WP. 23 Rev.1
3 October 2007

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint Eurostat/UNECE Work Session on Demographic Projections
(Bucharest, 10-12 October 2007)

Agenda item 4: Fertility

FAMILY AND FERTILITY EVENTS: AGE PROFILES ESTIMATION STARTING FROM SURVEY DATA

Invited Paper

Submitted by Bocconi University[1]

Abstract

This paper has been developed within the research project called "MicMac: bridging the micro-macro gap in population forecasting". The objective of MicMac is to develop a methodology that offers a bridge between aggregate projections of cohorts (Mac) and projections of the life courses of individual cohort members (Mic). The project is supported by the European Commission within the 6th Framework Programme[2].

In this paper we deal with the methodological background of the estimation of age profiles starting from survey data. In particular, we present a method for the estimation of age profiles relating to transitions between various living arrangements and on fertility behaviours. These age profiles for the relevant transitions in the field of family and fertility, will represent data to be included as input in the MicMac model.

The method is applied to micro-level data from the Household survey "Famiglie e Soggetti Sociali", conducted in Italy at the end of 2003 and from the Fertility and Family Survey conducted in the Netherlands during the same year. Data preparation for this specific application is also discussed. The methods developed here can be applied in general to a setting where micro-level data on transitions are available from a large-scale representative survey.

[1] Prepared by Roberto Impicciatore and Francesco C. Billari, "Carlo F. Dondena" Centre for Research on Social Dynamics.
[2] More information on MicMac is available on the website www.micmac-projections.org.

Acknowledgments: the data for this analysis have been provided by ISTAT (Italian National Institute of Statistics) and Centraal Bureau voor de Statistiek, The Netherlands.

.

1. Introduction

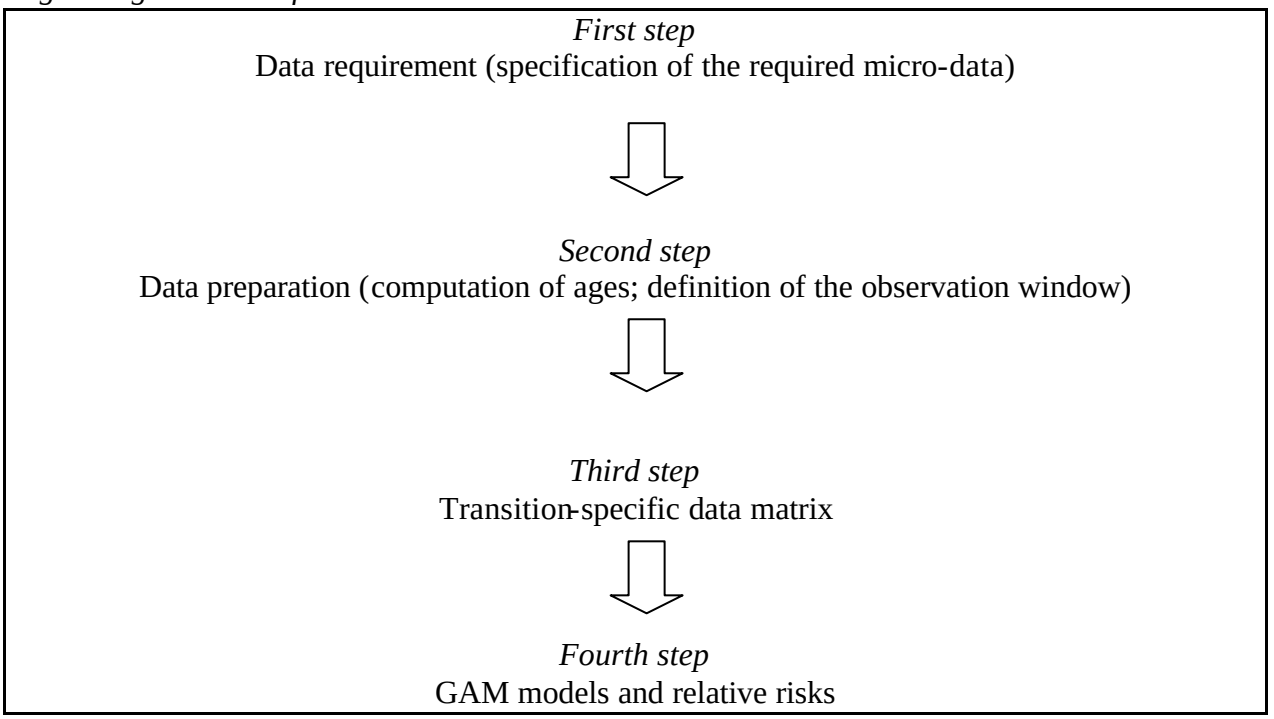
The new methodology for population forecasting that is being developed within the MicMac project does not only take into account macro demographic changes but also life course trajectories (see, e.g., Willekens, 2005; van der Gaag et al., 2006). In this framework, the life course is viewed as a sequence of states and events; each event marks a transition from one state to another state. The study of a single transition is based on the estimation of a transition rate (from the original state to the destination state). From the literature on living arrangements and fertility (see, e.g., Billari et al., 2005) we know that transition rates vary with age. Indeed, such variation with age has been traditionally exploited in demographic forecasting.

The present paper proposes a general method for the calculation of age profiles for the main transitions experienced by individuals, as far as living arrangement and fertility behaviours are concerned. Our analytical strategy starts from micro data, on the assumption that such data have to be collected allowing for a longitudinal (at least in a retrospective sense) reconstruction of the life course. This means that our method requires data that allow to reconstruct the biography of the individual and, for this purpose, all the dates of the more important events must be collected. Longitudinal data can be obtained both from retrospective surveys and from panel surveys. Here, we only refer to the former case. *Retrospective surveys* concerning family and fertility behaviours are rather available for most European countries (e.g., from the Fertility and Family Survey project, from the Generations and Gender Project and from other data collection ventures based on National Statistical Offices). Retrospective surveys permit to collect a wide range of information relating to the past experience of individuals with limited costs. On the other hand, the attention is limited uniquely to the survivors since we do not have information about deaths nor about emigrated individuals. However, this feature is not necessarily a disadvantage in methodological terms since we consistently reduce the number of events that drives the individual out of the observed sample over time, simplifying the calculation of transition rates. In the literature on demographic microsimulation, biographic information collected in retrospective surveys has often been used (see, e.g., Wachter et al., 1998).

The method can be segmented in four steps. The first step consists in the specification of the required micro-data for the evaluation of the age profiles for all the transitions relating to living arrangement and fertility presented in the Deliverable D2 “Report on data input requirement of MAC” (de Beer et al., 2006). In Section 2 we indicate all the information that the starting data files should contain in order to apply the method of calculation.
The second step relates to data preparation and includes the computation of ages at various events and the definition of the window of observation that we will consider for each individual. The third step consists in the specification of the transition-specific data matrix: within the window of ages, the events experienced and the time spent in a specific state contribute to the calculation of events and time of exposure for all the individuals in the sample. Starting from a data file where each record is a set of dates and status variables relating to the i-th individual, we obtain a matrix for each transition where the single row refers to a specific age x and containing the number of events and time of exposure. Second and third step are discussed in Section 3.
In the last step, presented in Section 4, we analyze how an observed set of events and time of exposure can be modelled by a smoother function, obtaining age profiles. In particular, we

develop GAM (Generalized Additive Model) that permits to evaluate the smoothed age profile as the transition rate baseline and, at the same time, to evaluate the effects of a vector of covariates as multiplicative changes from the baseline.
Finally, in Section 5 and 6 the method is applied to Italy and the Netherlands. As an example, we show the results relating to the transition "never married-first marriage".

Fig. 1 Organization of the method.



The data used in this report are the following:

- the ISTAT survey called "Famiglia e soggetti sociali" (FFS-IT) conducted in Italy at the end of 2003
- the Fertility and Family survey for Netherlands (FFS-NL) conducted between February and May of 2003.

The characteristics of the information contained in these datasets require some specific adaptation to the general indications given in Section 2. Besides, some transitions could not be considered because of missing information in the data set (for example, in both surveys there are no questions concerning the return in the parental home after a temporary exit) or because some transitions are very rare, thus implying a too small number of cases for the method to be correctly applied. For each transition that we are able to study, we report all the needed information for the analysis and, hence, we examine more deeply what has been discussed in Section 2. The resulting age profiles are plotted and included in this report. All the passages of the method are developed in R, a very suitable statistical free and open source software.

2. Input data requirements

This section describes the data required for the calculation of age profiles on the transitions between various living arrangements and concerning fertility. Considering micro level data originated from a retrospective survey, in the starting data file we need the individual complete biography concerning family and fertility behavior. More specifically, we assume to have information on the date of birth, the date of the interview and the date of each event experienced within the range considered. All dates are expected to be available on a monthly time scale, i.e. to be expressed in calendar month (format MM: 1 to 12) and year (format YYYY).

Moreover, for each specific transition, we need a *status* variable at the interview that is, for a generic transition *TR*:

- 0 if the individual has never experienced *TR* at the time of the interview
- *1* if the individual experienced *TR* before the interview
- *9* if the case is not applicable, i.e. the individual has never been at risk to experience *TR.*

As an example, we consider the transition *TR*="first child→second child". The status variable is 1 if the second child was born before the interview; is 0 if the individual has only one child; is 9 if the individual is still childless at the interview.

2.1 Marital status

The *state space* for marital status is composed by four states:

- never married
- married
- divorced
- widowed

These categories imply that we shall not consider explicitly the order of marriage. That is, the behavior of persons in their second marriage cannot be distinguished from persons in their first marriage. The qualitative shape of the transition matrix is as follows:

From \ to	Never married	Married	Divorced	Widowed
Never married	O	TR1		
Married		O	TR2	TR3
Divorced		TR4	O	
Widowed		TR5		O

"TRX" *transition between categories*
"empty space" *impossible event*
"O" *non-event*

All these transitions are experienced when one of these events (marriage, divorce, death of spouse) occurs. In the starting data file we need dates for these events (if any) and 5 status variables at the interview:

- date of first marriage
- date of second marriage

- date of divorce
- date of death of spouse
- status for TR1 (first marriage)
- status for TR2 (divorce)
- status for TR3 (death of spouse)
- status for TR4 (second marriage after a divorce)
- status for TR5 (second marriage after death of spouse)

2.2 Living arrangement

The choice of possible states for the living arrangements strongly depends on the characteristics of the country examined. As a general rule (see de Beer et al., 2006), we could consider the following states: living with parents, living without a partner, living with a partner, living with other persons, living in an institution. However, some of these states and/or transitions are relatively rare in a specific country. This is, for example, the case of Italy, where the distinction between living alone (without a partner), with other persons, and in institution cannot be adequately studied. Therefore, we prefer to aggregate some states in order to have an adequate number of cases. In detail, we consider: living with parents, living with a partner, living alone or with other people. The qualitative shape of the transition matrix is:

From \ to	at parental home	with a partner	alone or with other persons
at parental home	O	TR6	TR7
with a partner	TR8	O	TR9
alone or with other persons	TR10	TR11	O

"TRX" *transition between categories*
"empty space" *impossible event*
"O" *non-event*

We need all dates of the events that cause these transitions and a status variable for each transition:

- date of exit from parental home
- date of (first, second, third) return into parental home
- date of the beginning of (first, second, third) union
- date of the end of (first, second, third) union
- status for TR6 (exit from parental home with a partner)
- status for TR7 (exit from parental home alone or with other persons)
- status for TR8 (return into parental home after a union)
- status for TR9 ("alone or with other persons" after a union")
- status for TR10 (return into parental home after "alone or with other persons")
- status for TR11 (union after "alone or with other persons")

Here we need to underline that these transitions could be experienced twice or more by the same individual (for example, one can experience T6, then TR9, then TR10 and T6 again). If we have

a very detailed biography that gives us all the needed information, we can consider more than one record for each individual in the data set if he/she experienced a certain transition more than once. However, it results very difficult to rely on the complete set of information mentioned above. As we will see in the case of Italy, we are obliged to totally exclude some transitions and limit the others only at the first experience (for example, we can consider only the first exit from parental home).

2.3 Fertility (number of children ever born)

Women and men are distinguished by the number of children ever born. The possible states are then childless, 1 child, 2 children, 3 children, 4 or more children. The qualitative shape of the transition matrix is:

From \ to	childless	1 child	2 children	3 children	4+ children
Childless	O	TR12			
1 child		O	TR13		
2 children			O	TR14	
3 children				O	TR15
4+ children					O

"TRX" *transition between categories*
"empty space" *impossible event*
"O" *non-event*

Transition such as 0 → 2, 1→3, etc. caused by multiple births are not taken into account. A childless woman who has a twin birth simply experiences the transition 0→1 and 1→2 at the same date.
The *i*-th status variable accounts for the *i*-th child even born at the time of the interview (is 1 if the woman gave birth to at least *i* children, 0 if the woman gave birth to only *i-1* children, 9 if the woman has less than *i*-1 children). Evidently, for TR12 the status variable can only take value 0 or 1.
Then, we need:

- date at first birth
- date at second birth
- date at third birth
- date at fourth birth
- status for TR12 (first child)
- status for TR13 (second child)
- status for TR14 (third child)
- status for TR15 (fourth child)

2.4 Children in the household

Differently from the previous section, now we consider the presence of children in the parental home, thus not simply the births. Given that in many applications, a dichotomy is sufficient (see

D2), we consider two states: with and without children. The qualitative shape of the transition matrix is:

From \ to	Without children	With children
Without children	O	TR16
With children	TR17	O

"TRX" *transition between categories*
"empty space" *impossible event*
"O" *non-event*

The events that cause transition TR16 are the birth of the first child or the entry into a relationship with a partner who has one or more children. Transition TR17 is experienced when the last child exits from parental home (or dies) or following a divorce or separation causing children living at ex-partner's.
The status variables are dichotomous variables assuming value 0 when no event that causing transitions is experienced before the interview, and 1 otherwise.
Briefly, we need:

- date of entry into the status "with children"
- date of exit from the status "without children"
- status for TR16 (without children→ with children)
- status for TR17 (with children → without children)

2.5 Structure of the initial data set

The initial dataset should contain all the dates and variables specified in the previous sections. Then, the structure of records should be as follow:

ID	*Weight*	*m.birth*	*y.birth*	*m.inter*	*y.inter*	*m.event 1*	*y.event1*	*m.event 2*	*y.event2*	….
Identification number	weight	Month of birth	Year of birth	Month of interview	Year of interview	Month of event 1	Year of event 1	Month of event 2	Year of event 2	….

….	*TR1*	*TR2*	…	*TR17*	*Sex*	*Edu*
….	Status for transition TR1	Status for transition TR2	…	Status for transition TR17	Sex	Level of educational attainment

3. Data preparation

Given the initial dataset, the first step towards the estimation of age profiles is the calculation of exact ages in which each event occurred. Then, we must define a *window of observation* for our data. Our final aim is the estimation of the transition rates that will be used as input in the models

developed within MicMac for population forecasting. Nevertheless, we easily understand that in order to forecast, or "project", individual behaviors in the future, we need to start from recently observed behaviors and, in case, manipulate the outcomes of such observations. Since we are working with retrospective data, we have information on the whole biography of individuals interviewed at all ages, including events (and transitions) experienced many years ago. This is why we limit our observation to the more recent past, excluding all the events that happened before. The choice of this window of observation is evidently subjective. The third step is the computation of a new status variable, that we can call *censor* and that tells us if the individual must be considered in the analysis, if she/he has experienced the specific transition in the fixed window of observation or if it is left censored.

Through ages and status variable, we have all we need for the effective calculation of relative risks. Depending on the type of observational plan, transition rates can be distinguished in *period rates*, based on period-age observations (calendar year in which an event occurs and the age at the time of event), *cohort rates* based on cohort-age observations (cohort to which a person belongs and the age in calendar years) and *period-cohort rates* based on period-cohort observation (calendar year in which the event occurs and the cohort to which the person belongs). Since we are using retrospective data, we use cohort rates for which the observation period extends over two calendar years and the age is in completed years. The last step is the determination of a smoother function for the observed age profile.

3.1 Transformation of dates into ages

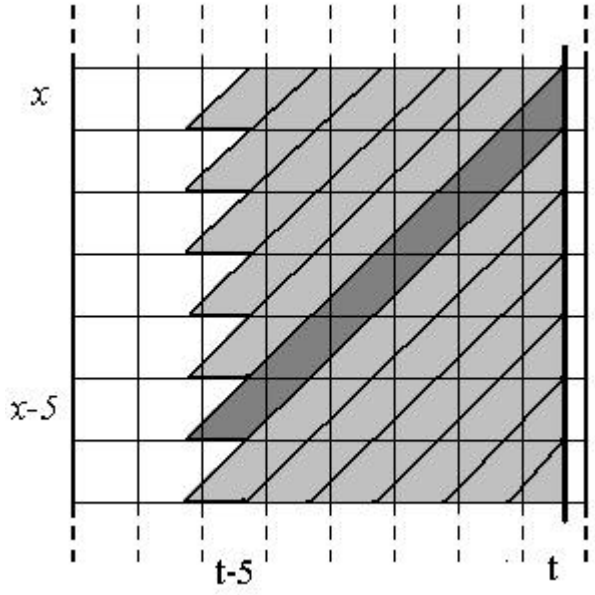
If we have, for each individual, the date of a specific event in the format month (MM: 1 to 12) – calendar year (YYYY), we can calculate the exact age at any event with a two-digits precision. First of all, we have to transform a date from the format month-year in a decimal expression considering the number of years and fractions of years since 1 January 1900. Then, the exact age at a specific event is given by the difference between this decimal expressions and the decimal expression of the date of birth. For example, if first marriage is experienced on November 1993 (decimal expression=93.88) and the individual was born on May 1965 (decimal expression=65.38), the exact age at first marriage is 38.50 years.

3.2 The window of observation

All the information on the events in the living arrangement and fertility fields, are collected through surveys based on interviews to respondents aged at least 15 or 18 years at the interview. This is consistent with the dynamics of such behaviors in contemporary Europe (see, e.g., Billari et al., 2005). From now on, we consider all the individuals aged at least 18 years at the interview. We have already stressed the necessity to focus our attention only on the more recent past. A plausible period could be the last five years before the interview. Then, we can consider all the events experienced within the age interval $[x\text{-}5, x]$, where x is the exact age at the interview. However, as we have already underlined, the type of data used requires transition rates calculated as cohort rates (based on cohort-age observations) and then it seems reasonable to consider the beginning of the window of observation at the $(x\text{-}5)$-*th* birthday. For example, if an individual were interviewed on November 2003 at the exact age 40.35, the window of

observation starts at the 35th birthday and ends at 40.35 years. Fig. 1 shows the windows of observation on the Lexis diagram.

Fig. 2. The window of observation in the Lexis diagram. Individuals are interviewed in a precise point in time during year t .



3.3 Episode limits and censoring

From this point onward, the procedure is transition-specific and it has to be repeated for each transition. Once we know the window of observation, we can define the limits of the episode that must be included in the window. The episode is the time interval spent in the initial state (the individual is at risk to experience the transition) and it is delimited by a starting age and a final age.

The starting age could be the age at the beginning of the window of observation (*agesw*) or a higher age, in case the individual enters the initial states after *agesw*. For example, in the transition TR1 (never married → married), the starting age is always the beginning of the window of observation, whereas for TR2 (married → divorced) the maximum between age at marriage (if any) and *agesw* (age at the beginning of the window of observation).
The age at the end of the episode is the minimum among the following ages:

- the age at the interview if the case is censored at the interview (*ageint*);
- the age at the considered event, if the individual experienced the transition (*ageev*);
- the age at any other event that provokes the exit from the observation (*agecens*).

The next step is the computation of a new status variable called *censor,* which indicates if the individual has to be included in the analysis, if it is censored or not and, in case, the kind of censoring (interview or other events). More in detail, for a generic transition TR*X*, *censor* assumes the following values:

0 censored at the interview
1 the event is experienced in the window of observation
2 the individual experienced another event in the window of observation that causes the exit of observation. Since we consider retrospective data, we do not have deaths and migrations but a competing event could cause a censoring. For example, in the transition TR1, death of spouse causes the exit from observation.
9 the individual cannot be considered in the analysis. This situation emerges when the transition or an event that causes the exit of observation are experienced before *agesw*. This means that in the period considered, the case cannot contribute to the time of exposure.

3.4 Transition-specific data matrix

In order to estimate transition rates, the number of events and time of exposure have to be measured. For every transition, we consider single years of age from 0 to 100+. Considering a generic transition TR*X* from a state *A* to a state *B*, for the *j-th* individual we have a window of observation included between $x_j^{IN\_WIN}$ and $x_j^{FIN\_WIN}$, and an episode that starts at the age x_j^{in} and ends in x_j^{fin}.
In general, we have that

$$\underbrace{x_j^{IN\_WIN} \le \overbrace{x_j^{in} \le x_j^{fin}}^{episode} \le x_j^{FIN\_WIN}}_{Window\ of\ observation}$$

The *transition rate* r_x at age x is the ratio between the number of events E_x experienced at age x and the amount of time spent in the initial states (time of exposure PY_x) by individuals at the same age. Therefore, the next step is the computation of E_x *and* PY_x for every age x. Considering N as the number of individuals, we have:

$$E_x = \sum_{j=1}^{N} E_{j,x}$$

$$PY_x = \sum_{j=1}^{N} PY_{j,x}$$

where, for the j-th individual:

$$E_{j,x} = \begin{cases} 1 & \text{if } x_j^{in} < x < x_j^{fin} \text{ and the transition is experimented at the exact age x} \\ 0 & \text{otherwise} \end{cases}$$

$$PY_{j,x} = \begin{cases} 1 & \text{if } x_j^{in} < x < x_j^{fin} \text{ and neither transition nor exit from observation are experienced at the age x} \\ d_{\text{j,x}} & \text{if } x_j^{in} < x < x_j^{fin} \text{ and transition or exit from observation is experienced at the age x} \\ 0 & \text{if } x < x_j^{in} \text{ or } x > x_j^{fin} \end{cases}$$

where $d_{j,x}$ is the fraction of year spent in the initial state at the exact age in which the individual experiences the transition or the exit from observation. For example, let us suppose that in the transition (married → divorced) the *j*-th individual has an episode that start at the 40th birthday and ends at the exact age 42.31 with a divorce, for the age 41 his contributions for the event is 0 and his contribution for the time of exposure is 1; for the age 42 he contributes 1 event and 0.31 years to the time of exposure.

We can also include individual post-stratification weights w_j in the computation. The formulas become:

$$E_x = \sum_{j=1}^{N} E_{j,x} \cdot w_j$$

$$PY_x = \sum_{j=1}^{N} PY_{j,x} \cdot w_j$$

It is also possible to take into account some covariates by computing events and time of exposure separately for any combination of two or more categorical variables. In order to do so, we need to select sub-samples (defined for each combination of levels) to which apply the previous calculations. For example, we can consider two *timed-fixed* covariates, in the sense that their values remain fixed for the whole window of observation, sex and level of education, coded as follow:

Sex: 1 Men; 2 Women.
Level of educational attainment: 1 primary education and less (ISCED 01); 2 lower secondary education (ISCED 2); 3 upper secondary education (ISCED 3-4); 4 tertiary (ISCED 5A-6, 5B).

In table 1, we show a segment of a transition-specific data matrix calculated taking into account sex and education. For the age *x* we have 8 rows, one for each combination of levels for the covariates sex and education. We have a column with the number of cases (CASES) relating to a specific row (combination of age *x* and the specific level of covariates); unweighted (EVENT and EXPOS) and weighted (EVENTW and EXPOSW) events and time of exposure.

Table 1. Transition-specific data matrix

ID	AGE	CASES	EVENT	EVENTW	EXPOS	EXPOSW	SEX	EDU
[110,]	27	46	3	2.7286948	41.83	45.99903	1	1
[111,]	27	669	41	44.4088040	596.41	622.97497	1	2
[112,]	27	783	32	34.2193542	693.11	715.09001	1	3
[113,]	27	275	6	4.9138153	252.44	269.64722	1	4
[114,]	27	21	3	3.6084344	19.27	16.23176	2	1
[115,]	27	324	36	37.1002991	280.89	262.49605	2	2
[116,]	27	607	47	46.0331352	528.09	516.12050	2	3
[117,]	27	347	31	35.4466726	306.03	312.34037	2	4
[118,]	28	45	6	7.0276129	37.83	43.06423	1	1
[119,]	28	658	46	52.2252426	579.47	602.05693	1	2
[120,]	28	717	53	58.0949265	615.20	653.72342	1	3
[121,]	28	272	7	5.3103960	250.33	265.40648	1	4
[122,]	28	23	2	1.1703283	19.18	16.29679	2	1
[123,]	28	300	34	33.5822526	258.96	245.48345	2	2
[124,]	28	528	53	54.3586481	444.91	447.92291	2	3
[125,]	28	320	17	23.9958731	285.22	291.48234	2	4

4. GAM models and transition rates

If we consider the transition rate for a specific event as the dependent variable, we should model it as a function of age and a set of covariates. However, age profiles for a specific transition should never be considered as a linear function. Smoothing or graduating rates, or more specifically the age profile of rates, has been a traditional issue in various disciplines, including demography and actuarial science. Traditional approaches based on polynomials have been criticized in the literature since a long time, with authors proposing the use of spline functions as a solution (see, e.g., McNeil et al., 1977); recent developments include Smith et al., (2004) and, on age-specific fertility rates, Schmertmann (2003).
For our purpose, a suitable solution are the so-called *Additive Models* (Hastie & Tibshirani, 1990; Chambers & Hastie, 1992; Hastie *et al*, 2001) that are a generalization of linear model where the dependent variable Y can be modeled as a sum of non-linear (smoother) functions.
The model structure is

$$g(m) = b_0 + f(age) + \sum_k b_k X_k \qquad (1)$$

where

$m = E(Y)$ and $g(.)$ is the link function

and Y is the response variable (distributed as some exponential distributions); X_k is a generic covariate and $ß_k$ the corresponding parameter; $ß_0$ is the intercept and $f(age)$ is the smooth function of age.
Since transition rates at age x for a specific event is given by the ratio between number of events (*Events*) and the time of exposure (*Exp.time*), considering natural logarithm as link function, for each i-th row of data matrix[3] we can write:

[3] We remember that each row of the data matrix is given by a specific combination of age x and the levels of categorical covariates.

$$\ln\left(\frac{Events_i}{Exp.time_i}\right) = b_0 + f(age_i) + \sum_k b_k X_{ki} + e_i$$

where e_i is a random error term. Then,

$$\ln(Events_i) = \ln(Exp.time_i) + b_0 + f(age_i) + \sum_k b_k X_{ki} + e_i$$

or, considering the expected value

$$\ln(E[Events]) = offset[\ln(Exp.time)] + b_0 + f(age) + \sum_k b_k X_{ki} \quad (2)$$

where $Events \sim Poisson$.

It is important that the term ln(*Exp.time*) has no coefficient to be estimated.
The smooth function f is a *piecewise cubic spline*, a curve made up of sections of cubic polynomial joined together so that they are continuous in value, as well as first and second derivatives. The points at which the sections join are known as the *knots* of the spline, that are placed at quantiles of the distribution of unique *x* values. The number of knots defines the *degree of smoothness* (i.e. number of knots + 2).
In order to avoid the choice of the is parameter, that is essentially arbitrary, the degree of smoothness of the f is estimated by Generalized Cross Validation[4] (Wood, 2006). The *mgcv* package contains a GAM implementation in which the degree of smoothness of model terms is estimated as part of fitting (see Wood, 2006). Calling $\hat{y}$ the fitted values of this model, the transition rate will be estimated as

$$\hat{r} = \frac{\hat{y}}{Exp.time} \quad (4)$$

We estimate relative risks for each transition separately for men and women. The covariates in the models are:

1. level of education at the interview (considered constant throughout the window of observation)
2. be married (Yes/No) (*time-varying*)
3. having children in the household (Yes/No)

A time-varying variable is obtained by splitting the episode at the point where the event occurs (see Blossfeld and Rohwer, 2002). Each sub-episode will be characterized by a unique value of

[4] The way to control smoothness by altering the basis dimension, is to keep the it fixed at a size a little larger than it is believed could reasonably be necessary, but to control the model's smoothness by adding a "wiggliness" penalty to the least squares fitting objective (penalized regression spline). (Wood, 2006).

the variable. For example, let's consider the transition to the first child and an episode that starts at 22 and ends at 25.83 years of age. If the individual married at 23.24 years, then the episode is splitted into the two sub-episodes (22; 23.24) and (23.24; 25.83). The covariate "be married" has value "No" in the first sub-episode and "Yes" in the second.
The variable "having children in the household" requires a particular procedure: we must account for the entry of the first child in the household and for the exit of the last child. Therefore, if needed, we will split our subinterval twice.

We must underline three crucial aspects emerging from this approach:

a. In our dataset, *Events* are calculated starting from individual weighted information. As a consequence, number of events and time of exposure are not integers. Since the *Poisson* distribution is defined only for integers, we need to round the number of weighted events. Empirical analyses (here not shown) suggest that this approximation appears acceptable.

b. The effect of covariates should be considered as differences from the grand mean, i.e. from the mean risk for the whole sample[5]. Therefore, we use the "deviation coding" system that permits to compare the mean of the dependent variable for a given level to the overall mean of the dependent variable. If we consider, for example, the categorical covariate *Education* with 4 levels (primary school, lower secondary school, upper secondary school, tertiary school), the deviation coding is accomplished by assigning value "1" to level 1 for the first comparison (because level 1 is the level to be compared to all others), to level 2 for the second comparison (because level 2 is to be compared to all others), and to level 3 for the third comparison (because level 3 is to be compared to all others). The value "-1" is assigned to level 4 for all three comparisons (because it is the level that is never compared to the other levels). The value "0" is assigned to all other levels (See table 2).

Table 2. Deviation coding for level of education

Level of education	Dummy 1 (Primary vs. mean)	Dummy 2 Low. sec. vs mean	Dummy 3 Upp. sec vs mean
Primary	1	0	0
Lower secondary	0	1	0
Upper secondary	0	0	1
Tertiary	-1	-1	-1

The contrast estimate gives the proportional effect to be applied to the baseline risk. Given that the expected values of the dummies specified in such a way are always zero[6], we can obtain the baseline transition rate as:

$$baseline_i = e^{b_0 + f(age_i)}$$

[5] We need this feature because in the development of the MicMac project we will use the effect of covariates estimated with micro data to the baseline age profile based on macro data.
[6] More precisely, the expected values are zero if the number of cases is (approximately) the same for each levels. In our analysis this condition is satisfied given the structure of our data-matrix (equal number of rows for each combination of levels of covariates).

c. The use of GAM models allows to include covariates in the equation and to evaluate their proportional effect on the smoothed age. Therefore, the estimated coefficients express multiplicative changes to be applied at the baseline age profile in order to evaluate the estimated risk for each year of age.
The simplest way to do so is to consider a vertical shift throughout the whole range of age. For example, in fig. 3 it is shown the multiplicative effect of the level of education on an unspecified transition.
However, very often the effect of a covariate shows a combination of vertical and horizontal shifts. In order to take into account this feature, a solution could be the estimation of the vertical shift for different specific sub-interval of age. In our analysis we split the range of age into 3 sub-intervals at two specific knots. The knots are fixed automatically at the 33rd and at the 67th percentiles (i.e. at the ages x_1 and x_2 at which, respectively, the 33% and the 67% of all the events are experienced before these ages).

The resulting model takes into account the baseline transition rate and the interaction between covariates and the age sub-intervals. The covariate deviation coding is adapted to this new feature: we evaluate the effect of each level of categorical variable within each subinterval.
In fig. 4 we can see an example of the effects of education on transition TR2 (married-divorced) obtained by dividing the age range into three subintervals.

Fig. 3 Multiplicative effects of covariates estimated with additive model.

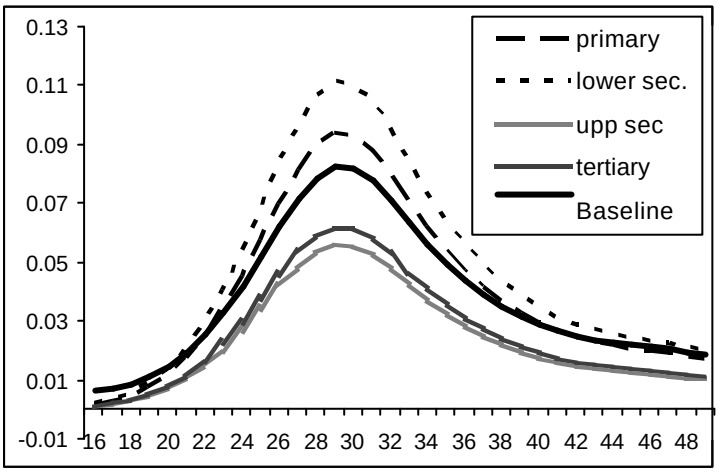
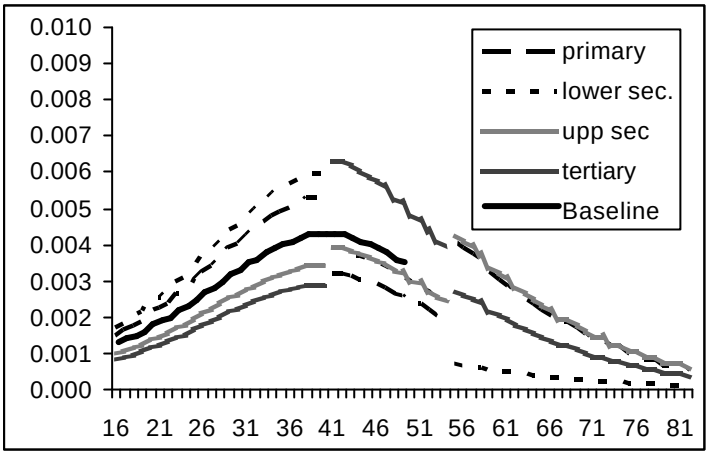


Fig. 4 Proportional effects of education on the transition TR2 (married→divorced).



Level of education. Relative risks

	15-40	41-54	55-90
primary	1.259	0.765	1.702
lower sec.	1.419	0.928	0.295
upp sec	0.812	0.936	1.774
tertiary	0.689	1.505	1.122

Finally, we can test the statistical significance of an additional covariate in the model by dropping it and noting the change in the deviance. The fitted models are compared using an analysis of deviance table. The tests are usually approximated, unless the models are un-penalized (Wood, 2006). In R this is accomplished by using the anova.gam function included in the MGCV package. We will only consider covariates that significantly increase the fit of the model when we add them into the equation.

5. Application examples in Italy and Netherlands

4.1 Data adaptations

Data for the Italian case come from the multipurpose survey called "Famiglia e soggetti sociali (FSS)". Carried out at the end of 2003, these data contain wide retrospective information on life course trajectories and the transition to adulthood, including data on the history of marital unions, cohabitations (followed by a marriage or not) and marital disruption, for a large sample of the resident population. The longitudinal nature of the survey makes it possible to update the collected information and to follow the same individual over time.
However, in this dataset we find limitations in the data available that requires some adaptation.

a. We do not have the date of birth of the respondent, but only the age (in completed years) at the interview (i.e. on 15 November 2003): an individual with the age x (say, 48) has an exact age that is in the interval (48.00, 48.99) and the date of birth is included between 16 November 1954 and 15 November 1955. This means that, even if we have dates in month and year for every event, we cannot define the exact age at which these events were experienced.
Since we do not have further information, we can avoid this indetermination adding a random month of birth extracted from a discrete uniform distribution $U(1,12)$.

b. For some events, only the calendar year has been asked. In particular, we do not have information on the month for the following events:

- exit from parental home

- divorce
- death of spouse
- exit from parental home (or death) of the last child

If the individual experienced the event, we can consider a random month for each event extracted from a discrete uniform distribution $U(1,12)$.

A remark is needed: in this case we add a second approximation to the one already introduced in point a. Since the available information concerns the respondent's age in completed years at the interview and the calendar year in which the divorce occurred, the exact age at this event is included in a two-years range. For example, an individual aged 48 at the interview who divorced during 1995, could have experienced this event at an exact age included in the interval (39.12, 41.11), i.e. at 39, 40 or 41. For the month of exit from parental home we can use the additional information given by the month of marriage: if the individual experienced both exit from home and first marriage and if the two events occurred in the same calendar year, we can assume that the two events occurred simultaneously. Therefore, the month of exit from home corresponds to the month of first marriage.

It is not possible to analyze all the transitions proposed in section 2 because of lack of data or very few events (rare transition). In particular we cannot consider:

- TR8, TR9, TR10 (no information are available for the events that cause these transitions);
- TR11 (We have information on the event that causes transition (entry into a union) but not on the event that causes exit from observation (return parental home);

Transition-specific age profiles are estimated separately for men and women. Moreover, we considered the effect of educational level and the following time-varying covariates:

- married (Yes/No), included in TR12, TR13, TR14, and TR15
- with children in the household (Yes/No), included in TR1,TR2, TR4 and TR5.

In table 4, we have p-values associated with the Null Hypothesis that a specific covariate does not increase the fit of the model. In the columns marked with (1) the significance Chi-square test related to the comparison between the model without covariate and the model with education is shown; in columns marked with (2) and (3) the model with education is compared with the model adding, respectively, "married or not" and "with/without children in the household". Results obtained for women suggest to include education everywhere excepting TR4, TR5 and TR15, where the difference from the base model (without covariates) is not significant (at 95% level). For men, education could be excluded in TR2, TR3, TR4, TR5 and TR15. Time-varying covariate "married" could be excluded in TR14 and TR15 for both sex whereas "with children" should be included in TR2, TR4 and TR5 for women and TR1 and TR2 for men.

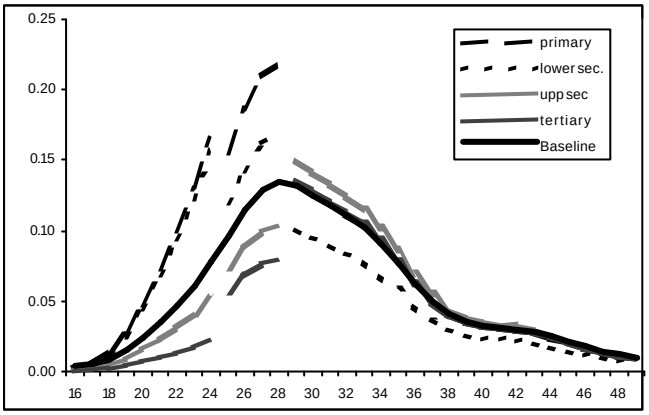
Table 4. Analysis of Deviance. Significance Chi-square test (p-value) comparing model without covariate and model with education (Column 1); model with education and model with "married or not" (column 2); model with education and model with "with/without children in the household" (column 3) by sex and transition.

	Women			Men		
	education	married	with children	education	married	with children
	(1)	(2)	(3)	(1)	(2)	(3)
TR1 never married -> married	0.000	-	0.066	0.000	-	0.000
TR2 married -> divorced	0.002	-	0.009	0.590	-	0.001
TR3 married -> widowed	0.000	-	-	0.124	-	-
TR4 divorced -> married	0.193	-	0.005	0.072	-	0.193
TR5 widowed -> married	0.084	-	0.003	0.051	-	0.428
TR6 parental home -> with a partner	0.000	-	-	0.000	-	-
TR7 parental home -> alone or with other persons	0.006	-	-	0.001	-	-
TR12 childless -> 1 child	0.000	0.000	-	0.000	0.000	-
TR13 1 -> 2 children	0.001	0.000	-	0.002	0.000	-
TR14 2->3 children	0.006	0.390	-	0.000	0.238	-
TR15 3->4 children	0.439	0.709	-	0.234	0.076	-
TR16 without children -> with children	0.000	-	-	0.000	-	-
TR17 with children -> without children	0.000	-	-	0.001	-	-

As an example of the output given by the method presented in this paper, we consider the transition TR1 (never married-married). First of all, we specify the needed information for transition rate calculations adapted to the ISTAT 2003 data. Similar indications could be defined for the other transitions. Fig. 5 shows the baseline risk and the relative risks by sex, level of education and the presence of children in the household.

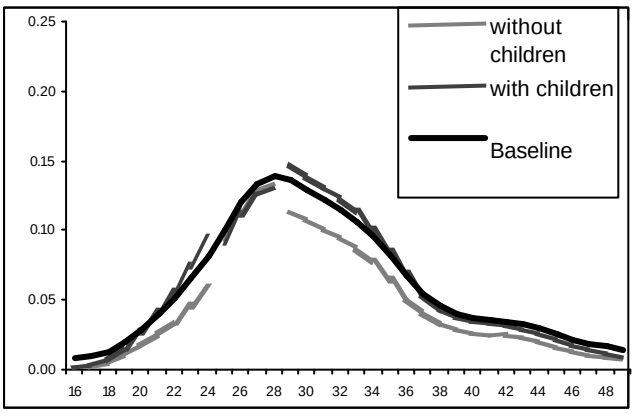
Transition	**Event**	**Events that implies the exit from observation**	**Cases in the analysis**	**Starting age of exposure**	**Final age of exposure**
TR1 (never-married → married)	First marriage	None	Who did not experience marriage before *agesw*	*agesw* (age at the beginning of the window of observation)	Age at marriage or Age at the interview

Fig. 5 Transition TR1 (never married->married). Baseline and relative risks according to sex, level of education and presence of children in the household. Italy. ISTAT 2003. (Events observed in the window of observation: 1306 women and 1279 men).



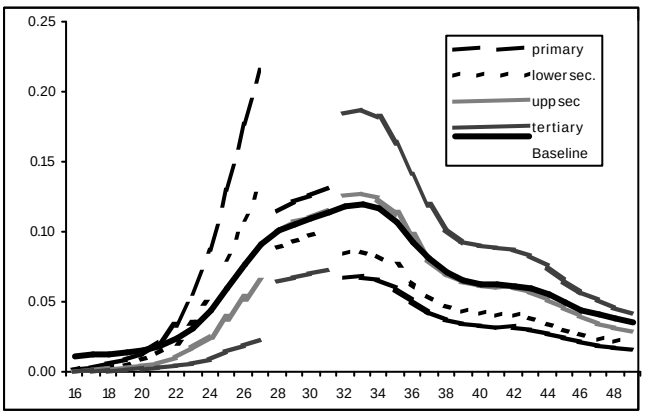
Women. Italy.
Level of education. Relative risks

	15-24	25-28	29-50
primary	2.237	1.657	1.046
lower sec.	2.108	1.273	0.779
upp sec	0.712	0.782	1.163
tertiary	0.298	0.607	1.055



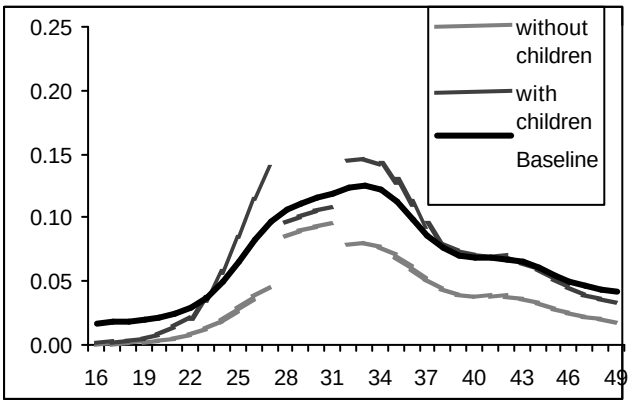
Women. Italy
Children in the household. Relative risks

	15-24	25-28	29-50
without children	0.785	1.012	0.876
with children	1.273	0.989	1.142



Men. Italy.
Level of education. Relative risks

	15-27	28-31	32-50
primary	2.694	1.277	0.624
lower sec.	1.648	0.983	0.791
upp sec	0.799	1.122	1.177
tertiary	0.282	0.71	1.721



Men. Italy.
Children in the household. Relative risks

	15-27	28-31	32-50
without children	0.573	0.939	0.739
with children	1.744	1.065	1.353

5. Age profiles of family and fertility events in the Netherlands

5.1 Data adaptations

We used data from Netherlands "Fertility and Family Survey" (FFS-NL). The 8145 interviews were carried out between February and June 2003. Since 1974, Statistics Netherlands organizes the Netherlands Fertility and Family Survey (FFS-NL) collecting longitudinal information on leaving home, cohabitation, marriage, and childbearing. Differently from the Italian data, in this case the exact date of birth is available, as well as the information on the date of each event, in the format month-year. Nonetheless, for some individuals it could happen that the information on the month is missing, since sometimes only a specification for the year of occurrence is provided. In this case, we input a random month extracted from a discrete uniform distribution $U(1,12)$.

For the Netherlands, data limitations and adaptation are similar to that needed in the Italian case. For example, we do not have all the required information for the analysis of all the transition proposed in Section 2. For the Dutch case we do not show detailed schemes for every transition since we still refer to the schemes proposed in Section 4 except for the following:

- children's dates of births are available only for women
- in the FSS-NL children's dates of exit (or death) are not available. Thus, we cannot study transitions related to the presence of children in the household. Besides, we have no information for evaluating the time-varying variable "with children in the household (Yes/No)".

Briefly, for the Netherlands we can analyze the following transitions: TR1 (never married → married); TR2 (married → divorced); TR3 (married → widowed); TR4 (divorced → married); TR5 (widowed → married); TR6 (parental home → with a partner); TR7 (parental home → alone or with other persons); TR12 (childless → 1 child) (only women); TR13 (2^{nd} birth) (only women); R14 (3^{rd} birth) (only women); TR15 (4^{th} birth) (only women).

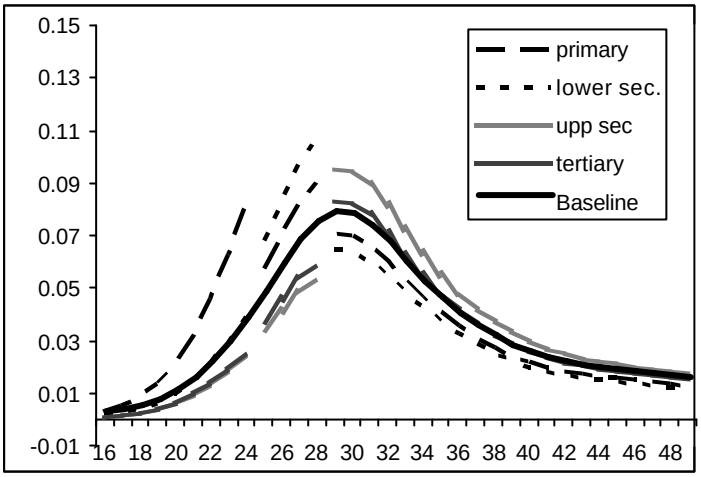
Transition-specific age profiles are estimated separately for men and women and the effect of educational level is considered. As time-varying covariates, we can consider "married (Yes/No)" and "having had first birth (Yes/No)" as an approximation of the missing covariate "with children in the household (Yes/No)".

p-values associated with the Null Hypothesis that a specific covariate does not increase the fit of the model, are shown in table 5. The meaning of this table is the same as table 4. Results suggest that, among women, education could be left out of the models in TR2,TR3, TR4; "with children" could be left out in all the three transitions that included this covariate whereas "married" could be excluded in TR14 and TR15. Among men, the introduction of education has a significant effect only for TR2 and TR7. We cannot consider time varying covariates for men since information on their child births is not available in the initial dataset.

Table 5. Analysis of Deviance. Significance Chi-square test (p-value) comparing model without covariate and model with education (Column 1); model with education and model with "married or not" (column 2); model with education and model with "with/without children in the household" (column 3) by sex and transition.

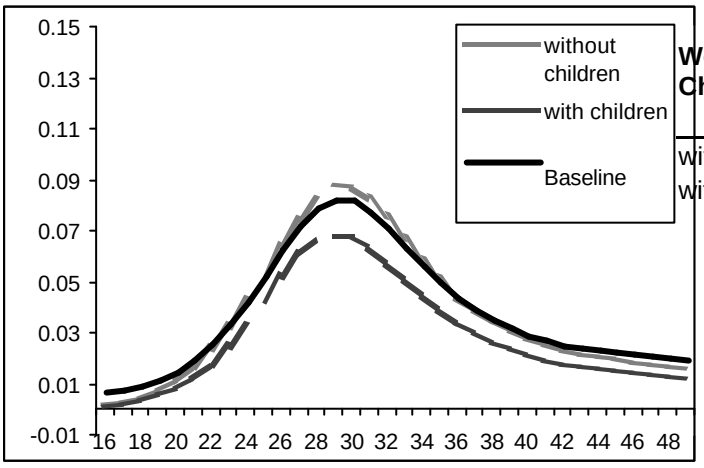
	Women			Men		
	education	married	with children	education	married	with children
	(1)	(2)	(3)	(1)	(2)	(3)
TR1 never married -> married	0.006	-	0.687	0.213	-	-
TR2 married -> divorced	0.67	-	0.36	0.046	-	-
TR3 married -> widowed	0.822	-	-	-	-	-
TR4 divorced -> married	0.269	-	0.413	0.241	-	-
TR5 widowed -> married	-	-	-	-	-	-
TR6 parental home -> with a partner	0.024	-	-	0.843	-	-
TR7 parental home -> alone or with other persons	0.000	-	-	0.000	-	-
TR12 childless -> 1 child	0.000	0.000	-	-	-	-
TR13 1 -> 2 children	0.000	0.000	-	-	-	-
TR14 2->3 children	0.012	0.124	-	-	-	-
TR15 3->4 children	0.011	0.446	-	-	-	-

Fig. 6. Transition TR1 (never married->married). Baseline and relative risks according to sex, level of education and presence of children in the household. Netherlands. FFS 2003. (Events observed in the window of observation: 295 women and 312 men).



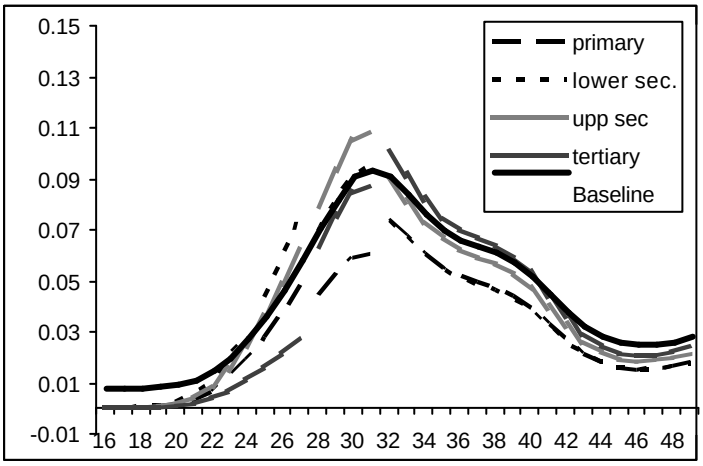
Women. Netherlands.
Level of education. Relative risks

	15-24	25-28	29-50
primary	2.223	1.212	0.913
lower sec.	1.036	1.44	0.833
upp sec	0.647	0.719	1.23
tertiary	0.671	0.797	1.069



Women. Netherlands.
Children in the household. Relative risks

	15-24	25-28	29-50
without children	1.153	1.109	1.141
with children	0.868	0.901	0.876



Men. Netherlands.
Level of education. Relative risks

	15-27	28-31	32-50
primary	0.958	0.707	0.889
lower sec.	1.529	1.103	0.879
upp sec	1.237	1.26	1.067
tertiary	0.552	1.017	1.198

Bibliographic references

Billari F.C., Philipov D., Toulemon L., 2005, *Report on fertility, family and household data, and current and future trends in fertility and household structure*, Deliverable D20, MicMac Project "Bridging the micro-macro gap in population forecasting".

Blossfeld H.P., Rohwer G. 2002. *Techniques of Event History Modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Chambers J.M. and Hastie T.J. (eds.), 1992, *Statistical models in S*, New York: Chapman and Hall.

de Beer J., van der Gaag N., Willekens F., 2006, *Report on Input Data Requirements of MAC*, Deliverable D2, MicMac Project "Bridging the micro-macro gap in population forecasting".

Hastie T.J and Tibshirani R.J., 1990, *Generalized additive models*, London: Chapman and Hall.

Hastie T.J., Tibshirani R.J. and Friedman J., 2001, *The elements of statistical learning. Data mining, inference and prediction*, New York: Springer-Verlag.

McNeil D.R., Trussell T.J., Turner J.C., 1977, "Spline interpolation of demographic data", *Demography*, 14 (2): 245-252.

Schmertmann C., 2003, "A system of model fertility schedules with graphically intuitive parameters", *Demographic Research*, 9 (5): 81-110.

Smith L., Hyndman R.J., Wood S.N., 2004, "Spline interpolation for demographic variables: the monotonicity problem", *Journal of Population Research*, 21(1): 95-98.

van der Gaag N., de Beer J., Ekamper P., Willekens F., 2006, *Using MicMac to project living arrangements: an illustration of biographic projections*, paper presented at the European Population Conference, Liverpool.

Venables W.N. and Ripley B.D., 1997, *Modern applied statistics with S-plus*, New York: Springer.

Wachter K.W., Blackwell D., Hammel E.A., 1998, *Testing the Validity of Kinship Microsimulation: An Update*, University of California, Berkeley, CA.

Willekens F., 2005, "Biographic forecasting: bridging the micro-macro gap in population forecasting". *New Zealand Population Review* 31 (1): 77-124.

Wood S.N, 2006, *Generalize Additive Models. An introduction with R*, Chapman and Hall/CRC

*****