Bridging the micro-macro gap in population forecasting

Contract no. SP23-CT-2005-006637

Deliverable D2

Report on Input Data Requirements of MAC

Work Package 1
Multistate Methods

May 2006

Authors:
Joop de Beer
Nicole van der Gaag
Frans Willekens

Netherlands Interdisciplinary Demographic Institute
P.O. Box 11650
2502 AR the Hague
The Netherlands

Preface

In an ageing population, the demand for adequate social protection systems is paramount. The sustainability of high-quality health care and pension systems, for example, is influenced to a considerable extent by demographic change and by the way people live their lives. Adequate monitoring and forecasting of demographic change and of the lifestyle and life course of the population, therefore, are indispensable for the provision of health and social security to the people of Europe. What is needed is a methodology that complements demographic projections with projections of the way people live their lives. For this purpose, in the autumn of 2003, within the 6th Framework Programme, Area 8.1 Policy-oriented research, Scientific Support to Policies (SSP), the European Commission launched a call for tenders for a research project entitled 'Development and testing of an innovative methodology for demographic projections'.

In January 2004, a consortium of European research institutes, coordinated by the Netherlands Interdisciplinary Demographic Institute, submitted a proposal to carry out this study. In the summer of 2004, the European Commission in principle approved this proposal, however, some revisions had to be made to the original research plan and the corresponding budget estimate. In the spring of 2005 a final agreement was reached between the European Commission and the Netherlands Interdisciplinary Demographic Institute (NIDI, The Hague, The Netherlands) as coordinator, as well as the following contractors: Vienna Institute of Demography (VID, Vienna, Austria), Institut National d'Etudes Demographiques (INED, Paris, France), Bocconi University (BU, Milano, Italy), Erasmus Medical Centre (EMC, Rotterdam, The Netherlands), Max Planck Institute for Demographic Research (MPIDR, Rostock, Germany), the International Institute for Applied System Analysis (IIASA, Laxenburg, Austria) and the University of Rostock (UROS, Rostock, Germany).

On the 1st of May, 2005, the project MicMac, Bridging the micro-macro gap in population forecasting, officially started. 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). MicMac is scheduled to be finished on 30 April 2009. More information on MicMac is available on the website www.micmac-projections.org.

The present report constitutes deliverable 2 of the project, 'Report on input data requirements of MAC', as specified in Annex 1 to the contract 'Description of Work'. This report is part of Work Package 1 'Multistate methods' and was the responsibility of NIDI.

We gratefully acknowledge the valuable comments on previous versions of (parts of) this report made by the contractors as well as the members of the Advisory Board of MicMac.

The Hague, May 2006

Contents

MicMac
Report on Data Input Requirements of MAC

1. Introduction

MicMac aims to develop a methodology that complements demographic projections with projections of the way people live their lives. The new methodology moves beyond conventional projections of the population by age and sex and can be used to monitor and forecast demographic change as well as the lifestyles and life courses of the population. The life course is viewed as a sequence of states and events that result in transitions from one state to another. Hence the focus is on incidence (transitions) rather than prevalence.

The present report provides an overview of the variables to be included in MicMac. The choice of the variables depends on the type of analyses for which MicMac will be used. As MicMac will produce demographic projections, it is obvious that age and sex are included in the model as basic variables.

MicMac contains modules on morbidity and mortality, fertility and living arrangements, and education. The first module determines the age profile of key events in morbidity and mortality in the life course. Hence, MicMac will include the variable disability status and will take into account the covariates smoking, body mass index (BMI), educational status (as proxy for socio-economic status), and household position. The module on fertility and living arrangements includes in a consistent way both household and fertility dynamics, as well as interactions with education. Therefore, MicMac will also include marital status, living arrangement and number of children. The module on education, finally, focuses on the estimation of educational transition rates (incidence) from data on the highest level of completed education by age and sex (prevalence). In addition to educational attainment, age at leaving school together with reason for leaving (graduation or drop out) and enrolment numbers will be used.

MicMac offers a bridge between aggregate projections of cohorts (Mac) and projections of the life courses of individual cohort members (Mic). Both Mic and Mac are multistate models with transition rates as parameters. Mac focuses on transitions among functional states by age and sex; Mic addresses demographic events and other life transitions at the individual level. For the project both macro level as well as micro level data are needed. In addition to macro level data like population numbers by age, sex and marital status, surveys can be used to estimate transition rates, for instance between different categories of disability status. In principle the same data can be used for Mic as well as for Mac. The present report, however, is limited to the data input requirements for the macro level projections of cohorts and focuses on the list of variables needed.

Apart from the overview of variables to be included in MicMac, the current report also takes into account the transitions between the different categories of the

variables. In addition, it presents the basic ideas behind the database in which the data for MicMac will be stored, together with a detailed list of data specifications.

Section 2 will discuss the state space of MicMac, *i.e.* the variables and their categories as well as the transitions between the categories. Rather detailed categories of each variable will be given.[1] However, this does not imply that all categories of all variables will be included in each application of MicMac, as not all cross-classifications will be relevant for each analysis. For many analyses some aggregation of categories may be justified. This paper *does not* discuss all kinds of cross-classifications. By way of an example section 3 only discusses the classification of the composite variable *household position*, based on cross-classification of marital status, living arrangement and number of children. Section 4 will pay attention to the development of a (synthetic) database, including an overview of the macro level data to be used in MicMac. As one of the aims of the overall project is that MicMac should be open to other possible applications, section 5 discusses other variables that may be included in future approaches of the MicMac approach, *viz.* ethnicity, place of residence, labour force participation, and disease-specific health. The report ends with a summary of the main conclusions (section 6).

2. State space: variables, categories and transitions

The state space is the set of possible states a person can occupy. States are mutually exclusive and exhaustive. The states are categories of a variable, *e.g.* marital status. A combination of variables may be used, resulting in a composite variable. For example, if marital status consists of four states (never-married, married, divorced or widowed) and living arrangement of five (living in the parental home, living alone, living with a partner, living with other persons, living in an institution) the combination of both variables results in twenty states, although the number can be reduced as some categories may be excluded (e.g. married while living at the parental home). The state space does not only define the states, but also the possible transitions between states.

MicMac will include 12 variables:

1. age
2. year of birth
3. sex
4. level of educational attainment
5. age at leaving school
6. marital status
7. living arrangement
8. number of children ever born
9. number of children in household
10. disability status
11. smoking
12. body mass index (BMI)

[1] This paper does *not* discuss measurement issues, such as data sources, and quality and availability of data.

This section describes the categories of these variables and the transitions between categories of the same variable. In addition the entries into and exits from the population, due to birth, death, immigration and emigration are included. This section does not discuss cross-classifications of these variables and the corresponding transitions.

For each variable a transition matrix is given which shows the possible events, i.e. transitions between the categories of each variable or entry to or exit from the population due to birth, death, immigration and emigration (indicated by 'X'). In addition, non-events (*i.e.* persons staying in the same category) are shown by 'O'. Empty cells indicate impossible events.

2.1 Age

Age can be measured in month and years or in completed years. If age is measured in months and years it can be recoded in *Century Month Code (CMC)*. In that coding scheme, dates of events are measured in months elapsed since January 1900. The age at the time of an event is simply the difference between the CMC at the event and the CMC at birth. The exact age in years is calculated as this difference divided by 12. The age in completed years is the age at last birthday.

For estimating transition rates, numbers of events and exposure time have to be measured. Different types of transition rates can be distinguished, based on the type of observational plan. Period rates are based on period-age observations: they record the calendar year in which an event occurs and the age at the time of the event; the age is recorded in completed years (age at last birthday). Cohort rates are based on cohort-age observations: they record the cohort to which a person belongs and the age in completed years at the time of the interval. The observation period extends over two calendar years. Period-cohort rates are based on period-cohort observations: they record the calendar year in which an event occurs and the cohort to which the person belongs. This covers two age groups, and is equivalent to recording age at the beginning or the end of the period and also equivalent to recording age in period difference.

For microsimulation the cohort-age observational plan is appropriate, whereas for macro projection purposes, a period-cohort observational plan is to be preferred.

MicMac will distinguish single years of age up to 100+.

Categories
- -1
- 0
- 1
- 2
- …
- 99
- 100+.

The category '-1' refers to events taking place in the year of birth.

Transitions
Table 1 shows the possible transitions for one-year intervals:
- births → age 0
- 0 → 1, 1 → 2, …, 98 → 99, 99 → 100+
- for each age:
    - emigration
    - death
    - immigration

Table 1: Transition matrix: age										
	To									
From	0	1	2	3	…	98	99	100+	emigration	death
0		X							X	X
1			X						X	X
2				X					X	X
…									X	X
…									X	X
98							X		X	X
99								X	X	X
100+								O	X	X
birth	X									
immigration	X	X	X	X	X	X	X	X		

Note. X: entry to or exit from population; O: non-event, i.e. person stays in the same category; empty cell: impossible event

2.2 Year of birth

Categories
- before 1900
- 1900
- 1901
- 1902
- …
- last observation year

Transitions
The only transitions are entries to and exits from the population:
- births by sex
- deaths by year of birth
- immigration by year of birth
- emigration by year of birth

Table 2: Transition matrix: year of birth								
	To							
From	before 1900	1900	1901	1902	…	last observation year	emigration	death
before 1900	O						X	X
1900		O					X	X
1901			O				X	X
1902				O			X	X
…							X	X
birth						X		
immigration	X	X	X	X	X			

Note. X: entry to or exit from population; O: non-event, i.e. person stays in the same category; empty cell: impossible event

2.3 Sex

Categories
- male
- female

Transitions
In MicMac we do not assume that there are changes of sex. Therefore the only transitions are entries to and exits from the population:
- births by sex
- deaths by sex
- immigration by sex
- emigration by sex

Table 3: Transition matrix: sex				
	To			
From	man	woman	emigration	death
man	O		X	X
woman		O	X	X
birth	X	X		
immigration	X	X		

Note. X: entries to and exits from population; O: non-event, i.e. person stays in the same category; empty cell: impossible event

2.4 Level of educational attainment

Categories
The categories are based on the ISCED classification (International Standard Classification of Education) developed by UNESCO. The level of educational attainment is defined by the highest completed level of education.

In MicMac five categories will be distinguished:
- Pre-primary and primary education and less (ISCED 01)
- Lower secondary education (ISCED 2)
- Upper secondary education (ISCED 34)

- Tertiary vocational (ISCED 5B)
- Tertiary general (ISCED 5A-6)

Transitions
Most transitions occur at schooling age. At older age transitions are quite rare, as people rarely upgrade to another level. Only upward transitions are possible, as the level of educational attainment is the highest completed level.
The detailed classification yields the following transitions:
- ISCED 01 → ISCED 2
- ISCED 2 → ISCED 34
- ISCED 34 → ISCED 5B
- ISCED 34 → ISCED 5A-6
- ISCED 5B → ISCED 5A-6

In addition entries to and exits from the population have to be distinguished by education:
- births → lowest education category
- immigration by educational attainment
- emigration by educational attainment
- deaths by educational attainment.

Table 4: Transition matrix: educational attainment

	To						
From	ISCED01	ISCED2	ISCED34	ISCED5B	ISCED5A-6	emigration	death
ISCED01	O	X				X	X
ISCED2		O	X			X	X
ISCED34			O	X	X	X	X
ISCED5B				O	X	X	X
ISCED5A-6					O	X	X
birth	X						
immigration	X	X	X	X	X		

Note. X: transition between categories or entry to or exit from population; O: non-event, i.e. person stays in the same category; empty cell: impossible event

Aggregation
Different aggregations may be appropriate depending on whether the analyses refer to young or old generations. In analysing mortality and morbidity the focus will be on elderly people for whom the distinction between the two lower categories is important, as more than 60 percent of people aged 60 or over belongs to these categories. For those generations aggregation of the two higher categories may be useful, because of small numbers.

For analysing fertility the focus will be on young generations, for whom aggregation of the two lower categories may be appropriate, whereas a distinction between the upper two categories is useful, especially as the age at graduation may differ between the latter two categories, which is particularly important in analysing postponement of fertility.

Aggregation A:
- Pre-primary and primary education and less
- Lower secondary education
- Upper secondary education
- Tertiary education

Aggregation B:
- Lower secondary education and less
- Upper secondary education
- Tertiary vocational
- Tertiary general

The transitions between these aggregate categories can simply be derived from the transitions between the separate categories.

2.5 Age at leaving school

In addition to the level of educational attainment, the age at leaving school is a useful variable. Many demographic events depend on age at graduation. For example, the increase in the age at leaving school is one main cause of the postponement of the birth of the first child.

Categories
- in school (including young children not yet in school)
- age at leaving school: 17, 18, 19, etc.

Transitions
- leaving school: in school → age at leaving school
  (assuming people leave school only once during their lifetime, this is a fixed value for each person)
- births → not yet in school (included in the category 'in school')
- immigrants in school or by age at leaving school
- emigrants in school or by age at leaving school
- deaths of persons in school or by age at leaving school

Table 5: Transition matrix: age at leaving school

		To						
From		in school	age at leaving school				emigration	death
			17	18	19	…		
in school		O	X	X	X	X	X	X
age at leaving school	17		O				X	X
	18			O			X	X
	19				O		X	X
	…					O	X	X
birth		X						
immigration		X	X	X	X	X		

Note. X: transition between categories or entry to or exit from population; O: non-event, i.e. person stays in the same category; empty cell: impossible event

2.6 Marital status

Categories
- never-married
- married
- divorced
- widowed

These categories imply that MicMac will not include the order of the marriage, i.e. the behaviour of persons in their second marriage cannot be distinguished from persons in their first marriage.

Transitions
- first marriage:
    - never-married → married
- end of marriage:
    - married → divorced
    - married → widowed
- remarriage:
    - divorced → married
    - widowed → married
- entries to and exits from the population:
    - birth → never-married
    - immigration by marital status
    - emigration by marital status
    - death by marital status

Table 6: Transition matrix: marital status

	To					
From	never-married	married	divorced	widowed	emigration	death
never-married	O	X			X	X
married		O	X	X	X	X
divorced		X	O		X	X
widowed		X		O	X	X
birth	X					
immigration	X	X	X	X		

Note. X: transition between categories or entry to or exit from population; O: non-event, i.e. person stays in the same category; empty cell: impossible event

Aggregation
In some applications it is sufficient to distinguish whether or not persons are married:
- not-married
- married

The transitions can be derived from the transitions between the four marital states.

In addition to the four categories mentioned here, several other (country-) specific categories may be possible, for instance ‘registered partnership’ in the Netherlands and ‘separation’ as legal status between ‘married’ and ‘divorce’ in Italy. Work Package 5 ‘Fertility and living arrangements’ will look at these special categories in more detail and will come with suggestions on how to cope with this issue.

2.7 Living arrangement

For the variable living arrangement, different classifications of categories can be defined. The categories given below form the point of departure for this variable, however, during the course of the project a (slightly) different classification might be chosen if it turns out that for practical reasons another classification is more appropriate. This will be one of the outcomes of Work Package 5 'Fertility and living arrangements'.

Categories
- child in the parental home
- living without a partner
- living with a partner
- living with other person(s)
- living in institution

Living without a partner includes persons living alone and persons living with their child(ren). *Living with other person(s)* excludes persons living with their child(ren), but includes elderly people living with their child, if the child (or his or her partner) is the head of the household.

Transitions
Transitions between all states are possible.
- child leaving the parental home:
    - child → without a partner
    - child → with a partner
    - child → with other person(s)
    - child → in institution
- person living without a partner:
    - without a partner → with a partner
    - without a partner → with other person(s)
    - without a partner → in institution
    - without a partner → back to parental home as child
- person living with a partner:
    - with a partner → without a partner
    - with a partner → with other person(s)
    - with a partner → in institution
    - with a partner → back to parental home as child
- person living with other person(s):
    - with other person(s) → without a partner
    - with other person(s) → with a partner
    - with other person(s) → in institution
    - with other person(s) → back to parental home as child
- person living in institution:
    - in institution → without a partner
    - in institution → with a partner
    - in institution → with other person(s)
    - in institution → back to parental home as child

- entries to and exits from the population:
    - birth → child living at parental home or living in an institution
    - death by living arrangement
    - immigration by living arrangement
    - emigration by living arrangement

Table 7: Transition matrix: living arrangement							
	To						
From	at parental home	without a partner	with a partner	with other persons	in institution	emigration	death
at parental home	O	X	X	X	X	X	X
without a partner	X	O	X	X	X	X	X
with a partner	X	X	O	X	X	X	X
with other persons	X	X	X	O	X	X	X
in institution	X	X	X	X	O	X	X
birth	X				X		
immigration	X	X	X	X	X		

Note. X: transition between categories or entry to or exit from population; O: non-event, i.e. person stays in the same category; empty cell: impossible event

In general the living arrangement category 'living in institution' covers also 'other institutions', among which 'living in prisons' etc. These 'other institutions' differ widely from the 'main' institutions like nursing homes and old people's homes. Although in theory, it would be desirable if we could eliminate these 'other institutions' from all institution, in practice, this would not be possible. As this refers to a very small number of persons, however, the impact of taking into account all institutions in one category, including the 'other institutions', will be negligible.

2.8 Number of children ever born

Women are distinguished by the number of children they ever had, whether or not they are still living at the parental home.

Categories
- childless
- 1 child
- 2 children
- 3 children
- 4 or more children

Transitions
- having children:
    - first child: 0 → 1 child
    - second child: 1 → 2 children
    - third child: 2 → 3 children
    - fourth child: 3 → 4+ children
    - fifth or following children: 4+ → 4+ children

The latter 'transition' is an event which does not lead to a transition to another category, as it concerns an 'aggregate' category.
If multiple births are taken into account, also transitions such as 0 → 2, 1 → 3, etc. should be included.

- entries to and exits from population:
    - immigrating women by the number of children ever born
    - emigrating women by the number of children ever born
    - death of women by the number of children ever born

Table 8: Transition matrix: number of children ever born							
	To						
From	0	1	2	3	4+	emigration	death
0	O	X				X	X
1		O	X			X	X
2			O	X		X	X
3				O	X	X	X
4+					O	X	X
birth	X[2]						
immigration	X	X	X	X	X		

Note. X: transition between categories or entry to or exit from population; O: non-event, i.e. person stays in the same category; empty cell: impossible event

2.9 Number of children in household

This variable refers to parents, distinguished by the number of children still living in the parental home, *not* to the number of children ever born.

Categories

- no children
- 1 child
- 2 children
- 3 children
- 4 or more children

Transitions

- having children:
    - first child: 0 → 1 child
    - second child: 1 → 2 children
    - third child: 2 → 3 children
    - fourth child: 3 → 4 or more children
    - fifth or following children: 4+ → 4+ children

  (if multiple births are taken into account, also transitions such as 0 → 2, 1 → 3, etc. should be included)
- enter into relationship with a partner who has child(ren): 0 → 1, 0 → 2, etc. children

[2] The row 'births' refers to the new-born children themselves, not to their mothers. A birth leads to a transition of the mother to another category (0 → 1, 1→ 2, etc.), but the new-born child itself enters the category 0.

(if it would be assumed that the person him- or herself had already one or more children, other transitions would be possible, such as 1 → 2, 1 → 3, 2 → 3, etc.)
- child leaving the parental home or death of a child : 4+ → 3, 3 → 2, 2 → 1, 1 → 0 (assuming only one child to leave the parental home within the time interval considered)
- entries to and exits from population:
    - immigrants who enter the country together with one or more children or who will live together with a partner with one or more children
    - emigrants who leave the country together with one or more children
    - death of person living together with one or more children

Table 9: Transition matrix: number of children in household							
	To						
From	0	1	2	3	4+	emigration	death
0	O	X	X	X	X	X	X
1	X	O	X			X	X
2		X	O	X		X	X
3			X	O	X	X	X
4+				X	O	X	X
birth	X[3]						
immigration	X	X	X	X	X		

Note. X: transition between categories or entry to or exit from population; O: non-event, i.e. person stays in the same category; empty cell: impossible event

Aggregation

In many applications, a dichotomy is sufficient:
- without children
- with children

Transitions in case of dichotomy with / without children
- without children → with children:
    - birth of first child
    - enter into a relationship with a partner who has child(ren)
- with children → without children:
    - last child leaving the parental home
    - divorce or separation after which children will live with ex-partner
    - death of child
- entries to and exits from population:
    - immigrants who enter the country together with one or more children
    - emigrants who leave the country together with one or more children
    - death of person living together with one or more children

[3] The row 'births' refers to the new-born children themselves, not to their parents. A birth leads to a transition of the parents to another category (0 → 1, 1→ 2, etc.), but the new-born child itself enters the category 0.

Table 10: Transition matrix: presence of children in household				
	To			
From	without children	with children	emigration	death
without children	O	X	X	X
with children	X	O	X	X
birth	X[4]			
immigration	X	X		

Note. X: transition between categories or entry to or exit from population; O: non-event, i.e. person stays in the same category; empty cell: impossible event

2.10 Disability status

Disability status is measured by self-reported or performance-based disability[5]. As currently it will be difficult to distinguish between levels of disability, to start with, only the distinction between disabled and non-disabled will be made:

Categories
- disabled
- non-disabled

Transitions
- incidence: non-disabled → disabled
- remission: disabled → non-disabled
- disabled → death
- non-disabled → death
- entries to and exits from the population:
    - new-born children by disability status
    - deaths by disability status
    - immigration by disability status
    - emigration by disability status

Table 11: Transition matrix: disability status				
	To			
From	non-disabled	disabled	emigration	death
non-disabled	O	X	X	X
disabled	X	O	X	X
Birth	X	X		
immigration	X	X		

Note. X: transition between categories or entry to or exit from population; O: non-event, i.e. person stays in the same category; empty cell: impossible event

Although we start with the distinction between disabled and non-disabled only, the MicMac model will include at least three categories to be able to include some further

[4] See note 3.

[5] During the MicMac project, Work Package 4 'Morbidity and mortality' will follow the developments of MHADIE (Measuring Health And Disability In Europe) and SHARE (Survey of Health, Ageing and Retirement in Europe). Given the incompatable time schedules of the different projects, however, it is unlikely that (new) developments of MHADIE and SHARE can be included in the current work of MicMac.

distinction in the future, for instance non-disabled, minor-disabled and severe-disabled[6].

2.11 Smoking

This variable should indicate at least whether someone smokes or ever smoked. Preferably more information on his or her smoking-history (timing of transitions) should be included, but for the time being that is unfeasible. In order to be able to include more detailed information in the future, however, in the state space the category 'current smokers' will be divided into 'heavy smokers' and 'incidental smokers'; to start with we will only distinguish never, ever and current smokers.

Categories
- never
- ever
- current

Transitions
If data allow it, transitions will be modeled. Otherwise prevalence of smoking will be included as a covariate.
- never → current
- ever → current
- current → ever
- entries to and exits from the population:
    - births → never
    - deaths by smoking history
    - immigration by smoking history
    - emigration by smoking history

Table 12: Transition matrix: smoking

	To				
From	never	ever	current	emigration	death
never	O		X	X	X
ever		O	X	X	X
current		X	O	X	X
birth	X				
immigration	X	X	X		

Note. X: transition between categories or entry to or exit from population; O: non-event, i.e. person stays in the same category; empty cell: impossible event

2.12 Body mass index

The variable body mass index (BMI) equals body weight divided by the squared body height (kg/m^2). The following categories will be distinguished:

[6] In Work Package 4 'Morbidity and mortality' existing disability projections of the OECD will be studied. If these projections involve other categories than disabled and non-disabled, MicMac will include in the state space the same categories as used in these projections.

Categories
- underweight: BMI<18.5
- normal weight: 18.5 ≤ BMI < 25
- overweight: 25 ≤ BMI < 30
- obesity BMI ≥ 30

Transitions
If data allow it, transitions will be modeled. Otherwise prevalence of obesity will be included as a covariate.
- gaining weight:
    - underweight → normal weight
    - normal weight → overweight
    - overweight → obesity
- losing weight:
    - normal weight → underweight
    - overweight → normal weight
    - obesity → overweight
- entries to and exits from the population:
    - births by body mass index
    - deaths by body mass index
    - immigration by body mass index
    - emigration by body mass index

Table 13: Transition matrix: Body Mass Index

	To					
From	underweight	normal weight	overweight	obesity	emigration	death
underweight	O	X			X	X
normal weight	X	O	X		X	X
overweight		X	O	X	X	X
obesity			X	O	X	X
birth	X	X	X			
immigration	X	X	X	X		

Note. X: transition between categories or entry to or exit from population; O: non-event, i.e. person stays in the same category; empty cell: impossible event

3. Aggregation and cross-classifications: household position

Cross-classification of these twelve variables results in a very large number of categories. Therefore in many cases cross-classifications will be based on some aggregation over categories. This section discusses one example.
On the basis of the three variables marital status, living arrangement and the number of children in the household the composite variable *household position* can be derived. Including all categories of the three variables in the cross-classification would result in a very large number of categories of household composition (100 categories) and an even much larger number of transitions. Hence some aggregation will be needed, but the choice which aggregation is appropriate may differ between analyses. For example, for some analyses a distinction between never-married, divorced and widowed persons may not be needed and therefore a distinction between married and not-married may be sufficient. Assuming that children living at the

parental home and persons living with other persons (i.e. not with a partner or a child) are not-married, this would lead to 8 categories:
- child living at parental home
- not-married, living without partner
- married, living without partner
- partner in married couple: married, living with partner
- partner in cohabitating couple: non-married living with partner
- living with other persons
- not-married, living in institution.
- married, living in institution.

Table 14: Cross-classification of marital status and living arrangement		
	not-married	married
at parental home	X	
without a partner	X	X
with a partner	X	X
with other persons	X	
in institution	X	X

Note. X: category included in cross-classification; empty cell: category *not* included in cross-classification

For many applications, it may be assumed that married persons live together with their partner. This would imply that only persons living with a partner are divided into married and not-married persons. That would reduce the number of categories to 6.

If in addition, it is sufficient to distinguish between households with and without children rather than to distinguish the number of children[7], the following 9 categories of household position can be distinguished, assuming that children living at home do not live together with their own child(ren) and that people in an institution and persons living with another person do not live together with their child(ren):
- child living at parental home
- living without partner without child
- living without partner with child(ren)
- partner in cohabiting couple without child
- partner in cohabiting couple with child(ren)
- partner in married couple without child
- partner in married couple with child(ren)
- living with other persons
- living in institution

[7] For projecting the size of households, the number of children needs to be specified. Moreover, this would require that for persons living with other persons it needs to be specified in which type of household they live. For example, a distinction should be made between elderly people living with the family of one of their children and persons who live with some relative, not being his/her child or parents (e.g. brothers or sisters) or friend, not being a partner (e.g. students living in the same household).

Table 15: Household position: cross-classification of marital status, living arrangement and the presence of children		
	without child	with child(ren)
at parental home	X	
without a partner	X	X
not-married, living with a partner	X	X
married, living with a partner	X	X
living with other persons	X	
living in institution	X	

Note. X: category included in cross-classification; empty cell: category *not* included in cross-classification

If several transitions can take place in one time interval, transitions between all these categories are possible.

Cross-classification of these 9 categories with sex will result in 18 categories. Cross-classification with sex and age will result in maximum 1800 categories. However, as most household positions will be concentrated at certain age intervals, many categories will contain hardly observations, if at all. Taking this into account, the total number of categories with reasonable numbers of persons would be around 1300. For many analyses level of educational attainment is a useful covariate also. Adding five categories of educational attainment would result in 6500 categories. Clearly, if only a few variables are used, analyses based on stratification imply that a large number of categories have to be distinguished.

The above example showed that household position can be derived as composite variable of marital status, living arrangement and number of children in the household. Based on two-way interactions together with the age and sex distribution for each of these variables separately, the full cross-classification can be estimated using methods of iterative proportional fitting. An alternative approach is to assume independence between the variables at macro level and to use micro level data to derive the composite variables (at micro level all cross-classifications will be available). How to deal with household position in MicMac will be further explored in Work Package 5 'Fertility and living arrangement'.

4. The database: introduction, design and tables required for MAC

One of the activities of MicMac is the development of a (synthetic) database. This database will be part of the overall software application. The software application will be object-oriented, with three broad objects distinguished: a pre-processor to produce the (synthetic) database for projection (the input), a processor that represents the projection engine and a post-processor to store and process the projection results (the output). For sake of user-friendliness the three main objects will be part of the same overall programme shell as far as possible.

In principle, for the pre-processor four steps can be distinguished: 1) preparing a database with raw data; 2) estimation of missing data; 3) calculation of indicators and 4) preparation of scenario parameters. The last step – preparation of scenario parameters – will not be integrated with the projection module. This implies that a future value of a scenario parameter is independent of the projection results (there is

no feedback mechanism). As a consequence, the preparation of scenario parameters may take place apart from the software application, with the scenario parameters as input to the model.

Although in principle the preparation of scenario parameters is part of the pre-processor, for practical reasons this module will not be included in the pre-processor at first. It might be integrated, however, at a later stage. The scenario input, on the other hand, whether or not prepared within the pre-processor, will be part of the (synthetic) database.

The current section deals with a first proposal for preparing the database. The idea of a synthetic database may be used to distinguish a set of 'complete data' (including estimates) and a set of observed (raw) data. We will discuss the design of the database referring to both observed and complete data. The estimation of missing data and the calculation of indicators, however, are beyond the scope of this paper.

4.1 Data storage: a relational database

Data can be stored in different types of files, for instance text files, database files, spreadsheet files, DIF files (Data Interchange Format), SPSS files, binary files, etc. For MicMac we will store in principle all data in (a) relational database(s). For this purpose we will use SQL-server together with MS-Access.

Using a relational database means that the data will be stored by subject in various tables which are interrelated and can be brought together in several ways.
Advantages of a relational database are a.o.:

- data of different tables can be linked, for instance indicators stored in different input-tables, can be used in the same output-table or in the same calculation schemes
- input-tables can be extended with more recent data, for instance a new year in a series, a new country or an additional data source
- tables can be extended with new indicators
- new tables can be added to the database
- data can be easily extracted from the database
- output tables can be defined in any desired format
- individual data can be aggregated in various ways, for instance 1-year age groups in various broad age groups
- different aggregation levels can be stored in one and the same table
- the number of records in a table does not affect the amount of work
- different labels can be used, for instance codes or names or labels in different languages, and it is very easy to switch between labels
- different persons can simultaneously work with the database without having to make copies

An additional advantage of storing all data and tables within one database is that the tables are stored relatively safe, flexible, and well organized. Storing for example all tables in different Excel or ASCII files will have the following implications:

- a choice has to be made on how to organize the tables, for instance by country, or by subject; once a choice has been made this cannot be easily changed later on

- the more tables there are, the more difficult it will be to keep the set of tables well organized
- it is much more difficult to extract data of different tables as this is highly dependent on where exactly in the files the data are stored (no direct access) and how the files are connected to each other
- more records in a table result often in more work
- Excel or ASCII files are 'single-user' only
- in general there will be a difference between the tables that are stored and the tables that are worked with (for instance copies or linked tables)

To sum up, the risk to make errors is much higher using a set of separate tables compared to the use of a relational database. Preparing a relational database, on the other hand, may be more time consuming and somewhat more complicated than the use of separate Excel or ASCII tables. Thus, data storage may be more time-consuming, but data manipulation may be less. We believe that in the end preparing a relational database is worth the investment, however if during the course of the project it turns out that preparing a relational database will be too time-consuming, it might be necessary to adjust the work programme accordingly.

4.2 Database design

All data tables will be stored in .mdb format. While working on the database, we will use SQL-server for data storage and MS-Access for data manipulation. SQL means 'Standard Query Language'. SQL is a computer language, while SQL-server is a package that uses SQL. The main advantages of SQL-server are that 1) in principle all data can be made available for others through internet and 2) a previous version of the database can be restored whenever necessary. This means that you can go back to any previous state of the database and that every update can be rectified. However, as MicMac should be totally open source software, we have to make sure that MicMac can work with a file format that is not dependent on a commercial software programme, i.e. we can create a database in SQL-server/MS-Access, but we should be able to work with the database without having to use these programmes.

Data categories and sources

In principle, for MicMac input as well as output data have to be stored in a database. Generally we can distinguish the following data categories:

1. raw data (observations)
2. base population and values for the projections } input
3. projection assumptions
4. projection results → output

For both input and output data we can distinguish between population numbers, events and transition rates. We either can develop different databases for different categories (or sources or types) of data, or we can try to store different types of data within the same database.

Anticipating the possibility of including a scenario setting module at a later point in time, it seems to be preferable to include raw data for at least a number of years. An additional advantage of this is that we will be able to present not only projected, but also observed trends. As far as possible we will include data from 1995 onwards. In principle, the number of years for which data can be included in the database will not be limited, therefore the period for which observed (raw) data are included might be extended through the years as data for more recent years will become available.

Data will be collected and stored for the Netherlands and Italy. We will use both data of Statistics Netherlands and ISTAT as well as data of Eurostat's database New Cronos. In general, a number of projections could be compiled and stored in the database.

Store comparable numbers in one table

The aim is to store comparable information in one and the same table. The same information of different data sources, for instance population numbers for a specific country originating from the National Statistical Institute (NSI) can be stored in the same table as population numbers originating from Eurostat's New Cronos or the UN database. By including a 'table of preferences', one can select what data one would use, for instance: at first data of the NSI, then, if the NSI data are missing, data of Eurostat and finally data of the UN. This table of preference can be standard for all countries, but can also be country-specific: for instance for those countries with good quality data always use NSI data and for those with hardly any data, use UN estimates. For MicMac we can use this concept for instance to choose between country-specific data or default patterns included in the programme.

Apart from different data sources, also the nature of the numbers can be different: data can be observed, be estimated or refer to projections. Again, these data can be stored in the same table and a choice can be made of what data to be used, using a table of preference, for instance use for period xxxx-xxxx observed figures, and for period yyyy-yyyy projected ones. This can be used to produce output tables or figures in which observed and projected figures are combined.

How to import or upload data

To upload data in the database, an import procedure will be developed. Notwithstanding that the data will be stored in database format, it is not necessary that the data will be delivered in .mdb-files. In principle, tables can be imported in different formats (different kind of files and different data structure). For each format, however, a specific import procedure will be needed. Import procedures will be prepared for ASCII, Excel and Access files. Possibilities to directly import and export data of a relational (Access) database using R will be explored.

For every table a template will be given, specifying the fields and corresponding labels of the table. An example of a template for the table with population figures by age is the following:

GEO	YEAR	SEX	AGE	IND	VALUE	NATURE	SOURCE	NOTE
NL	1980	M	0	PJAN	...	...	NewCronos	
NL	1980	M	1	PJAN	...	...	NewCronos	
...								

In which

GEO	geography: country (or region) code
YEAR	period for which data are stored (calendar year)
SEX	males or females
AGE	age codes, for instance 0,1,2,...,100+, or 0-4,5-9,...,95-99,100+
IND	indicator: for instance PJAN: population at 1 January
VALUE	value of indicator PJAN for GEO by YEAR by SEX by AGE
NATURE	nature of the data: for instance observation, estimation, or projection
SOURCE	source of the data: for instance NSI, Eurostat, UN
NOTE	field for meta data

Meta data will be numbered and will be stored in a separate table. The note field of the data table will include the numbers that refer to the metadata. Each note field can contain several numbers. By including all information at cell level, you can extract all relevant meta data together with each data extraction, even if you only download one number of the database. Moreover, each separate note has to be included in the note table only once (which not only might save time, but also may be very practical if for instance the phrasing of a specific note has to be changed).

How to export or download data

Apart from the import procedure, an export procedure will be developed to download data of the database. To start with, this should refer to those data and tables that will be used as input for the projections. If we take LIPRO as point of departure for Mac, then this could be a number of text files, Excel files or binary files. As for the import procedure, also export procedures will be developed for ASCII, Excel and Access files.

The aim is to generate together with each data table extracted, an information table containing all relevant meta data related to all cells in the data table as well as an overview of all variables in the selection, together with their occurring values. Finally, we strive for the possibility of producing a list of the latest updates, containing for each variable when the data have been added to the database or when the latest changes have been made.

4.3 Data needed for MAC

In order to be able to run full detailed macro level projections with MAC, we need the following data tables[8]:

[8] We have to be aware of the fact that measurement issues may hamper data collection, especially if we would like to collect comparable data across countries. For instance if we would like to use data on the number of children ever born, we might run into trouble if for a specific country the number of children is only registered within marriage. Comparable issues may play a part for other components as well.

Raw data (observations)

The following demographic data are needed:

Population data (population numbers and events):
- population by year of birth (before 1900, 1900, 1901, ...), sex (M, F), and marital status (never-married, married, divorced, widowed) for recent years[9]
- population by year of birth (before 1900, 1900, 1901, ...), sex (M, F), and living arrangement (child in parental home, living without partner without children, living with partner without children, living with partner with child(ren), living without partner with child(ren), living with other person(s), living in institution[10]) for recent years

Event data (in principle we aim for double-classified data):
- number of live births by age (-15,16,...,50+), year of birth (before 1955, 1955, 1956, ...1989, 1990 or later) of the mother, marital status of the mother and parity (1, 2, 3, 4+) for recent years
- number of deaths by age (0,1,...,100+), year of birth (before 1900, 1900, 1901, ...), marital status and sex (M, F) for recent years
- number of immigrants by age (0,1,...,100+), year of birth (before 1900, 1900, 1901, ...), marital status and sex (M, F) for recent years
- number of emigrants by age (0,1,...,100+), year of birth (before 1900, 1900, 1901, ...), marital status and sex (M, F) for recent years[11]
- number of marriages by age (18, 19,..., 50+), year of birth (before 1955, 1955, 1956,..., 1989, 1990 or later), marital status and sex (M, F) for recent years
- number of divorces by age (18, 19,..., 100+), year of birth (before 1900, 1900, 1901, ...), sex (M, F) and marital duration (if available) for recent years
- number of 'widowing' by age (18, 19,..., 100+), year of birth (before 1900, 1900, 1901, ...) and sex (M, F) for recent years

With the exception of live births, the above tables are also required by living arrangement (instead of by marital status). For live births, we assume that all births take place in the living arrangements 'with partner', i.e. either in the state 'living with partner without children' for first births, or in the state 'living with partner with child(ren) for higher order births.

Transition data:
- a cross-classification of the categories of the variable living arrangement for at least two subsequent years by age and sex

[9] Preferably from 1995 onwards, but in case this is not possible for as many years back as data are available.
[10] Note that this classification differs from the one discussed in section 2.7 'Living arrangement' as far as living with(out) child(ren) concerns.
[11] We have to note that emigration data for the NL includes net administrative corrections; as a consequence, emigration can be negative.

Apart from demographic data, the following non-demographic data are needed:

Education[12]:
- population by year of birth (before 1900, 1900, 1901, ...), sex (M, F), level of educational attainment (ISCED 01, 2, 34, 5B and 5A-6) and age at leaving school and reason for leaving for recent years
- enrolment rates by year of birth (single years), sex (M,F), and level of education (ISCED 01, 2, 34, 5B and 5A-6) for recent years

Disability status:
- population by year of birth (before 1900, 1900, 1901, ...[13]), sex (M, F) and disability status (disabled and non-disabled (if available further distinguished by minor and severe disabled)) for recent years

Smoking:
- population by year of birth (before 1900, 1900, 1901, ...), sex (M, F) and smoking behaviour (never, ever, current (if available further distinguished by incidental and heavy smokers)) for recent years

Body mass index:
- population by year of birth (before 1900, 1900, 1901, ...), sex (M, F) and BMI (underweight (<18.5), normal weight (18.5≤BMI≤25), overweight (25≤BMI≤30), obesity (BMI≥30)) for recent years

In addition the following cross-classifications are needed:
- disability by education by marital status
- disability by education by living arrangement
- living arrangement by marital status

If principle, cross-classifications are needed distinguished by age and sex. However, if it turns out that this will be too detailed to collect, then the age and/or sex distribution will have to be estimated.

Base population

If we follow the principle that in general the model should be able to handle all possible interactions, then the base population will contain an enormous number of cells. Taking into account the following (limited number of) cross-classifications: age (101) x sex (2) x level of educational attainment (5) x marital status (4) x living arrangement (5^{14}) x number of children in the household (5) already results in more than 100.000 different combinations. Even though this number is somewhat inflated (not all combinations will be possible or relevant), it will be obvious that it is impossible to work with such a detailed set of data. Therefore, aggregations will be needed. However, even using aggregations, it will be difficult to directly collect all cross-classifications. If we take into account for instance the composite variable

[12] Data on education will only be limited.
[13] Year of birth can be a problem for data on disability status; most often instead of year of birth, age at survey is given, usually in broad age groups.
[14] Here we refer to the five categories of living arrangements as discussed in section 2.7.

'household position' with 9 categories, derived from the variables 'marital status', 'living arrangement' and 'number of children in the household' as discussed in section 3, we can reduce the number of cross-classifications to slightly less than 10.000. Notwithstanding this reduction, it still will be infeasible to collect the full detailed set of data.

To solve this problem, we can collect data by age and sex for the main variables only together with the two-way interactions between the main variables. Subsequently we can estimate the age and sex distribution of the full table using methods of iterative proportional fitting. For instance, if we have tables on 1) population by age (A), sex (S) and household position (HP; 1818 cells in total); 2) population by age, sex and level of educational attainment (E; 1010 cells); and 3) population by age, sex and disability status (D; 404 cells), together with the two-way interactions of household position and educational attainment (HPxE; 45 cells), household position and disability status (HPxD; 18 cells) as well as educational attainment and disability status (ExD; 10 cells), we can estimate the missing cross-classification of household position by educational attainment by disability status by age and by sex (HPxExDxAxS; 18180 cells in total; of which 14875 have been estimated). Please note that for research purposes we will aim for as complete data as possible in order to study how well the estimations fit the data.

A key question here is: do we have to start with the most detailed table thinkable and then use application-specific aggregations? To first estimate the full framework of all cross-classifications and then aggregate the estimations to those categories we need, looks like we move into a circle. Nevertheless this might be necessary to guarantee that we will be able to capture all possible aggregations.

One part of the pre-processor will be a module to estimate the full base population from the data available.

5. Future extensions of MicMac

Other variables than discussed in this report might be relevant for demographic projections as well. Mainly due to budgetary constraints, no further variables are included in MicMac for the time being. Nevertheless it would be advisable to include them in the further development of MicMac. In this section we pay attention to four possible future extensions.

5.1 Ethnicity

Due to immigration, the size of the non-native population in many European countries has increased strongly. As the non-native population may differ from the native population (e.g. level of fertility, level of education, socio-economic status) it may be useful to distinguish the native and non-native population by distinguishing the population by country of birth or by citizenship. As it may be useful to distinguish the second generation, also the country of birth of the parents should be included.

5.2 Place of residence

Even though the current project is not aimed at making regional projections, the MicMac methodology may be used for that purpose, if place of residence is included. For regional population projections, individuals of a given age will be grouped by region of residence to determine the number of residents of a given region (and age). We may distinguish various grouping algorithms to generate population projections on the basis of the biographies generated by Mic. Obviously regional projections require assumptions about internal migration as well.

5.3 Labour force participation

The original project proposal included a work package on the economic dimension. This work package, however, was deleted from the proposal because of budgetary constraints. Nevertheless labour force participation is an important covariate, for instance in explaining fertility trends. Possibly labour force participation may be included as an exogenous variable in making projections.

5.4 Disease-specific health

To start with the variable disability status distinguishes two categories: disabled and non-disabled persons. However, for many purposes more detailed information is useful, specifically a distinction between diseases, as there are big differences in the survival rate and the demand of health care between different types of disease. In addition, therefore, a distinction of mortality by cause of death would be useful for analysing the relationship between disease and mortality.

6. Summary

In this section we summarize the basic issues discussed in this report.

6.1 The state space

The state space defines the possible states a person can occupy as well as the possible transitions between the states.

MicMac will include the following 12 variables (between brackets the estimated number of categories):

1. age (101)
2. year of birth (101)
3. sex (2)
4. level of educational attainment (5)
5. age at leaving school (a limited number of years)
6. marital status (4)
7. living arrangement (between 5 and 7)
8. number of children ever born (5)

9. number of children in household (5)
10. disability status (2)
11. smoking (3)
12. body mass index (4)

As cross-classification of these twelve variables results in a very large number of categories in many cases some aggregations over categories will be needed. However, even if only a few variables will be used, analyses based on stratification imply that a large number of categories have to be distinguished. Methods of iterative proportional fitting can be used to estimate the full cross-classification based on two-way interactions together with the age and sex distribution for each of these variables separately. Alternatively, micro level data can be used, as at micro level all cross-classifications are available.

6.2 The database

All data for MicMac will be stored in (a) relational database(s). This database will include all input for the projections as well as all output of the projections. The aim is to store comparable information in one and the same table. Data import and export procedures will be developed for ASCII, Excel and Access files. Possibilities to directly import and export data of a relational database using R will be explored. As far as possible full detailed data will be collected from 1995 onwards. In principle, meta data will be stored at cell level. If this will be unnecessarily complicated, however, each table in the database, meta data will be stored in separate files.

Data will be collected on population numbers and events (live births, deaths, immigrants, emigrants, marriages, divorces and 'widowing') by age, sex and marital status or living arrangement. Furthermore a cross-classification will be needed of the categories of the variable living arrangement for at least two subsequent years by age and sex. In addition, the following non-demographic data will be collected: data on education, disability status, smoking behaviour and body mass index. Finally, cross-classifications are needed for disability by education and marital status, disability by education and living arrangement ass well as living arrangement by marital status.

Apart from data collection, a module will be developed to estimate the most detailed base population possible, on the basis of data by age and sex for the main variables together with the two-way interactions between these variables.