In search of a modelling strategy for projecting internal migration in European countries. Demographic versus economic-geographical approaches.

Leo van Wissen[1], Nicole van der Gaag[2],
Phil Rees[3] and John Stillwell[3]

[1] Population Research Centre, University of Groningen, PO Box 800, 9700 AV Groningen, email l.j.g.van.wissen@rug.nl
[2] Netherlands Interdisciplinary Demographic Institute NIDI, PO Box 11650, 2502 AR The Hague, email: gaag@nidi.nl
[3] School of Geography, University of Leeds, Leeds LS29JT, UK, emails: J.C.H.Stillwell@leeds.ac.uk, and P.H.Rees@leeds.ac.uk

Abstract

Internal migration is the most volatile and difficult to predict component of regional demographic change. A pure demographic approach using age and sex-specific parameters of migration intensities cannot fully capture the migration trends over time. One of the approaches that can be used for a better description of past trends and forecasting of future trends is to use additional non-demographic information such as regional economic indicators. In this paper we compare the predictive performance of pure demographic and extended economic-geographical models using data of four European countries at the so-called NUTS 2 level. The models are nested within a GLM specification that allows both demographic and extended models to be written as specific cases of log-linear models. Therefore model fit and performance can be compared directly.

Keywords: Internal migration, Poisson regression models, subnational population projections

Acknowledgments
This paper is based on the report: Study of Past and Future Interregional Migration Trends and Patterns within European Union Countries. In Search of a Generally Applicable Explanatory Model**.** Report on behalf of **Eurostat** (Invitation to tender no: 2002/S 67-052015/EN; Lot 5), by Nicole van der Gaag, Leo van Wissen, Phil Rees, John Stillwell and Marek Kupiszewski.

1 Introduction

In subnational population projections, internal migration is the most volatile and difficult to project component of regional population change. Population migration involves the relocation of individuals between geographical locations. It is a complex phenomenon, not only because of the complexity of spatial patterns of movement that are involved, but because of the myriad of motivations that influence the size and the composition of the flows between any two discrete areas. Moreover, the imprecision of the data that are used to analyse spatial patterns of population relocation is a severe obstacle for scientific advances in this field. The interest of demographers in this field is also quite volatile. Subnational population projections received ample attention in the seventies and eighties as a consequence of the development of multi-state demography, in which several interacting populations may be projected simultaneously (Rogers, 1966, 1967, 1968, 1973; Rogers and Willekens, 1986; Willekens and Drewe, 1984; Rees and Wilson, 1973, 1975, 1977). In the last decades the attention has waned however, and this may have to do with some of the practical difficulties in implementing the methodology. The multi-regional demographic model is a perfect accounting framework in which all demographic components of subnational population change are integrated consistently. The problem with this model is that it demands huge amounts of data and parameters: for instance, every origin-destination flow is decomposed in age and sex categories. Many countries lack the data to fill this matrix. For a system with 10 regions, and 20 age groups, the migration component is driven in theory by 1800 distinct parameters, and for projection purposes, in principle assumptions about the future time path of each parameters are required. The question how to deal with this problem leads to two types of approaches, which may be labelled demographic and explanatory. The demographic approach tries to find a structure in the parameters, which may be used for modelling future migration patterns. For instance, by studying the age patterns of regional outmigration rates over time, it may be concluded that this age pattern is relatively stable over time, or is structurally different between groups of regions, but stable within each group. Baydar (1983) studied the intertemporal stability of migration intensities in the Netherlands. Van Imhoff et al. (1997) developed a methodology to detect structure in interregional migration patterns by age and sex over time, which was subsequently used to formulate a projection model of internal migration (van der Gaag et al., 2000). The other approach follows the route of explanation in order to detect structure in the migration process. Push and pull factors of migration explain why certain regions have a higher outmigration rate or inmigration share, and spatial interaction factors such as distance explain why certain destination regions are more attractive for specific origin regions than others. This approach has its roots in the gravity models and spatial interaction literature in which the migration process is decomposed into factors related to the region of origin, factors related to the region of destination, and an interaction component, which is usually a distance decay factor (Wilson, 1970). A recent state-of-the-art example of this approach is the MIGMOD project, where a spatial interaction model has been developed for internal migration in the UK, which includes a large set of policy-sensitive variables (Champion et al., 2002, Fotheringham et al., 2004, Rees et al., 2004).

The demographic approach is aimed at projection making as such, whereas the explanatory approach first primarily answers the *why*-question about the migration flows. Explanatory models may under certain conditions also be used in projecting into the future, although this usually is in the form of scenarios. So, both approaches have partly similar aims, but different routes to achieve this. Methodological advances in the last decades have also shown that both approaches may use the same type of methodology, which is the Poisson regression model

(Willekens, 1980, 1983, Scholten and van Wissen, 1985; van Wissen and Rima, 1988; Congdon, 1991; Flowerdew, 1991; Flowerdew and Lovett, 1988). Therefore, the choice whether to choose a pure demographic approach or an explanatory approach is a matter of model specification. This also opens the possibility to specify a mixed model, whereby some elements in the model are treated using a pure demographic approach, and others specified as an explanatory submodel.

In this paper we will compare and evaluate demographic and explanatory approaches in projecting internal migration in a number of European countries. The motivation for this comparison derives from a European project, sponsored by Eurostat[1] in preparation of new subnational population scenarios for all countries of the European Union, at the so-called NUTS 2 regional level (European Communities, 1999). Since the late eighties subnational population scenarios have been carried out on behalf of the European Commission, and these projects were evaluated by Rees et al. (2001). One of their key conclusions was that previous attempts in projecting internal migration were too mechanical and did not take into account country- and region-specific information that may be important. In other words, their recommendation was to move away from a pure demographic approach to an explanatory approach. Moreover, since migration motives differ according to the stage in the life cycle, the model should be segmented accordingly. This makes it possible to specify submodels for student migration, labour market migration, family migration, retirement migration, and so on. Before deciding on using such an explanatory approach for inter-NUTS 2 migration, its feasibility should be investigated in some detail. Therefore, four countries were chosen to compare and evaluate both types of models: the Netherlands, Sweden and the UK[2]. It is also important to note that the current practice in European countries reflects the broad array of possibilities between pure demographic and explanatory approaches (van Imhoff et al., 1994; van der Gaag et al., 1997, 2003). Some countries use the simplest of demographic approaches, which involves only net migration totals, whereas other countries, such as the Netherlands, use a complex model with housing and labour market variables, and specific modules for students and other special groups. An important requirement for the models to be used by the European Commission was that one common methodology should be used in all countries. This does not necessarily mean that 'one size fits all', which was one of the main points of criticism of the Rees et al. evaluation report, but that within one approach simpler and more complicated variants should be possible, depending on the specific circumstances and data availability of each country.

This paper is organized as follows. In the next section the data are discussed. In section 3 the methodology is presented. Section 4 gives the results, and the final section concludes.

2 The data

The variety of regional classifications across Europe makes it very difficult to compare migration levels and patterns between countries. Each country has its unique set of sub-national areas. Regions can differ significantly both in terms of size and structural characteristics and those differences may have implications for the measurement of migration. Large regions, for example, may subsume within their boundaries as intra-regional migrants many of the flows that might be inter-regional if the regions were smaller. Thus, for countries with only a relatively small number of large regions, a lower rate of inter-regional migration

[1] Eurostat Invitation to tender no: 2002/S 67-052015/EN; Lot 5
[2] In the project Spain was also taken into account, but due to limited data availability for this country it is left out of the evaluation in the current paper

may be expected in comparison with countries that have a large number of small regions. It is appropriate to recognize that the huge variation in population size across and within countries is likely to have a significant impact on the modelling outcomes, and that this factor alone is sufficient argument to move away from the 'one size fits all' approach. Figure 1 shows the NUTS 2 regions in each of the four countries, as well as the main inter-NUTS 2 migration linkages. The number of regions is 8 in Sweden, 12 in the Netherlands, and 32 in the UK.

The source of the migration data in each country is different. The Netherlands and Sweden have a complete population register, whereas in the UK use a register of patients in the National Health System (NHS) re-registering with doctors in different regions. Migration patterns are broken down into 6 age groups, and in Sweden and the Netherlands also by sex. A longitudinal approach is possible due to the availability of 8 to 9 years of observation.

Based on existing migration theory, hypotheses were formulated about the main driving factors of migration flows, and from these hypotheses, a set of variables was collected, mostly from Eurostat's REGION database. The distance variable was not available from REGIO or from national statistical offices. In the case of the UK, the matrix of distances (in kilometres) between NUTS 2 areas was built up as an average of distances between smaller spatial units for which information was available using the database of the MIGMOD project. For Sweden and the Netherlands, straight-line distances were calculated between the centres of NUTS 2 regions. The contiguity matrices of ones (indicating contiguity) and zeros (indicating no contiguity) were produced manually. Table 1 gives an overview of all variables used.

Table 1: The variables selected for modelling

Demographic	Economic	Other
Population	GDP at ppp per inhabitant**	Distance
Density	Unemployment rate**	Contiguity
Immigration	Employment*	
Accessibility	Housing stock	

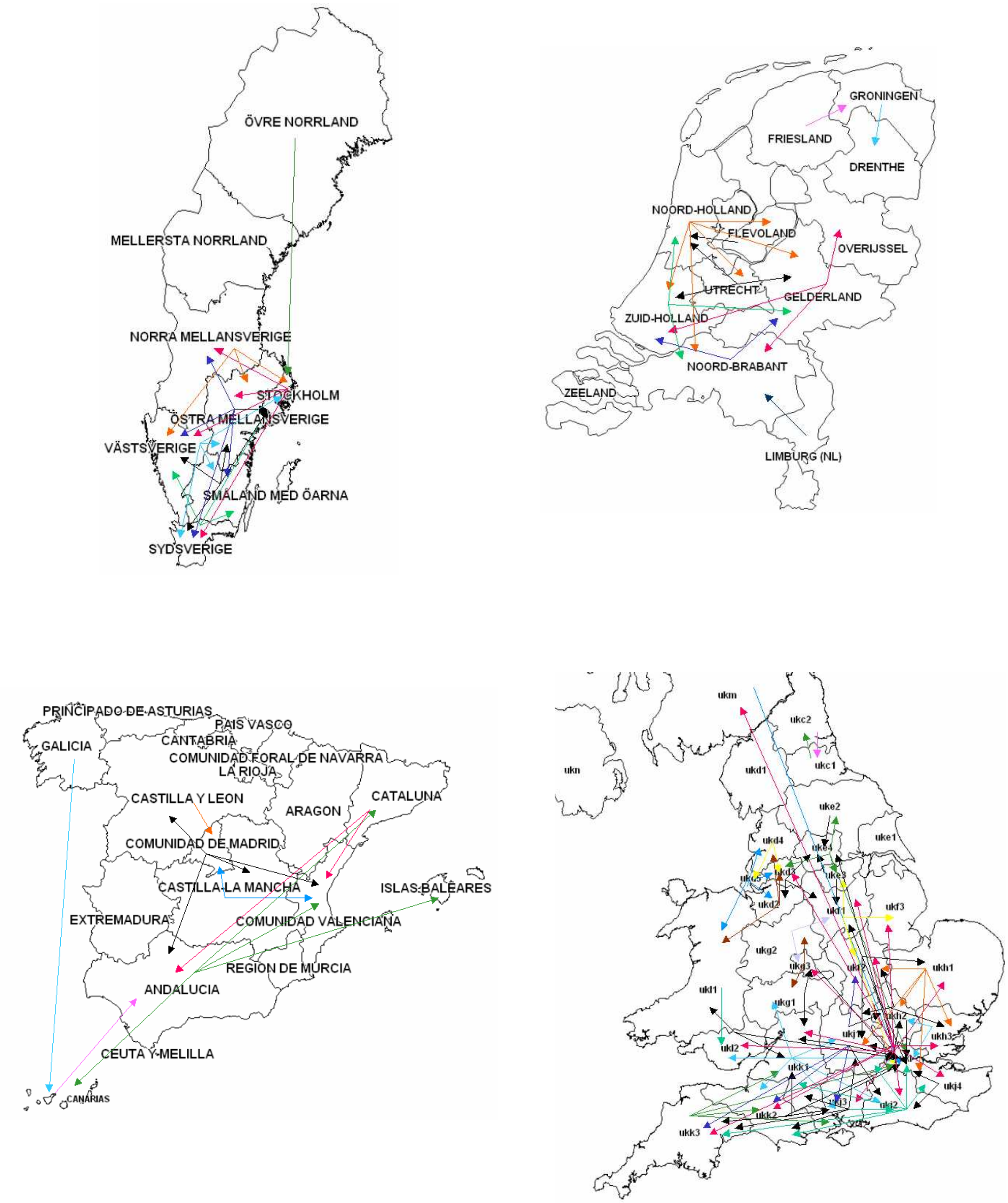
* Change variable computed

** Lagged and change variables computed

.

Each of the variables identified in this table is assumed to have an influence on migration. Hypotheses can be formulated that specify the nature of the relationships in more detail. However, we must be aware that relationships between migration and explanatory variables tend to be *time and scale specific*. In addition, relationships will vary by *age* and *other selective influences*, e.g. a large population size may be important for young workers but not necessarily for young students. Moreover, we should not be surprised to find that *signs* of regression model parameters as well as their *significance*, will vary from place to place.

Figure 1: Main origin-destination flows (> 3000 persons) in each of the case study countries in 1998 between NUTS 2 regions (source: van der Gaag et al., 2003)



3 Model specification

As stated in the introduction, the Poisson regression model is the vehicle that allows the specification of demographic as well as explanatory migration models. Following standard practice in migration modelling, we model the internal migration process in two steps: first the out-migration rates, and second, conditional on out-migration, the probabilities of in-migrating to destinations:

$$m_{ij}^{as}(t) = m_i^{as}(t)\, p_{j|i}^{as}(t) \qquad (1)$$

All rates *m* and probabilities *p* are age- and sex-specific (*a*,*s*). $m_{ij}(t)$ is the rate of migration from region *i* to region *j*, $m_i(t)$ is the total rate of out-migration from region *i*, and $p_{j|i}(t)$ is the conditional probability of choosing destination region *j* after out-migration from region *i*. We construct models separately for $m_i^{as}(t)$ and $p_{j|i}^{as}(t)$.

Both the outmigration model and the destination choice model were estimated on migration data of the Netherlands, the UK and Sweden. First, using only data pertaining to the first half of the period, a pure demographic model was estimated, and next explanatory models were tried, as well as mixed forms. A purely demographic model is usually a good description of reality, but contains no causal "drives" that might produce change over time in the migration process. Comparing the goodness of fit of these two models for each country gives insight into the relative descriptive or explanatory power of non-demographic variables *vis-à-vis* a pure demographic approach. The best demographic and explanatory models, in terms of (a) model fit, (b) parsimony, and (c) similarity in specification with the models for the other countries, were then used to for prediction. This prediction phase is a form of external validation of the models. We used observed values of exogenous variables to predict migration outcomes, thus leaving aside for the moment the (important) question how to predict these exogenous variables. Of course, three years is a short period for prediction, but the available time series do not allow longer time periods for validation. The next section presents the main results, both of the model fitting as well as the prediction stage for the outmigration model as well as the destination choice model.

Models for outmigration

A Poisson regression model for the outmigration rates has the form:

$$\ln O_i^{as}(t) = \ln P_i^{as}(t) + \mu_i^{REG} + \mu_a^{AGE} + \mu_s^{SEX} + \mu_t^{TIME} + \mu_{as}^{AGE*SEX} + \mu_{ia}^{REG*AGE} + \mu_{it}^{REG*TIME} + \ldots \qquad (2)$$

The dependent variable $\ln O_i^{as}(t)$ is the expected number of outmigrants out of region *i*, for age-sex group (*a*,*s*) at time *t.* The $\ln P_i^{as}(t)$ is the log of the population size of region *i*, age group *a*¸ and sex *s* at time *t.* It is called an *offset* since it has a fixed parameter value of 1. Subtracting the offset from both sides of the equation gives:

$$\ln m_i^{as}(t) = \ln\left(\frac{O_i^{as}(t)}{P_i^{as}(t)}\right) = \mu_i^{REG} + a_a^{AGE} + \ldots,$$

which is a log-rate model. The parameters μ pertain to the categories of each of the dimensions in the multidimensional contingency table of the migration matrix. The main effects $\mu_i^{REG}, \mu_a^{AGE}, \mu_s^{SEX}$ and μ_t^{TIME} relate to the marginal totals of each of the variables. If there was no interaction between each of the variables, the full migration table could be described by a model with only main effects. Each cell entry would be the product of the marginal probabilities of the contributing variables. Of course in reality there are strong interaction effects between the variables. The age distribution of outmigration is different for both sexes, which makes the inclusion of the interaction parameter $\mu_{as}^{AGE*SEX}$ necessary. Moreover, age profiles of outmigrants are likely to be different across regions, which necessitates the inclusion of a separate interaction term $\mu_{ia}^{REG*AGE}$. Other interaction terms may be necessary as well, and the method to decide which interaction term is necessary is through a statistical analysis of the multidimensional migration matrix over time. The model gives parameter estimates, standard errors of the estimates, and an overall fit measure, the Deviance *D*, which may be used to judge whether the inclusion of one parameter gives a significant improvement of fit (the difference in Deviance between two nested models is approximately Chi^2 distributed with degrees of freedom equal to the difference in the number of free parameters between both models. The number of interaction terms with 4 variables (region, age, sex and time) is substantial: there are 6 two-way interaction terms, 4 three-way interaction terms, and 1 four-way interaction term. Shorthand notation of the different possible models specifications is as follows:

- a model with only main effects is: LPOP + A + S + R + T, where LPOP is the log of the population, or offset
- a model including all two-way interaction effects is: LPOP+AS+AR+AT+SR+ST+RT (the inclusion of a higher order interaction term implies the inclusion of the lower order interaction terms and the main effects, therefore the main effects A+S+R+T are not specified here, but they are included by definition)
- a model with all two-way interaction effects plus one three-way effect is: LPOP+AT+ST+ RT+ASR (note that only the non-implied two-way interaction effects are specified separately here). This model specifies that in addition to the two-way interaction effects the age- and sex profiles of outmigrants are different over the regions.
- a model with the four-way interaction term is simply denoted as LPOP+ASRT, but it contains as many parameters as there are cells in the longitudinal migration table. This is the saturated model, and if the four-way interaction term is significant, we cannot simplify the table by leaving out specific interactions, without loosing significant information.

These log-linear models, or their multiplicative equivalent, in which multiplicative terms $\pi = \exp \mu$ replace the additive terms in the log scale, are purely descriptive. They describe the structure among the dimensions age, sex, region and time. For projection purposes, a time-invariant model structure is preferred over a time-varying structure. Time-invariant structures are those specifications that do not involve T-terms. For instance, the model LPOP+ASR does not contain a time dimension, and this implies that there are region-specific age-and sex-specific outmigration profiles, but that these profiles are time-invariant. Likewise, the model LPOP+ASR+T implies that there is a generic time effect (i.e. a time-specific scaling factor) but the overall ASR effect is stable over time. For projection purposes, the question would then become how to project the T-terms into the future. Inclusion of a REG*TIME interaction effect would imply that there are region-specific changes of outmigration rates over time. For projection purposes this would mean that we have to project these regional specific factors into the future. The log-linear analysis makes it clear which terms are time-invariant and which terms are varying over time.

The explanatory models within this Poisson regression framework are a straightforward extension. The general form of an explanatory model of outmigration rates is:

$$\ln O_i^{as}(t) = \ln P_i^{as}(t) + X_i^{as}(t)\beta_i^{as}(t) \qquad (3)$$

where $X(t)$ is a vector of explanatory variables, and β a vector of coefficients to be estimated. Mixed forms of demographic models including non-demographic information are easily specified, by combining equations (2) and (3). For instance, we may specify a model of outmigration rates by including a demographic part of age- and sex-specific coefficients that specify the different levels of outmigration per age-sex combination, and an explanatory part of regional-specific and time-varying covariates to explain the differences in rates between regions and over time:

$$\ln O_i^{as}(t) = \ln P_i^{as}(t) + \mu_a^{AGE} + \mu_s^{SEX} + \mu_{as}^{AGE*SEX} + X_i(t)\beta + Z(t)\alpha \qquad (4)$$

Using the shorthand notation introduced above this model may be written as:
LPOP+A*S+X+Z.

One of the variables that may be included as an explanatory variable is the population size of the region. In that case, the variable is included twice: first as an offset, to reduce the dependent variable to a rate, and second as an indicator of region size, that captures the statistical artefact that the larger the region, the smaller the outmigration rate. Population size as an explanatory variable is therefore assumed to have a negative effect on outmigration rates.

Models for destination choice

For the specification of the destination choice model there are multiple options. The model includes a distance function which measures the friction of interaction between origin *i* and destination *j*. Here we have the choice between a demographic solution in the form of a historical migration matrix, and a functional form that involves geographical distance. The attractiveness function describes the attractiveness of the destination *j*, and here again the choice is between a purely demographic approach, where a region-specific dummy represents each region, and an explanatory approach with covariates. The combination of these options leads to three types of models as made clear in table 2. We will deal with three of the four model types, since one combination is not very useful in this respect (although it is the Poisson regression equivalent of the doubly constrained spatial interaction model using a distance function).

Table 2 Four different specifications of the destination choice model

	Attractiveness function	
Distance function	demographic	exogenous information
Demographic	1	2
spatial interaction function	not used	3

The model form is multinomial logit, which has the following form:

Model 1: Demographic model with distance function between origin i and destination j:

$$p_{j|i}^{as}(t) = \frac{\overline{M_{ij}} A_j^{as}}{\sum_{k=1}^{I} \overline{M_{ik}} A_k^{as}} \qquad (5)$$

where $\overline{M_{ij}}$ is the historical flow from *i* to *j* (e.g. the average of the last five years), and A_j^{as} is an attractiveness factor for region *j*, which may be age- and sex-specific, for instance: $A_j^{as} = \pi_{jas}^{DEST*AGE*SEX}$, or less complicated: $A_j^{as} = \pi_{as}^{AGE*SEX} \cdot \pi_j^{DEST}$ .The π's are the multiplicative equivalents of the loglinear μ terms. Note also that we have left out here the lower order interaction terms and main terms here but their effect is included in the three-way interaction terms that are released by their non-inclusion. The present specification still contains as many parameters as there are combinations of destination, age and sex.
This model merely says that out-migration flows out of *i* are distributed over the destination regions according to historical destination shares, adjusted by destination specific constants.

Technically this model is estimated as a Poisson model of the flows $M_{ij}^{as}(t)$ in GLIM using the following specification:

$$\log M_{ij}^{as}(t) = \mu_{iast}^{ORIG*AGE*SEX*TIME} + \log \overline{M_{ij}} + \mu_{jas}^{DEST*AGE*SEX} \qquad (6)$$

which is a doubly-constrained spatial interaction model for each age-sex combination (*a*,*s*) and with a historical migration matrix as the interaction term. The $\log \overline{M_{ij}}$ is the offset that remains outside of the estimation procedure (i.e. it is subtracted from the dependent variable *before* the estimation procedure) The $\mu_{iast}^{ORIG*AGE*SEX*TIME}$ parameters, one for each (*i*,*a*,*s*,*t*) combination, fit the expected outflows exactly equal to the observed outflows from *i* for each (*a*,*s*) combination in each year, and the $\mu_{jas}^{DEST*AGE*SEX}$ are proportional to the log of the observed inflows into *j*. Equation (7) is transformed into (6) as follows:

$$p_{j|i}^{as}(t) = \frac{M_{ij}^{as}(t)}{\sum_{k=1}^{I} M_{ik}^{as}(t)} = \frac{\exp\{\mu_{iast}^{ORIG*AGE*SEX*TIME} + \log \overline{M_{ij}} + \mu_{jas}^{DEST*AGE*SEX}\}}{\sum_{k=1}^{I} \exp\{\mu_{iast}^{ORIG*AGE*SEX*TIME} + \log \overline{M_{ik}} + \mu_{kas}^{DEST*AGE*SEX}\}} = \frac{\exp\{\log \overline{M_{ij}} + \mu_{jas}^{DEST*AGE*SEX}\}}{\sum_{k=1}^{I} \exp\{\log \overline{M_{ik}} + \mu_{kas}^{DEST*AGE*SEX}\}}$$
$$= \frac{\overline{M_{ij}} \pi_{jas}^{DEST*AGE*SEX}}{\sum_{k=1}^{I} \overline{M_{ik}} \pi_{kas}^{DEST*AGE*SEX}} \qquad (8)$$

Note that the parameters pertaining to the out-migration flows cancel out. A shorthand notation of this model is:

```
OAST + DAS + {OD}
```

where {OD} is the historical migration matrix, included as an offset.

Model 2: Explanatory model with OD distance function:

$$p_{j|i}^{as}(t) = \frac{\overline{M_{ij}} \exp\{\mu_{jas}^{DEST*AGE*SEX} + X_j(t)\beta_i^{as}\}}{\sum_{k=1}^{I} \overline{M_{ik}} \exp\{\mu_{kas}^{DEST*AGE*SEX} + X_k(t)\beta_i^{as}\}} \qquad (9)$$

Here we have added explanatory variables X_j to the demographic model, in order to explain the relative attractiveness of the destinations. Some of the μ terms need to be constrained to zero in order to make the model identifiable. The coefficients β may be age-, sex- and origin-dependent. The model may be estimated in GLIM similarly as model (6). It may be abbreviated to:

```
OAST + DAS + {OD} + ASX
```

where X refers to the set of explanatory variables.

Model 3: Spatial interaction model:

The formula is:

$$p_{j|i}^{as}(t) = \frac{\exp\{\mu_{jas}^{DEST*AGE*SEX} + X_j(t)\beta_i^{as} + F_i^{as}(W_{ij})\}}{\sum_{k=1}^{I} \exp\{\mu_{kas}^{DEST*AGE*SEX} + X_k(t)\beta_i^{as} + F_i^{as}(W_{ik})\}} \qquad (10)$$

Here, the historical flow matrix is replaced by a spatial interaction function $F_i^{as}(W_{ij})$ of distance W_{ij}. The function may be origin-, age- and sex-specific. The implementation in GLIM is similar to equation (6). This model may be abbreviated to:

OAST + DAS + OASX + OASF

where F refers to the distance function, and X as before to the set of explanatory variables for the attractiveness of the destination region.

4 Results

Results for the outmigration model

In line with the results of van Imhoff et al. (1997) the demographic model for outmigration was estimated as: OAS+T (for the UK the S dimension was not available in the data). This corresponds to a model for the outmigration rate of the form: $m_i^{as}(t) = A_t B_i^{as}$, i.e. a time factor and an origin-specific factor for each age-sex combination.

The best explanatory models for each country were slightly different. In shorthand notation they were as follows:

Sweden: `O+AS+T+LPOP+A*GDPLAG +A*UNEMPLAG +A*DENS`
UK: `O+A +T+LPOP+A*GDPLAG +A*UNEMPLAG +A*DENS`

Here GDPLAG is GDP of time t-1, and UNEMPLAG is unemployment one year lagged. DENS is population density of the origin region. This model includes dummies for age- and sex-specific rates (AS), an origin-specific factor (O), and GDP, unemployment and population density with age-specific coefficients. This means that we have age-specific coefficients of the explanatory variables. In general, the higher the GDP level, the lower the outmigration rates in Sweden, but not in the UK. Moreover, the higher the unemployment, the higher the outmigration rate for the younger ages in both countries, but for higher ages the effect is not significant (Sweden) or reverse (UK). Here, high unemployment is associated with low outmigration rates for middle aged and older people.

The best model with explanatory variables for the Netherlands was slightly different. Instead of lagged GDP and unemployment rate, the lagged regional differences with the national averages of GDP and unemployment were included in the model (GDPZLAG and UNEMPZLAG respectively), both not age-specific:

Netherlands: `O+AS+T+LPOP+GDPZLAG+UNEMPZLAG+A*DENS`

Here, GDP and unemployment have the expected signs and are significant. Density is marginally significant, and is positive: the higher the density, the higher the outmigration rate, especially for the older ages.

Table 3 shows the goodness of fit of the best models for the three countries. Although for Sweden and the UK the best explanatory model captures the same variables, the goodness of fit of these models compared to the best demographic models does not point to one overall conclusion. For Sweden the explanatory model gives a better fit to the data, whereas for the UK the reverse is true. Taking also the results of the Netherlands into account, we may conclude that in the Netherlands the demographic model gives an exceptionally good fit, when judged from the mean LR. This is an indication that in the Netherlands the structure of the out-migration process is relatively time-invariant.

Table 3 *Likelihood ratio test statistic (Deviance) results for out-migration models in three countries*

	demographic model AO(S)+T			'best' explanatory models		
	LR test stat.	**d.f.**	**mean LR**	**LR test stat**	**d.f.**	**mean LR**
Sweden	5706	380	15.0	4272	438	9.75
UK	31474	764	41.2	45168	900	50.2
Netherlands	1805	572	3.2	13330	684	19.5

Next we used these models for prediction purposes. Figure 2 shows scatterplots of predicted and observed migration rates (by region, age and sex) for each country. A perfect fit would mean that all points lie on the diagonal in the scatterplot. A heuristic fitting statistic is the R^2 and regression line between predicted and observed rates. For a good prediction, the intercept should be 0 and the slope should be 1, which can be verified when looking at the coefficients and confidence bounds of the regression output. This is only the case in the Netherlands. For all other countries and models the intercept is indeed 0 but the slope is less than 1, indicating that the higher the observed rate, the larger the underestimation of the rates. This is true for both Sweden and the UK.

We may conclude from the UK and Swedish results that the best fitting models for migration, based on information from the first half of the nineties, do not give adequate results in terms of prediction for the second half of the nineties: they underpredict, and this bias is linearly related to the size of the migration rate. There is not much difference here between the pure demographic model, and the model including explanatory variables. For the Netherlands, the situation is different: the predictive power of the models is higher, and there is no structural bias in the results. Both the demographic and the explanatory model give satisfactory predictions of the rates.

Results for the destination choice model

The three models presented above were estimated for Sweden, the UK and the Netherlands. The specification of the demographic model was derived from the results of van Imhoff et al. (1997) and can be abbreviated as follows:

OAST+DAS+{OD}

(for the UK no S-dimension was available). The OAST terms denote the outmigration totals for each subcategory per year, and,as explained above, are not related to the destination part. The true destination model terms are DAS (each destination has an age- and sex-specific attractiveness term) and {OD} which is the historical interaction pattern. In the explanatory models the explanatory variables used were a regional mass indicator (population plus employment summed: LMASS), unemployment, gross regional product GDP, accessibility and population density. The model specification for the explanatory models for each of the three countries turned out to be:

For the explanatory model with historical interaction matrix:
OAST+DA+A*GDPLAG+LMASS + {OD}

For the explanatory model with spatial interaction function:
OAST+DA+A*GDPLAG+LMASS + O*W+O*Cont

where W is the straight line distance between origin and destination, and Cont is a dummy indicating contiguity of adjacent zones. Both parameters are included in interaction with region of origin, which means that we have origin-specific distance functions. GDPLAG is the lagged value of GDP, and LMASS is the log of POP+EMP, or population and jobs in the region. The DA terms represent the age-profiles of inmigration in each region, and they are time-invariant. The value of GDP is generally in line with expectations, but specifically for age groups 20-49. For younger and older age groups the coefficients are not significant or negative. LMASS is negative, indicating that migrants at the NUTS-2 level are not attracted to larger regions in terms of population and/or jobs.

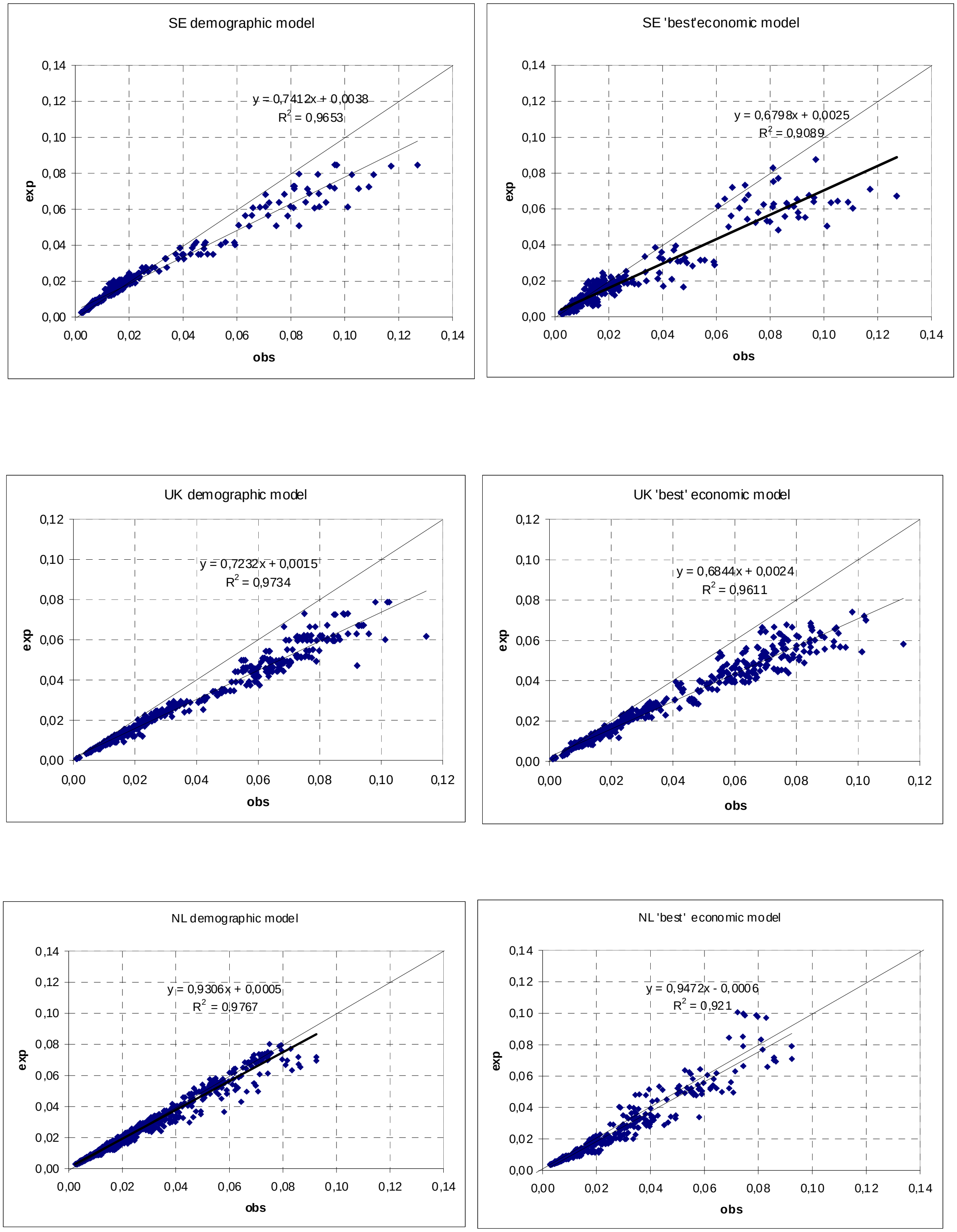


Figure 2 Observed and predicted outmigration rates for two models in Sweden, the UK and the Netherlands.

Table 3 shows the fit results of the 3*3=9 models.

Table 3 Goodness of fit of three destination choice models for three countries

	Demographic model			Explanatory model +OD			Explanatory model plus spatial interaction function		
	LR	d.f.	Mean LR	LR	d.f.	mean LR	LR	d.f.	Mean LR
Sweden	7629	2838	2.68	7542	2831	2.66	16901	2850	5.93
UK	175777	28614	6.14	175389	28608	6.13	483110	28539	16.9
Netherlands:	18749	7134	2.63	18714	7128	2.63	53803	7157	7.58

The goodness of fit of the models indicates that the demographic model, as well as the explanatory model with historical interaction parameters perform much better than the pure explanatory model with distance function. In the Netherlands the performance of the explanatory model including {OD} is much better than the other models, in Sweden and the UK there are hardly any differences between this model in fit with the demographic model.

These models, estimated on data of the period 1992-1995 were subsequently used to predict destination choice in the period 1996-1998 in each of the three countries. Figures 3, 4 and 5 present the results, in the form of scatter plots of observed and predicted destination probabilities for each country.

Figure 3: Observed and predicted destination probabilities 1996-1998, Sweden

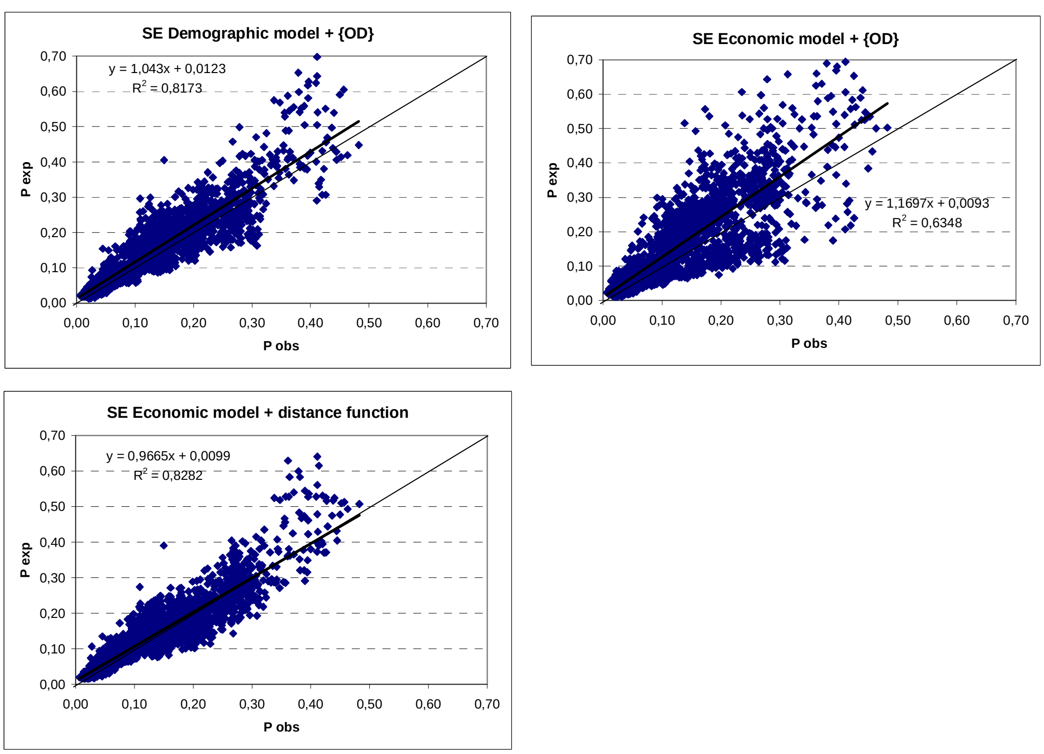


Figure 4: Observed and predicted destination probabilities 1996-1998, UK

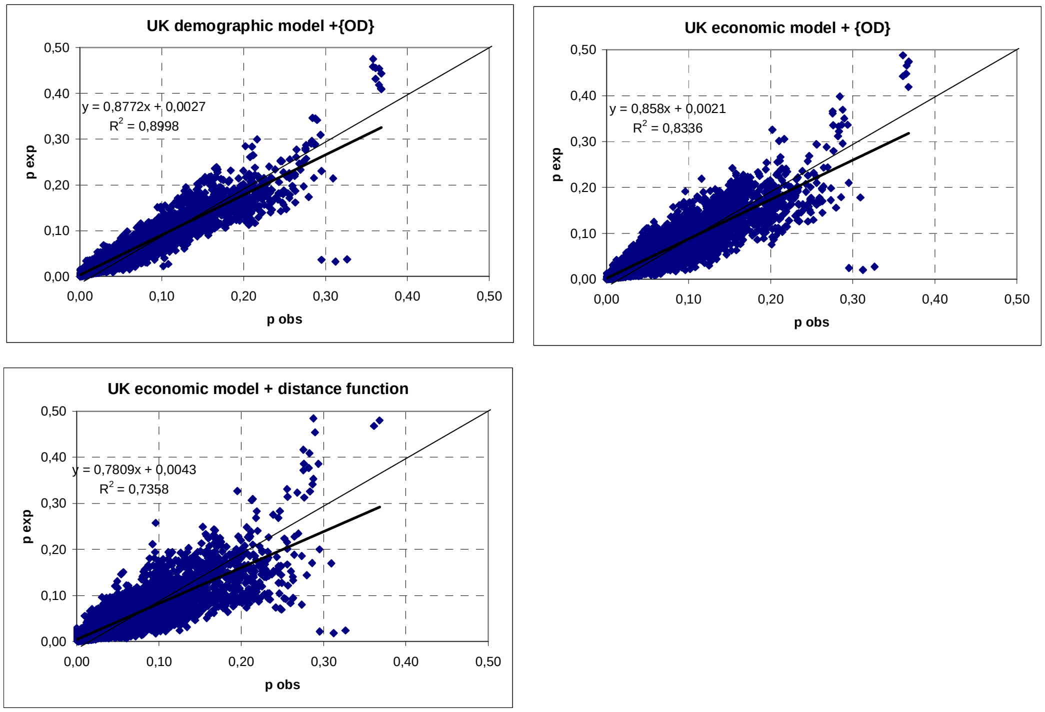
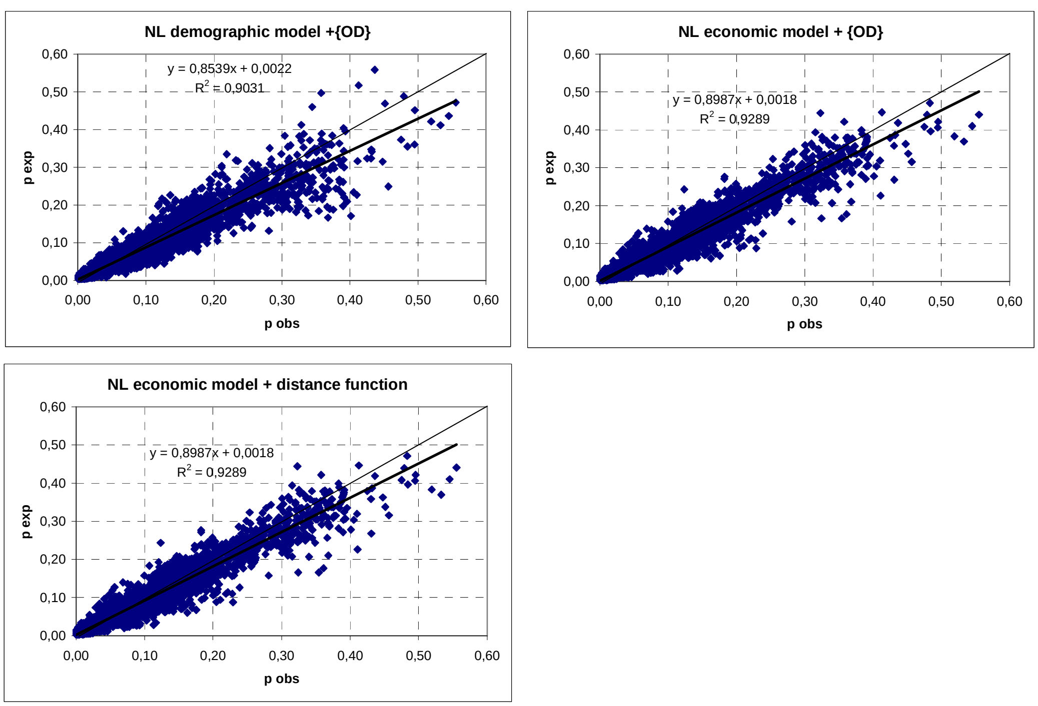


Figure 5: Observed and predicted destination probabilities 1996-1998, the Netherlands



The prediction results are different per country, and not always totally parallel to the model fit results reported in table 3. In Sweden we find that the explanatory model with distance function gives both the best model fit for the period 1992-1995, and the best prediction for the period 1996-1998, when judged from figure 3. The R^2 is higher than the other models, and the points cluster around the diagonal. For the UK we find the best fit for the demograhic model, and this model is also best in terms of prediction. In the Netherlands, the demographic model and the explanatory model with OD terms perform almost equally well in model fit. The predictive performance of all three models is not too different, with a slight advantage for the explanatory model including the OD term (when judged from both R^2 and slope of the regression).

Conclusions

We may conclude from the UK and Swedish results that the best fitting models for outmigration, based on information from the first half of the nineties, do not give adequate results in terms of prediction for the second half of the nineties: they underpredict, and this bias is linearly related to the size of the migration rate. There is not much difference here between the purely demographic model, and the model including explanatory variables. For the Netherlands, the situation is different: the predictive power of the models is higher, and there is no structural bias in the results. Both the demographic and the explanatory model give satisfactory predictions of the rates.

When looking at the destination choice models, we find different results per country. When judged from a predictive point of view, in Sweden and the UK we would prefer the pure demographic model, in the Netherlands we would choose the economic model with OD term. The common denominator in all these models is that historical interaction parameters perform better in prediction, but for the attractiveness function a similar conclusion may not be drawn. Results are different per country, although the purely demographic model in the Netherlands is not much worse here than the optimal model. This would support the conclusion that a pure demographic destination choice model is a good option for -short term- prediction. Note however, that we do not include the important problem of predicting the explanatory variables itself here, since we used observed values of these variables for the years 1996-1998. In a scenario study this is not a problem, but for prediction it surely is.

In the short run, destination patterns are quite stable, and may be predicted using historical patterns. If this is still true for long-term prediction cannot be judged from these analyses. In the long term the spatial structure of a country may change, and this is by definition not captured in purely demographic approaches. In that case we have to rely on explanatory variables, for instance in a scenario setting.

The major problem with prediction turns out to be in the outmigration rates. We found except in the Netherlands, where these rates are more stable than in the other countries, that the optimal models for the early nineties seriously underpredict outmigration in the second half of the nineties. The models appear to be good in predicting the regional differences in rates, but the time dimension is not captured well using these variables. This is equally true for explanatory and demographic models of outmigration. It is likely that overall internal mobility levels are related to the business cycle, for instance in the form as used by van der Gaag and van Wissen (2001). Such a model would be needed to set the overall migration level in each period. Future research is necessary to show if such an approach is feasible.

References:

Baydar, N. (1983) Analysis of the temporal stability of migration in the context of multi-regional forecasting, *Working Paper No 38,* Netherlands Interuniversity Demographic Institute, Voorburg.

Champion, A.G., Bramley, G., Fotheringham, A.S., Macgill, J. and Rees, P.H. (2002) A migration modelling system to support government decision-making, in Stillwell, J. and Geertman, S. (eds.) *Planning Support Systems in Practice*, Springer Verlag, Berlin, pp.257-278.

European Communities (1999) Regions, nomenclature of territorial units for statistics – NUTS.

Flowerdew, R. (1991) Poisson regression modelling of migration, in Stillwell, J.C.H. and Congdon, P. (eds.) *Migration Models: Macro and Micro Approaches*, Belhaven Press, London, pp. 92-112.

Flowerdew, R. and Lovett, A. (1988) Fitting constrained Poisson regression models to interurban migration flows. *Geographical Analysis,* 20: 297-307.

Fotheringham, A.S., Rees, P., Champion, A., Kalogiou, S. and Tremayne A.R. (2004) The development of a migration model for England and Wales I: overview and modelling out-migration, *Environment and Planning A*, 36 (forthcoming).

Rees, P.H., Fotheringham, A.S. and Champion, A.G. (2004) Modelling migration for policy analysis, Chapter 14 in Stillwell, J. and Clarke, G. (eds.) *Applied GIS and Spatial Analysis*, Wiley, Chichester, pp. 259-296 (forthcoming).

Rees, P., Kupiszewski, M., Eyre, H., Wilson, T. and Durham, H. (2001) The evaluation of regional population projections for the European Union, *Eurostat Working Paper, Population and Social Conditions* 3/2001/E/no 9.

Rees, P.H. and Wilson, A.G. (1973) Accounts and models for spatial demographic analysis 1: aggregate population, *Environment and Planning*, 5: 61-90.

Rees, P.H. and Wilson, A.G. (1975) A comparison of available models of population change, *Regional Studies*, 9: 39-61

Rees, P.H. and Wilson, A.G. (1977) *Spatial Population Analysis*, Edward Arnold, London.

Rogers, A. (1966) Matrix methods of population analysis, *Journal of the American Institute of Planners,* 32: 40-44.

Rogers, A. (1967) Matrix analysis of interregional population growth and distribution, *Papers of the Regional Science Association,* 18: 17-196.

Rogers, A. (1968) *Matrix Analysis of Interregional Population Growth and Distribution*, University of California Press, Berkeley.

Rogers, A. (1973) The multi-regional life table, *Journal of Mathematical Sociology*, 3: 127-137.

Rogers, A. and Willekens, F. (eds.) (1986) *Migration and Settlement: A Multiregional Comparative Study,* Reidel, Dordecht.

Scholten, H.J. and van Wissen, L.J.G. (1985) A comparison of the loglinear interaction model with other spatial interaction models. In: P. Nijkamp, H. Leitner and N. Wrigley (eds.) *Measuring the Unmeasurable*, Martinus Nijhoff Publishers, Dordrecht, pp. 177-196.

Van der Gaag, N., van Imhoff, E. and van Wissen, L. (1997a), Regional population projections in the countries of the European Economic Area, update of the 1992-Questionnaire. *NIDI Working paper* no.: 97/1, The Hague

Van der Gaag, N., van Imhoff E. and van Wissen, L. (2000) Internal migration scenarios and regional population projections for the European Union, *International Journal of Population Geography*, vol. 6(1):. 1-19.

Van der Gaag, N. and van Wissen, L. (2001), Economic developments and internal migration propensities. Paper presented at the European Population Conference 2001, Helsinki, 7-9 June.

Van Imhoff, E., van der Gaag, N., van Wissen, L. and Rees, P.H. (1997) The selection of internal migration models for European regions, *International Journal of Population Geography*, 3: 137-159.

Van Imhoff, E., van Wissen, L. and Spiess, K. (1994), *Regional Population Projections in the countries of the European Economic Area*, NIDI CBGS Publications no. 31, The Hague, Swets and Zeitlinger.

Van Wissen, L.J.G. and Rima, A. (1988) *Modelling Urban Housing Market Dynamics,* Studies in Regional Science and Urban Economics, North Holland-Amsterdam.

Willekens, F.J. (1980) Entropy, Multiproportional adjustment and analysis of contingency tables, *Sistemi Urbani*, no 2/3, pp 171-201

Willekens, F.J. (1983) Specification and calibration of spatial interaction models. A contingency table perspective and an application to intra-urban migration in Rotterdam, *Tijdschrift Economische Sociale Geografie TESG*, vol. 74, pp. 239-252.

Willekens, F. and Drewe, P. (1984) A multi-regional model for regional demographic projection, in ter Heide, H. and Willekens, F. J. (eds.) *Demographic Research and Spatial Policy: The Dutch Experience*, Academic Press, London, pp. 309-334.

Wilson, A.G. (1970) *Entropy in Urban and Regional Modelling*, Pion, London.

Wilson, A.G. (1972) *Papers in Urban and Regional Analysis*, Pion, London.

Wilson, A.G. and Rees, P.H. (1974) Accounts and models for spatial demographic analysis 2: age-sex disaggregated populations, *Environment and Planning A*, 6: 101-116.