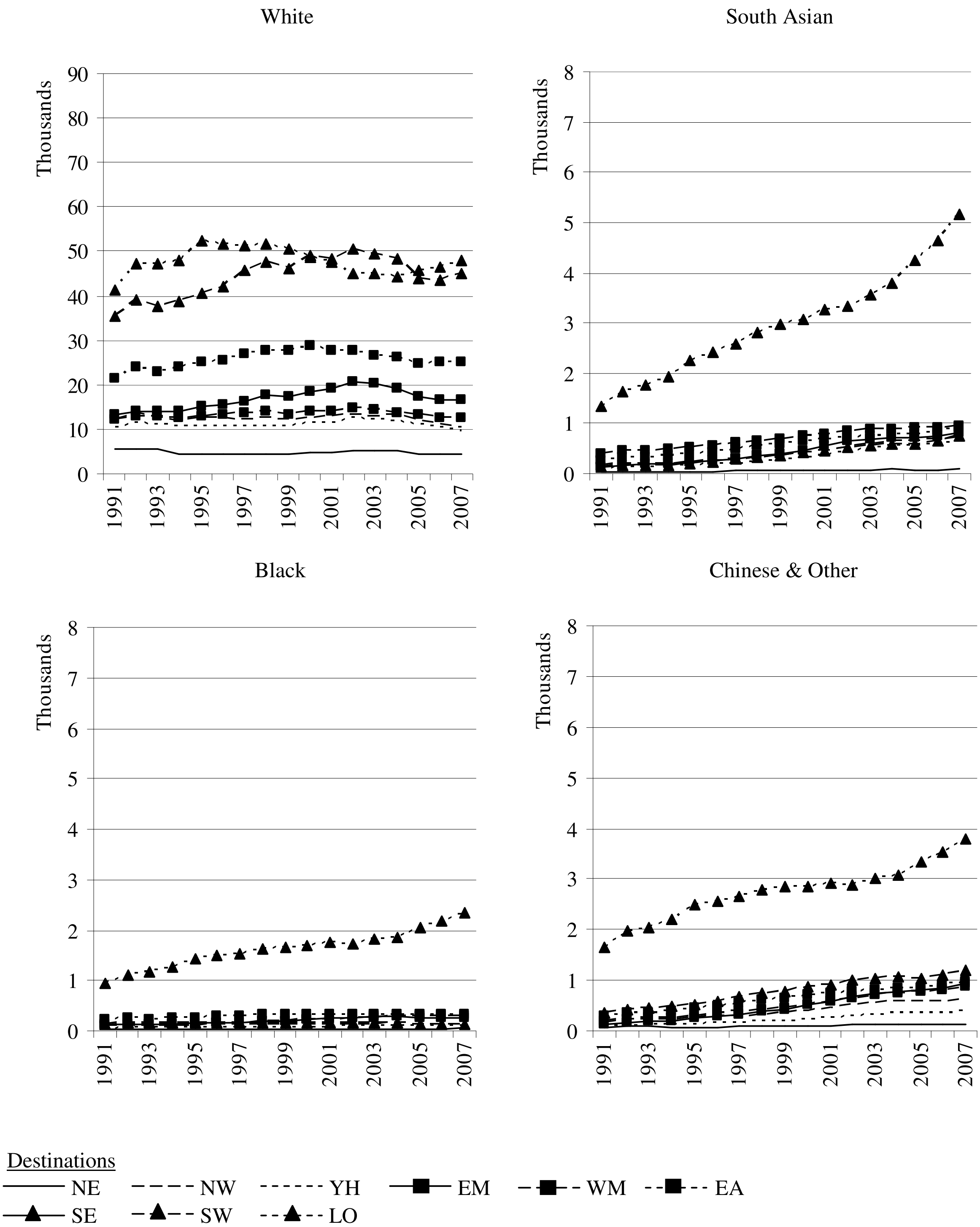

Figure 4. Interregional migration from South East by ethnicity, 1991-2007

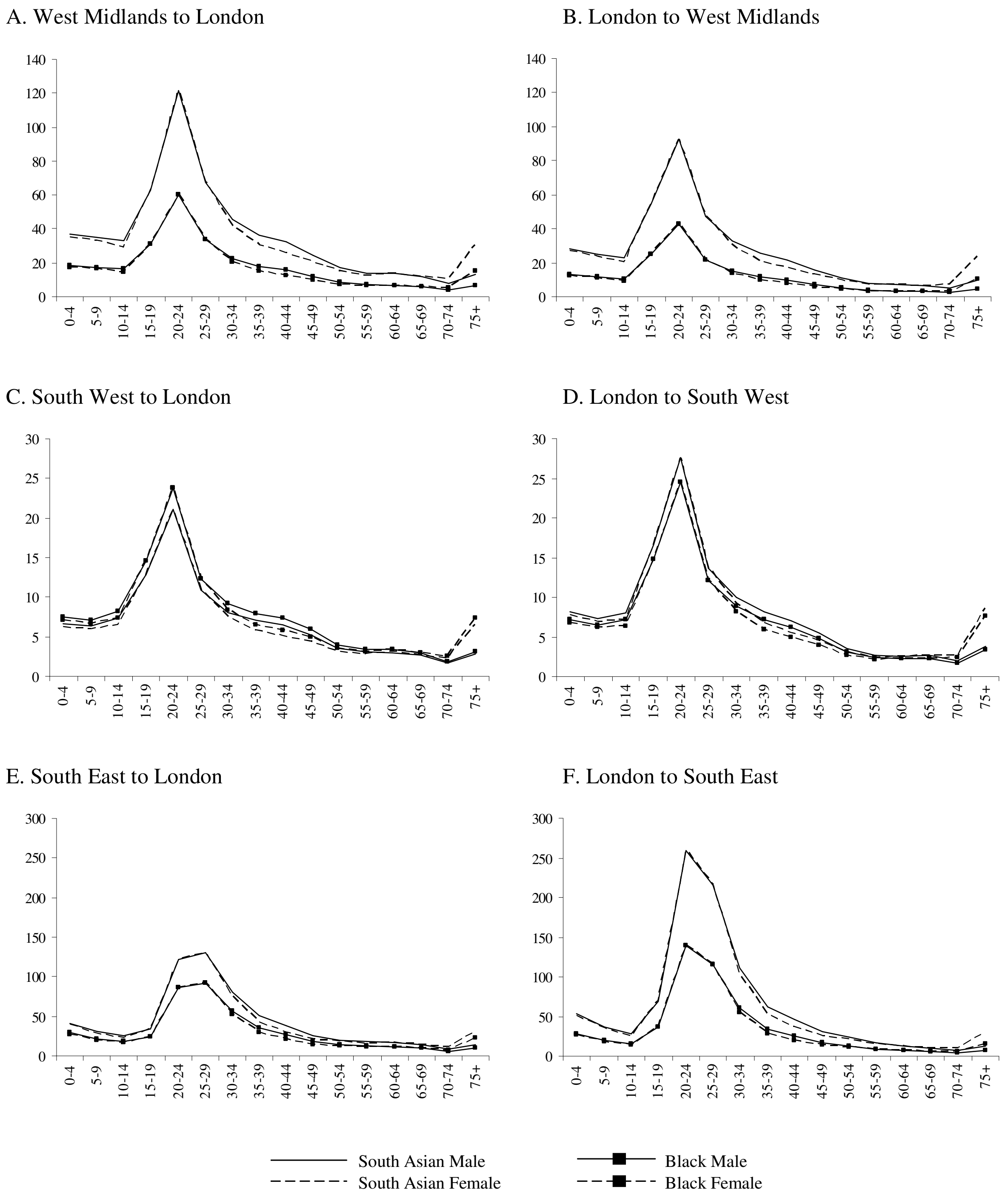

Figure 5. Age- and sex-specific migration of South Asians and Blacks between West Midlands and London, South West and London, and South East and London, 1991

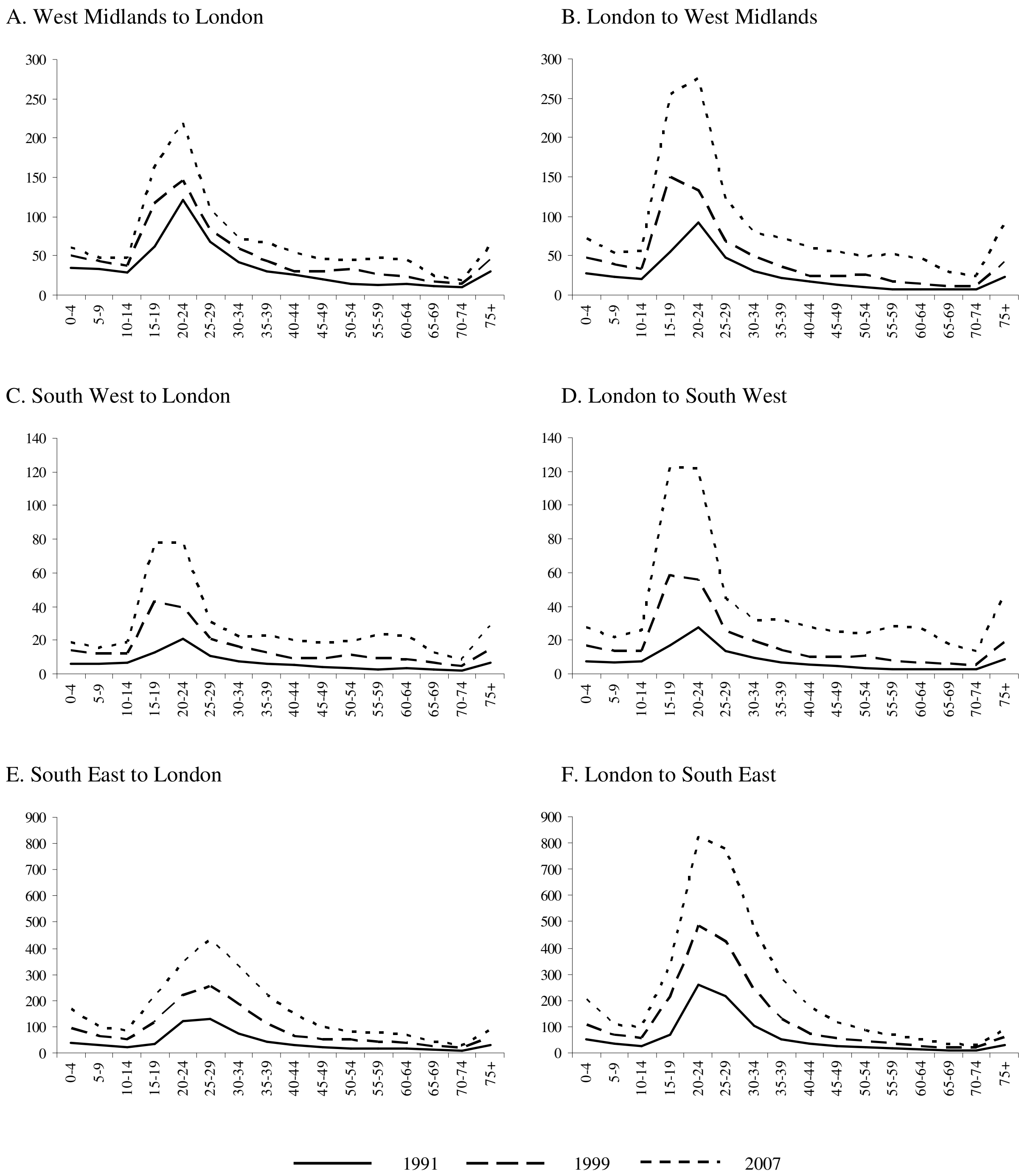

Figure 6. Age-specific migration of female South Asians between West Midlands and London, South West and London, and South East and London: 1991, 1999 and 2007