# 21 Including level of education in regional population estimates using hybrid loglinear models

*Leo van Wissen*

## Abstract

The term hybrid loglinear model was introduced by Frans Willekens and Nazli Baydar in 1985 (Willekens & Baydar 1985), as an extension of the class of loglinear models for the analysis of contingency tables. The term hybrid refers to nonstandard forms of loglinear models. This model is used here to estimate the population by age, sex, NUTS3 region and level of education. Educational attainment is only available without age detail at the NUTS2 level at the European scale. An aggregated logit model, based on a small set of predictors of regional educational attainment is used to estimate a prior distribution at the NUTS3 level. Using Iterative Proportional Fitting we fit this prior distribution to known marginal totals at the regional and national scale, following the principles set out in Willekens' paper from 1980.

## 21.1 Introduction

Whereas traditionally age and sex were the defining dimensions of a population, the recognition of variation in demographic behavior resulted in the introduction of more dimensions of interest, such as country, region, household position, and more recently also education. The cross-classification of the population using multiple dimensions leads to a large number of cells in the contingency tables describing these populations, which may be the reason that many view demography as a data-hungry activity, for which you have to be a data-fetishist to enjoy working with

it.[1] Because the data demands for these large tables are substantial, a common problem in this type of analysis is incomplete data: there is partial information available, but not for every cell. Demographers have developed or adopted various methods to deal with these problems. Few know however, that Frans Willekens has been at the forefront of these developments in the seventies and eighties. This is quite understandable as a result of his involvement in the development of multidimensional demographic methods at the time, for which these methods were very well suited.

In this contribution I will present a method to estimate the NUTS3 population of European countries by age, sex and level of education using partial available information from various international statistical sources. The method is well known and already described in the comprehensive paper by Frans in 1980 in Sistemi Urbani, entitled "Entropy, multiproportional adjustment and the analysis of contingency tables"; see (Willekens 1980). The usefulness of this approach is still undisputed, as my contribution demonstrates, and my paper is a tribute to his original paper, which, strangely enough, was only cited 49 times, according to Google Scholar.

In the next section a short overview of the method is presented. In section 22.3 the estimation problem is explained. The loglinear model formulation is introduced in section 22.4. Section 22.5 presents an aggregated logit formulation to estimate a crucial missing interaction between region and educational attainment, that is not present in the given partial information. Next, in section 22.6, the estimated model for the Netherlands is applied and tested for the case of Norway. Section 21.7 gives some results at the European level, and section 21.8 concludes.

## 21.2 The analysis of contingency tables: an overview

In demographic research, contingency tables are very common. Contingency tables of the counts of the number of persons by age and sex are the bread and butter of every demographic analysis of populations. The two-way table by age and sex can be extended to include other dimensions of demographic interest, such as position in the household, country or region of residence, or education. The number of

[1] At least, some researchers do enjoy this activity, and a high concentration of such persons is to be found among the authors of this LA.

cells in such multiway tables multiplies with every additional dimension included. Loglinear models were developed in the 1970s by Bishop et al. (2005) to deal with these complex discrete multivariate distributions. They provided the statistical background for the analysis of such distributions. Frans was among the first to discover the added value of this approach in demographic analysis. He wrote a comprehensive overview in his 1980 article in Sistemi Urbani referred to above, in which he showed that many existing models for the analysis of interaction between multiple variables could be brought under this umbrella. These earlier approaches included for instance the iterative proportional fitting procedure introduced in demography by Deming & Stephan (1940) for combining census with sample information, the Fratar-Furness method (Fratar 1954) in transportation science to update a spatial interaction table to known marginal totals, Stone's RAS method (Stone 1962) for updating input-output tables in economics, entropy maximization (Wilson 1970) in spatial interaction models, information minimization (Kullback 1959), and linear programming (Nijkamp 1975). A central element in this general approach is iterative proportional fitting (IPF), which can be viewed as a method to find a multivariate distribution that is as close as possible ("adds as little as possible information") to a given prior distribution, given a set of marginal constraints. The method is since then widely accepted in demography, and well-suited to estimate the cell counts of an incomplete contingency table where only partial information is available, in the form of marginal and certain interaction totals. This is exactly the problem to be addressed in this contribution.

## 21.3 The estimation problem

In the project PREMIUM_EU[2], a regional demographic database at the NUTS3 level for all EU27 countries was built, that includes age (by 5-year age categories, with the upper age category 85+), sex, and level of education. There are, although the exact classification changes regularly, more than 1200 NUTS3 regions. Information about age (5-year age categories) and sex is available through the Eurostat REGIONS database. The European Joint Research Centre JRC has harmonized these data, which resulted in a time series at the NUTS3 level starting in 1990.

[2]The research on which this contribution is based was conducted as part of the Horizon Europe project PREMIUM_EU (Grant Agreement number: 101094345).

The research problem that we try to solve here is to expand this table of three dimensions (region $R$, sex $S$, age $A$) with a fourth dimension: educational attainment $E$, in three categories: Low, Middle and High educated. For the Netherlands for example, having 40 NUTS3 regions (so called COROPs) this table contains 40x3x2x18 = 4320 cells. The level of education is not available for most European countries at the NUTS3 level. Eurostat provides information on level of education at the NUTS2 level[3] but not broken down by age[4]. More detailed information of educational attainment is provided at the national level by the Wittgenstein Centre for Demography and Human Capital WIC. The available partial regional information is summarized in Table 21.1. As already mentioned, Eurostat collects annual information about the population by age and gender for each NUTS3 region[5]: $R \times S \times A$. In the PREMIUM_EU project the years 2010, 2015 and 2020 are used. Educational attainment is a more problematic dimension. WIC has made estimates for all countries in the world of the population by sex, age and level of education, but not for regions within countries. There is some information at Eurostat/JRC, based on the ongoing European labor force surveys: The population of working age (between 15 and 75 years) by sex and educational attainment, at the NUTS2 level. The notation in the first column of the table will be explained below.

**Table 21.1:** Data resources for estimation

| Notation | Source | Description |
|---|---|---|
| $R \times S \times A$ | Eurostat/JRC | Population by age (0-5, ..., 85+) and gender for each NUTS3 region |
| $S \times A \times E$ | WIC | Educational attainment by age (0-5, ..., 100+) and gender |
| $R2 \times S \times E$ | Eurostat/JRC | Population 15–75 by gender and educational attainment for each NUTS2 (R2) region |

[3]There are currently 244 NUTS2 regions in the EU. NUTS3 regions are a subdivision of NUTS2 regions.

[4]After the analysis for this chapter was finished, Eurostat released the data of the Census of 2021, which contains a table at the NUTS2 level by broad age groups and sex. It was too late to include this table in the current estimation.

[5]This information is even available at the finer detailed regional level of the LAU (Local Area Units).

## 21.4 Model specification

The problem can be specified as follows, using the modelling language that was introduced in the programming language GLIM (an acronym for Generalized Linear Interactive Modelling) in the 1970s (Aitkin et al. 2005). In this modelling language a contingency table with dimensions sex $S$ and age $A$ can be written as $S \times A$. If we only have the marginal totals of age and sex, we can estimate a table $S + A$, assuming independence between both dimensions. The formulation of independence between the two dimensions in a loglinear formulation is given by:

$$\log \hat{M}_{ij} = \mu + \mu_i^A + \mu_j^B$$

where $\hat{M}_{ij}$ is the expected value of cell $(i, j)$ in the two-way table under the assumption of independence of factors $A$ and $B$, and the coefficients are given by:

$$\mu = \frac{1}{IJ} \sum_{i,j} \log \hat{M}_{ij}, \mu_i^A = \frac{1}{J} \sum_j \log \hat{M}_{ij} - \mu, \mu_j^B = \frac{1}{I} \sum_i \log \hat{M}_{ij} - \mu. \quad (21.1)$$

If $\mu$ is determined, only $I$-1 coefficients of factor $A$ can be estimated, and one is redundant, and similarly for factor $B$. Usually, the constraint that all coefficients sum to 0 is used, but other designs can be imposed as well, for instance that the coefficient of the first or last category of the factor is 0. Note that the parameters are estimated from the expected values $\hat{M}_{ij}$. This means that the expected values under the model have to be estimated first, and from these expected values the parameters are derived. The sufficient statistics to estimate the expected values are the marginal totals implied by the specified main and interaction effects. For the model of independence (21.1) these are the row- and column totals of the $A \times B$ table.

By comparing the expected values $\hat{M}_{ij}$ with observed values $M_{ij}$, the hypothesis of independence can be tested, using the well-known Chi-square test, with appropriate degrees of freedom: in this case $I \times J - 1 - (I - 1) - (J - 1)$. However, many contingency tables of data suffer from overdispersion, which heavily inflates the test statistic. Moreover, many tables are based on register or census data of the whole population, not samples. Therefore, the value of the test statistic is only

indicative of the fit of the model.

If the hypothesis of independence is rejected, an interaction term is necessary. In a two-way table this leads to a saturated model, with the number of coefficients equal to the number of observations (cells) in the table. The model of interdependence is $A \times B$, or, in a loglinear formulation:

$$\log \hat{M}_{ij} = \mu + \mu_i^A + \mu_j^B + \mu_{ij}^{AB}$$

and the interaction term in this model is estimated as:

$$\mu_{ij}^{AB} = \log \hat{M}_{ij} - \mu - \mu_i^A - \mu_j^B.$$

This example readily extends to more than two dimensions. Our table to be estimated has four dimensions, with cell entries $M_{ijkl}$, where $i$ is the index for region, $j$ for sex, $k$ for age and $l$ for education. The available partial information allows only to estimate a model with restrictions on many parameters.

$$R \times S \times A \times E \approx R \times S \times A + S \times A \times E + R2 \times S \times E$$

This is a hybrid loglinear model, since $R2$ is an aggregate of NUTS3 regions and therefore not a standard factor in a loglinear model. In parametric form the model reads:

$$\begin{aligned} \log \hat{M}_{ijkl} = \quad & \mu + \mu_i^R + \mu_j^S + \mu_k^A + \mu_l^E + \mu_{ij}^{RS} + \mu_{ik}^{RA} + \mu_{jk}^{SA} + \\ & \mu_{jl}^{SE} + \mu_{kl}^{AE} + \mu_{kl}^{R2E} + \mu_{ijk}^{RSA} + \mu_{jkl}^{SAE} + \mu_{jkl}^{R2SE} \end{aligned} \tag{21.2}$$

Note that in this hierarchical model, if a higher dimensional interaction is included, the lower-level interactions between these variables are also included. It is clear that the crucial interaction at the NUTS3 level is missing. Although $R2 \times S \times E$ provides some information between the regional dimension and the level of education, it does not differentiate between NUTS3 regions within a NUTS2 region. This is most clearly seen if we observe a single sex-age combination, say 25-30 Females. For this group we have the distribution over the regions $R$ in a country, as well as the nation-wide level of education $E$. The interaction $R2 \times S \times E$ assigns values of educational attainment to each NUTS3 region based

on the corresponding NUTS2 regional context, and for all ages 15-75. As a result, all 25-30 aged Female categories in the NUTS3 regions of the same NUTS2 region will be assigned the same values. We therefore need an $R \times E$ interaction term (i.e., at the NUTS3 level), and possibly such interaction term could also be age- or sex-dependent. This information exists only for specific countries, but not European wide. To fill in this missing interaction we need a model that predicts the educational distribution of each NUTS3 region as a function of regional characteristics. This model, an aggregated logit model (a member of the family of loglinear models), is explained in the next section in some detail. The estimates of this model generate a distribution of level of education as a function of the regional characteristics. We include this term $R \times E$, or, if it is age and sex-specific $R \times S \times A \times E$, in the model as an *offset*. An offset is a covariate in the model with a fixed parameter value of 1. This could be specified as:

$R \times E \times S \times A \approx R \times S \times A + E \times S \times A + \{R \times E \times S \times A\}$

where the notation {..} is used to denote the offset. This offset can be interpreted as prior information to be included in the estimation. The parametric form of this model is:

$$\begin{aligned} \log \hat{M}_{ijkl} &= \mu + \mu_i^R + \mu_j^S + \mu_k^A + \mu_l^E + \mu_{ij}^{RS} + \mu_{ik}^{RA} + \mu_{jk}^{SA} + \\ &\quad \mu_{jl}^{SE} + \mu_{kl}^{AE} + \mu_{ijk}^{RSA} + \mu_{jkl}^{SAE} + \log \check{M}_{ijkl}. \end{aligned}$$

As will be explained in the next section, we include the $R2 \times S \times E$ table as a covariate in the estimation of the prior information $\log \check{M}_{ijkl}$. The formula has one intercept, four main effects, five two-way interactions, and two three-way interactions. This formulation makes clear that because of the partial information available, neither the two-way interaction $\mu_{il}^{RE}$, nor the three-way interactions $\mu_{ijl}^{RSE}$ and $\mu_{ijl}^{RAE}$ are included (i.e., these are set to zero). Instead, the prior distribution $\check{M}_{ijkl}$ contains an approximation of the interactions between these factors, derived from the logit predictions to be discussed below. Iterative Proportional Fitting (IPF) starts with the prior distribution $\{R \times S \times A \times E\}$, which is scaled in a number of rounds to fit it to the marginal totals $[R \times S \times A]$ and $[E \times S \times A]$, until convergence is reached.[6]

[6]The IPF model was programmed, not in Fortran (which would have been suitable given the timing of the origin of the approach), but, since all data are stored in Excel files, in an almost equally

## 21.5 The aggregated logit model to estimate the regional distribution over educational attainment

An aggregated logit model is equivalent to a loglinear model. In a loglinear model such as equation (21.2) there is no dependent or independent variable. The model estimates a multivariate discrete distribution. But we can designate one factor, say education, as the dependent variable, and estimate its value, conditional on the value of the other factors. For instance, we could model the probability that a person living in region $i$, with sex $j$ and age $k$ will have educational attainment $l$. This probability is

$$Prob_{l|ijk} = p_{l|ijk} = \frac{M_{ijkl}}{\sum_{l'} M_{ijkl'}}. \tag{21.3}$$

Substitution of (21.2) in (21.3) gives:

$$p(l|ijk) = \frac{\exp(\mu^{SE}_{jl} + \mu^{AE}_{kl} + \mu^{SAE}_{jkl} + (\log \check{M}_{ijkl}))}{\sum_{l'}[\exp(\mu^{SE}_{jl'} + \mu^{AE}_{kl'} + \mu^{SAE}_{jkl'} + (\log \check{M}_{ijkl'})]}. \tag{21.4}$$

All terms not related to $E$ cancel out. The $p$ can be either interpreted as the probability of having a certain educational level, or as population shares. In the current approach the interpretation of shares is more to the point. The $\mu$-terms describe differences in educational attainment between age- and sex categories, based on the national distribution from the WIC table $S \times A \times E$. By using data from countries that have all variables $R$, $S$, $A$ and $E$, we can model the prior distribution $\log \check{M}_{ijkl}$ with a set of region-specific explanatory variables. The estimated coefficients of such a model can then be applied to predict the value of the prior distribution for other countries without a complete $R \times S \times A \times E$ - distribution, but with the explanatory variables. We present the results of the estimation for the Netherlands, as an example of a country with all variables available. We estimate the following model on Dutch data:

$$p(l|ijk) = \frac{\exp(\mu^{SE}_{jl} + \mu^{AE}_{kl} + \mu^{SAE}_{jkl} + \sum_m \beta_{jkl,m} X_{i,m})}{\sum_{l'}[\exp(\mu^{SE}_{jl'} + \mu^{AE}_{kl'} + \mu^{SAE}_{jkl'} + \sum_m \beta_{jkl',m} X_{i,m})]}, \tag{21.5}$$

old-fashioned language Visual Basic, which was introduced by Microsoft in 1993.

where $\sum_m \beta_{jkl,m} X_{i,m}$ is a linear sum of $m$ region-specific variables $X_{i,m}$, with weights $\beta_{jkl,m}$.

The educational attainment data, by region, sex and age (i.e. the observed table $E \times R \times S \times A$) for the Netherlands come from the Statline database of Statistics Netherlands. Statistics Netherlands provides information for 10-year age categories (15-25, ... 65-75). We use the 2015 and 2020 data for the estimation. Table 21.2 gives an overview of the explanatory variables used in the model. Note that the educational attainment at the NUTS2 level is used as a predictor in this model.

**Table 21.2:** Explanatory variables of the regional NUTS3 share of the population by educational attainment

| Variable | Description |
|---|---|
| Educ15–75 | Share of the population aged (15–75) by educational attainment and sex |
| %Hightech | Share of employment in professional, scientific and technical activities; administrative and support service activities |
| Econ_index | The economic index as calculated in Arnold (2024). It is a composite index of regional product per capita, and unemployment |

The model was estimated using *R* function glm as a hybrid log-linear model including all available terms $R \times S \times A$ and $E \times S \times A$, plus quantitative predictors of the regional educational attainment distribution. Table 21.3 shows the deviance fit for a number of nested models for 2015 and 2020. It gives an impression of the contribution of the explanatory variables to the overall fit between observed and predicted regional educational attainment shares.

**Table 21.3:** Model fit of aggregated logit estimates of educational attainment in NUTS3 regions in the Netherlands

| Model | Deviance (2015) | Df | Deviance (2020) | Df |
|---|---|---|---|---|
| 1: $E \times R \times S \times A + E \times S \times A$ | 298666 | 936 | 321639 | 936 |
| 2: Model 1 + Educ15–75 | 234582 | 933 | 240933 | 933 |
| 3: Model 2 + %Hightech | 207490 | 931 | 232546 | 931 |
| 4: Model 3 + Econ_index | 184016 | 929 | 183516 | 929 |

Table 21.3 shows that the three explanatory variables have a substantial effect on the model outcomes. The economic index, one of the dimensions of the regional development concept, is the strongest variable. Further interactions of these variables with age and sex do not add much to the fit. This means that the $\beta_{jkl}$'s can be simplified to $\beta_l$.

The parameter estimates $\beta_l$ of these three explanatory variables are given in Table 21.4.

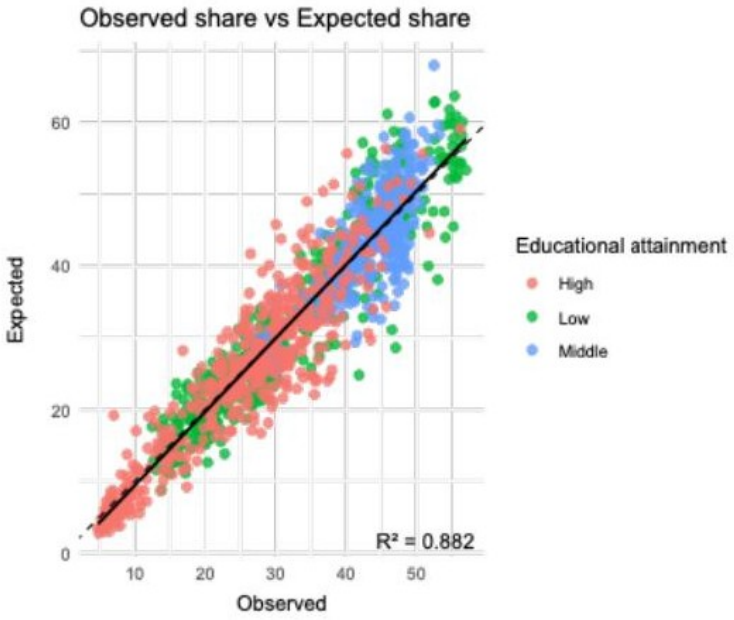
**Table 21.4:** Parameter estimates of explanatory variables in 2015 and 2020

| | 2015 | | | 2020 | | |
|---|---|---|---|---|---|---|
| Variable | Low | Middle | High | Low | Middle | High |
| Intercept | 3.145 | 3.035 | (ref) | 3.176 | 3.430 | (ref) |
| NUTS2 Educ 15–75 (/10000) | -0.15 | -0.12 | (ref) | -0.25 | -0.15 | (ref) |
| %Hightech (/100) | 0.255 | (ref) | 0.818 | 0.638 | (ref) | -2.162 |
| Econ_index | -0.178 | (ref) | 1.684 | 0.004 | (ref) | 2.760 |

The most important variable is Econ-index. The higher the economic index of a region, the higher the share of highly educated, as expected. In 2015 it also correlates negatively with the share of low educated. The percentage of employed in the Hightech sector is positive for the low, and mixed for the share of high educated: positive in 2015 but negative in 2020. The effect of the NUTS2 educational distribution is marginal and negative for low and middle educated shares, relative to the high educated shares at the NUTS3 level. This implies that on average higher shares of low or middle educated at the NUTS2 level correlate with lower shares at the NUTS3 level. Including interaction effects with either age or sex does not greatly improve the fit of the model, while making it substantially more complicated.

Using these parameter estimates, a prior distribution $\log \hat{\hat{M}}_{ijkl}$ can be estimated. Figure 21.1 shows the expected and observed shares based on the 2015 data. The figure shows a decent fit ($R^2$ is 0.88 for 2015, and similarly 0.87 in 2020).

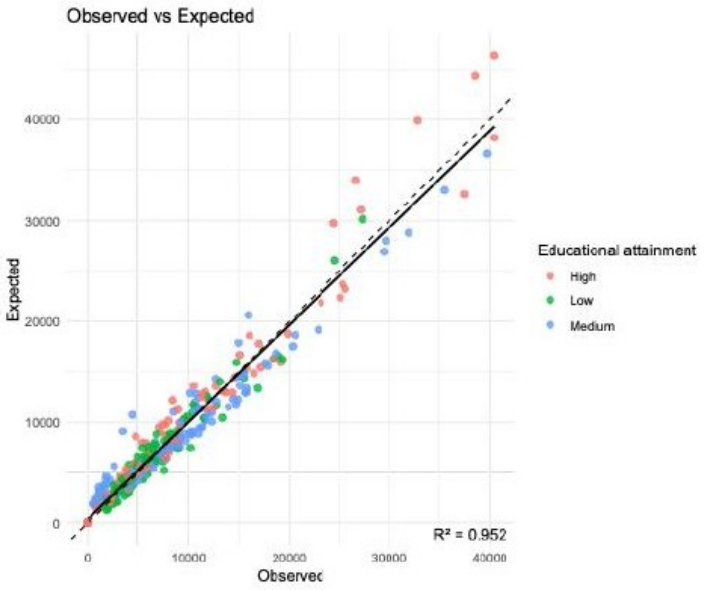
**Figure 21.1:** Observed and expected shares of levels of education for NUTS3 regions by age and sex, 2015, in the Netherlands.




## 21.6 How well does this prior distribution work for other countries?

We use the estimated values of the logit model for the Netherlands to construct the prior distribution for Norway. The estimated values can be compared with the observed counts, which are available from Statistics Norway. Norway has 11 NUTS3 regions, and Statistics Norway publishes data on regional educational attainment for 6 age categories. In total there are therefore 11x2x6x3 = 396 cell counts. Figure 21.2 shows the fit of the resulting model. Each dot represents observed and expected shares of low, middle and high educated for each combination of region, age and sex. The fit is very satisfactory. Of course, this does not indicate that the model can be transferred to any country. The Netherlands and Norway are both North-western European countries with a highly educated population. Nevertheless, as an illustration we used the model estimates, to estimate the level of education by age, sex and NUTS3 region for all European countries.

**Figure 21.2:** Observed and Expected counts of low, middle and high educated by age, sex and region, Norway, 2015




## 21.7 Some results at the European level

The variables Econ_index, %Hightech and NUTS2, together with the estimated coefficients from the Dutch logit model will generate a prior distribution $\{R \times E\}$ for each country. Unfortunately, data for Austria, Germany and Spain are not available in the Eurostat Regions database. For all other countries, the educational distribution, by NUTS3 region, age and sex can be estimated. Figure 21.3 shows the resulting pattern for highly educated 25-30 years old women in 2015.

## 21.8 Conclusion

Frans Willekens was the first to provide an integrated overview of the various methods that existed in the 1970s in such diverse fields as transportation science, economics, and geography to estimate missing values in a multi-way contingency

**Figure 21.3:** Estimated share of high educated women aged 25-29 in NUTS3 regions in Europe in 2015

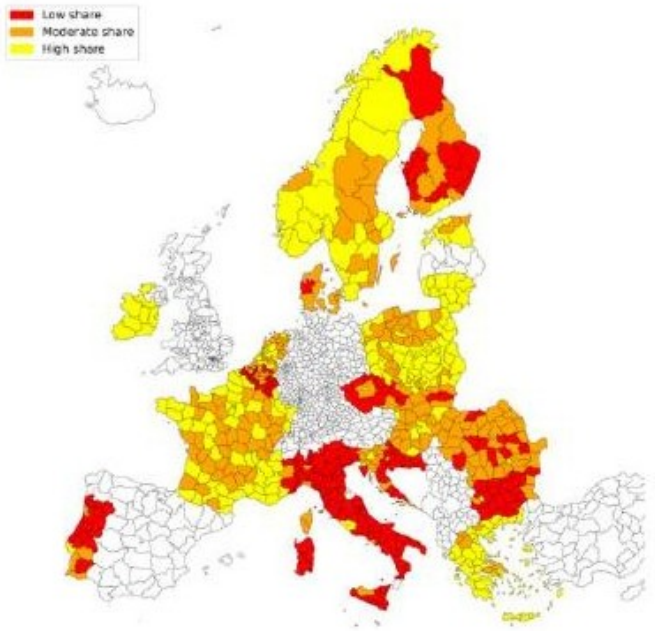



table. The method of Iterative Proportional Fitting is used today by many demographers for filling in the cells of such tables. Few know that Frans was at the heart of the introduction of this family of techniques in the discipline. Loglinear models for the analysis of contingency tables have become less popular these days than in the 1970s and 1980s. Logit models, where one of the variables is explicitly treated as the dependent variable, such as discussed in Section 22.5, are closely related and can easily be derived from the results of a loglinear model of a contingency table. Nevertheless, detecting the most informative interactions in a multidimensional distribution has its own merits, and should be in every demographer's toolbox. In this contribution we show that the method is still very useful, and can accommodate hybrid forms such as used here, by including prior distributions of interactions between variables, itself the result of a separate modelling strategy.

Today alternative methods exist, such as Bayesian analysis or machine learning techniques. It would be interesting to compare the outcome of such models with the more traditional approach in this chapter. The advantage of the current method

is that it is transparent. The fit of the model for the two countries where the predications could be compared with observations, The Netherlands and Norway, is quite satisfactory.

## Acknowledgements

Nico Keilman gave useful comments on an early version.

## Afterword

I came to know Frans when I started my PhD in 1981. Multiregional population models were quite new at the time, and I (together with Annemarie Rima) made extensively use of these new methods. My PhD involved the specification, estimation and projection of a regional household and housing market model, for which it was necessary to copy - by hand - published tables of the 1971 Dutch census, to combine it with housing market survey information, and to program everything in Fortran on a main frame. These multiregional models were very data hungry, and population registers or surveys at that time were totally insufficient to provide the information on a cell-by-cell basis. Frans saw the potential of loglinear models to estimate the missing information for such data-hungry models, and I thankfully made use of his ideas. Since then, a lot has changed: there is much more detailed information available, today's personal computers are much more powerful than the large mainframes of the eighties, and there are new techniques and much more powerful programming languages to choose from. Still, the problem of incomplete information remains, which is partly due to the increased ambitions of researchers. These ambitions, for some reason, always keep ahead of what is possible.

Since that first encounter between a PhD student and a leading demographer, we have stayed in frequent contact: at NIDI, where I started working in the early 1990s, and at the University of Groningen, where I developed, stimulated by Frans, the field of the Demography of Firms. Much later, in 2010, I followed in his footsteps as his successor as Director of NIDI. Under his directorship the institute developed a strong academic character, which turned out to be a very important asset to survive the difficult period of budget cuts in 2013 through the affiliation with the University of Groningen. So, NIDI owes a lot to Frans, but clearly my

career would have been quite different without him as well. Many, many thanks Frans.

## Bibliography

Aitkin, M. A., Francis, B. & Hinde, J. (2005), *Statistical modelling in GLIM 4 (Vol. 32)*, Oxford University.

Arnold, B. (2024), 'Deliverable 4.1 Workpackage 4, Horizon Europe'. **URL:** *https://premium-eu.org/fume-publications/atlas-of-regional-development-mapping-the-landscape-of-regional-development-in-europe/*

Bishop, Y., Fienberg, S. & Holland, P. (2005), *Discrete Multivariate Analysis: Theory and Practice*, M.I.T. Press, Cambridge (Mass.).

Deming, W. E. & Stephan, F. F. (1940), 'On a least squares adjustment of a sampled frequency table when the expected marginal totals are known', *The Annals of Mathematical Statistics* **11**(4), 427–444.

Fratar, T. (1954), 'Vehicular trip distribution by successive approximations', *Traffic Quarterly* **8**(1).

Kullback, S. (1959), *Information Theory and Statistics*, John Wiley, New York.

Nijkamp, P. (1975), 'Reflections on gravity and entropy models', *Regional Science and Urban Economics* **5**(2), 203–225.

Stone, R. (1962), 'Multiple classifications in social accounting', *Bulletin de l'Institut International de Statistique* **39**(3), 215–233.

Willekens, F. (1980), 'Entropy, multiproportional adjustment and the analysis of contingency tables', *Sistemi Urbani* **2**, 171–201.

Willekens, F. & Baydar, N. (1985), Hybrid log-linear models, *in* Nijkamp.P. & H. Leitner, eds, 'Measuring the Unmeasurable', Martinus Nijhoff, pp. 141–176.

Wilson, A. (1970), *Entropy in urban and regional modelling*, Routledge.