Arbeidsnotater

STATISTISK SENTRALBYRÅ

WORKING PAPERS FROM THE CENTRAL BUREAU OF STATISTICS OF NORWAY

IO 70/9 Oslo, 23. juni 1970

A DEMOGRAPHIC MODEL FOR THE NORWEGIAN POPULATION AND ITS TECHNICAL CHARACTERISTICS

BY EIVIND GILJE AND SVEIN NORDBOTTEN

C O N T E N T S

I. Introduction

Following efforts concentrated to improve quantitative analysis of the nation's economic aspects, it now seems appropriate to intensify the investigation of population distribution by economic and non-economic classes in order to comprehensively describe and explain the nation's general welfare.

The socio-demographic population model system is a framework within which such an investigation may be carried out. It consists of two main components, the first being a set of systems for classifying the population. These systems include classification by sex, age, family status, residential area, health status, educational status, labour force class, social status, etc. By definitions the application of these systems will establish certain accounting identities, which may be called a population accounting system. The second set of components consists of structural relations which in addition to the accounting relations explain changes in population distribution. Because we are considering models with the transition of persons among a number of classes, the structural relations will be relations which explain the expected transitions from a given class with a specified population size during a certain period. The relations may be very simple, as in the model to be discussed in this paper, with only one parameter or probability to be estimated or they may be more complicated as e.g. when the transition is influenced by factors outside the system, such as economic and educational policies, supply of medical care, etc.

The model system has so far been considered as family of models with the following members:

1. The demographic model
2. The health model
3. The educational model
4. The labour force model
5. The social model

The demographic model will explain how the population is distributed by such characteristics as sex, age, family status and residential area, and how this distribution is changing by its own self-generating influence, and from changes due to factors outside the models e.g. economic policy.

The health model will explain the distribution of population by health status and major illnesses, and how this distribution changes under influence from the demographic distribution and outside factors such as medical care.

The educational model will describe the distribution of the population by its past and present educational activities, and how this distribution reacts to changes in the educational system.

The labour force model will describe the population distribution by its labour force availability and its qualifications, and how the distribution will be influenced by the demographic and educational distributions and the economic environment. The demand for labour as well as the realized employment are, however, variables not considered within the range of this model.

The social model will explain the distribution of the population by different social classes as a result of the above factors and other.

Because there are obvious interactions between the different models, they have to be developed with the total system in mind. The health model must e.g. be developed in such a way that it is compatible with the demographic model. The interface between a pair of models may, however, raise specific theoretical problems. In the development of the structural equations, other theoretical problems will be encountered of which some estimating problems seem still unsolved.

The system or the individual models may be used for analyzing past development, or for making predictions, subject to given assumptions for the factors influencing the probabilities.

The present paper is dealing with the demographic model with special reference to its technical characteristics. Even though it is rather primitive at this stage, it is developed with the above objectives in mind, and may be considered as the first step in the direction of a socio-demographic model system.

II. The demographic model

1. Concepts and symbols

The model is discussed in detail elsewhere (Gilje, 1969), and only a brief summary of the relations will be given in this paper.

In the description of the relations the following conventions and definitions will be used:

Left superscript:

Sex: M for males, F for females

Right superscript:

Municipality no.: k, where $k = 1,2,\ldots,K$.

Right subscript:

Age: x, where $x = 0,1,\ldots,\omega$. (ω is the highest age.)

In addition, n in parenthesis after the symbols represents January 1 of the year n for stock concepts, and the year n for flow concepts.

The definitions are specified explicitly for women only, as they will be analogous, or would have no meaning for men in our model.

The following symbols are needed to explain the model:

(a) ${}^{F}L_{x}^{k}(n)$ represents the actual number of x-year old women January 1 of the year n in municipality k.

(b) ${}^{F}D_{x}^{k}(n)$ represents the actual number of deaths in municipality k in year n among women who were x-years old per January 1 of the year n.

(c) ${}^{F}_{t}U_{x}^{k}(n)$ and ${}^{F}_{t}I_{x}^{k}(n)$ denote the number of out-migrants and the number of in-migrants, respectively and are defined similarly to D.

(d) $B_{x}^{k}(n)$ represents the number of children the x-year old women at the beginning of year n in municipality k, actually will bear during the year n.

(e) ${}^{F}_{t}q_{x}$ is the probability that a woman being x years old, at the beginning of a year, die in the municipality before she is x+t years old.

(f) ${}^{F}_{t}u_{x}^{k}$ denotes the probability that a woman living in municipality k and being x years at the beginning of a year, will move out of the municipality before she is x+t years old.

(g) ${}^{F}_{t}i_{x}^{k}$ is the probability that a female out-migrant of age x at the beginning of a year will move into municipality k before she is x+t years old.

(h) ${}_{t}f_{x}^{k}$ is the expected number of live children born by a woman before she obtains the age x+t years, and who is x years old and living in municipality k at the beginning of a year.

(i) ${}^{F}c$ denotes the proportion of girls in one birth cohort. We assume this to be a constant.

(j) Furthermore, we shall introduce the convention that if X is a ramdom variable, $\tilde{X}$ shall denote a predictor for X. Thus, for example, ${}^{F}\tilde{L}_{x}^{k}(n)$ will be a predictor for ${}^{F}L_{x}^{k}(n)$.

(k) A symbol like ${}^{F}_{t}\hat{q}_{x}^{k}$ denotes an estimator for ${}^{F}_{t}q_{x}^{k}$.

The above symbols can now be used for introducing the simple regional model we intend to use. Our first intention was to estimate specific death probabilities for each municipality. These estimates showed great variations, however, and we decided to use the same set of age-specific estimates for all municipalities.

2. Model relations

For typographical simplification symbols with no sex designation represent either sex, when the relations are analogous for men and women. When the subscript t is omitted, it is always assumed to be 1.

The model contains the following relations:

(2.1) $\tilde{L}_{x+1}^{k}(n+1) = \tilde{L}_{x}^{k}(n) - \tilde{D}_{x}^{k}(n) - \tilde{U}_{x}^{k}(n) + \tilde{I}_{x}^{k}(n).$

(2.2) $\tilde{D}_{x}^{k}(n) = \hat{q}_{x} \cdot \tilde{L}_{x}^{k}(n).$

(2.3) $\tilde{U}_{x}^{k}(n) = \hat{u}_{x}^{k} \cdot \tilde{L}_{x}^{k}(n).$

(2.4) $\tilde{U}_{x}(n) = \sum_{j=1}^{K} \tilde{U}_{x}^{j}(n).$

(2.5) $\tilde{I}_{x}^{k}(n) = \hat{i}_{x}^{k} \cdot \tilde{U}_{x}(n).$

(2.6) $\tilde{B}^{k}(n) = \sum_{x=15}^{44} \hat{f}_{x}^{k} \cdot {}^{F}\tilde{L}_{x}^{k}(n).$

(2.7) $\tilde{L}_{0}^{k}(n+1) = \{1 - ({}_{\frac{1}{2}}\hat{q}_{-1} + {}_{\frac{1}{2}}\hat{u}_{-1}^{k})\} \cdot c \cdot \tilde{B}^{k}(n) + \hat{i}_{-1}^{k} \cdot \tilde{U}_{-1}(n)$

for $k = 1,2,\dots,K$; $0 \leq x \leq \omega$ and $n = 0,1,\dots$. As age 0 applies to children born in the year n-1, we have set x equal to -1 for children born in year n. With this convention, and using a modified version of (2.3):

(2.3') $\tilde{U}_{-1}^{k}(n) = {}_{\frac{1}{2}}\hat{u}_{-1}^{k} \cdot c \cdot \tilde{B}^{k}(n).$

(2.4) and (2.5) will also be valid for $x = -1$.

III. Estimation of probabilities

Let μ and σ be the force of mortality and out-migration respectively. (A definition of the force of mortality can for example be found in Berger, 1939, p. 40. σ is defined similarly.) If we have data for each of the years m to (m+t-1), estimators for average forces of mortality and migration in the period are

$$(3.1)\qquad \hat{\mu}^k_x(m,t) = \frac{\sum_{j=1}^{t} D^k_x(m+j-1)}{\sum_{j=1}^{t} M^k_x(m+j-1)}$$

$$(3.2)\qquad \hat{\sigma}^k_x(m,t) = \frac{\sum_{j=1}^{t} U^k_x(m+j-1)}{\sum_{j=1}^{t} M^k_x(m+j-1)}$$

$M^k_x(m+j-1)$ is the observed, aggregated lifetimes in municipality k and calendar year (m+j-1) of persons of age x at the beginning of the year. This quantity may be approximated by

$$(3.3)\qquad M^k_x(m+j-1) \approx L^k_x(m+j-1) - \tfrac{1}{2}(D^k_x(m+j-1) + U^k_x(m+j-1) - I^k_x(m+j-1)).$$

The following relation:

$$(3.4)\qquad \hat{u}^k_x(m,t) = \frac{\hat{\sigma}^k_x(m,t)}{\hat{\mu}^k_x(m,t) + \hat{\sigma}^k_x(m,t)} \cdot \{1-\exp(-\hat{\mu}^k_x(m,t)-\hat{\sigma}^k_x(m,t))\},$$

is an estimator for the average probability for out-migration in the years m to (m+t-1). A similar formula can be used for $\hat{q}^k_x(m,t)$, but we are satisfied with an estimate of the average mortality in the whole country. In this context, we assume that the net migration from Norway is zero. This gives us:

$$(3.5)\qquad \hat{\mu}_x(m,t) = \frac{\sum_{k=1}^{k} \sum_{j=1}^{t} D^k_x(m+j-1)}{\sum_{k=1}^{k} \sum_{j=1}^{t} M^k_x(m+j-1)},$$

and

$$(3.6) \quad \hat{q}_x(m,t) = 1 - \exp(-\hat{\mu}_x(m,t)).$$

As an estimator for $i^k_x(m,t)$ we use:

$$(3.7) \quad \hat{i}^k_x(m,t) = \frac{\sum_{j=1}^{t} I^k_x(m+j-1)}{\sum_{j=1}^{t} U_x(m+j-1)}$$

$f^k_x(m,t)$ is estimated by:

$$(3.8) \quad \hat{f}^k_x(m,t) = \frac{\sum_{j=1}^{t} B^k_x(m+j-1)}{\sum_{j=1}^{t} {}^{F}L^k_x(m+j-1)}$$

IV. Processing system

1. Planning and preparatory work

By processing system we mean here all data specifications and operative instructions needed to apply the theoretical model set down above on empirical observations and problems.

When we started building this processing system, our main aim was to make it as flexible as possible. To change smaller or greater parts of the programs should be easy for several reasons.

First, changes would be necessary for technical reasons, e.g. the data situation could change; the model application output, as for example population projections, could prove to be inadequate, etc.

Another reason for flexibility is the demand for population projections based on sets of hypothetical probability values as alternatives to the estimated values of the probabilities. Such alternative assumptions do not change the system basically. In many cases it is sufficient to change input data. However, situations may appear when it is necessary to change parts of the system. If this can be done by substitution without damaging the functions of other parts, much programming work and time may be saved.

Lastly, the planned extension of the model had to be taken into account. We may want to split the population into further groups not included this far in the model. Such groups may for instance be a splitting by different marital status, by economically active and non-active population, etc. Large parts of the system will not be affected by such extensions and should therefore be left undisturbed.

All this leads to a modular building of the system of which we will give a description in the next section.

To test our methods and the mathematical model, a simplified version of the system was first made. We shall not go into details about this version, but only state that the experiences we thus obtained gave us a basis to proceed with the general system.

2. System description

The general system can be divided into three separate parts.

The first part consists of aggregation of the "raw" data from the central population register into the more compact form convenient for our special purposes. (Blocks 1 through 7, and 9 and 11 in fig. (4.1).)

In the second part the estimates for probabilities for death and migration and expected number of births are calculated. (Blocks 8 and 10 in fig. (4.1).)

Finally, the system includes one part for applying the numerical model for projection calculations. (Blocks 12, 13 and 14 in fig. (4.1). See also fig. (4.2).)

In the following we will give a short explanation to each of the blocks in fig. (4.1).

Block 1: All data needed on which this model is based, exist as identified, individual data on magnetic tapes. They are identified by the official Norwegian identification numbers introduced October 1, 1964 (Selmer 1967, Skaug 1967).

Block 2: In each record with data of an observed event, the number of the municipality of residence is included. (For instance at the data of death in a death record.) Because of several municipality mergers in recent years, a certain area on the map may not be identified by the same municipality-number during the period we are interested in. Such mergers have therefore been anticipated by changing all municipality-numbers at the beginning of the projection period. One consequence of this is that many movements previously recorded as inter-municipality migration, will be disregarded as such because the new borders surround both municipalities involved.

Block 3: A migration record contains both the municipality-number from which person is moving and that of the municipality into which he is moving. For all migrations in one year the sequence of the file is determined first by the migrants' identification numbers then by the dates in the year that the migrations were registered. Thus, if more than one migration is registered for the same person, they will lie in sequence on the file according to the registration date. In the model we have account for the fact that persons can migrate several times in the course of one year. As we are only interested in where a person is living at the end of each calendar year, we can avoid this problem by disregarding all these intermidiate moves.

The following procedure is used: The first record which in most cases is also the last, for one person is read and the identification number of the municipality in which he was staying at the beginning of the year memorized. Then any more migration records for the same person except the last one is passed. From the last record the number of the municipality of residence at the end of the year is memorized. If the number of the municipality of residence at the beginning and at the end are equal, no migration is recorded. If the numbers are different, the first record is copied into a file which can be called the out-migration file, and the last record (most often equal to the first one), is copied into the in-migration file.

These two files are sorted into sequences determined by the numbers of the municipality of residence at the beginning of the year and at the end of the year, respectively.

Block 4: At this stage sums are developed from the individual data. This is done similarly for deaths, out-migrants, and in-migrants after having sorted the deaths according to the municipality of residence at the time of death.

The sums developed are $D^k_x(m+j-1)$, $U^k_x(m+j-1)$ and $I^k_x(m+j-1)$ for $x = -1,0,1,\dots,\omega$; $k = 1,2,\dots,K$; and $j = 1,2,\dots,t$. (For the use of $x = -1$, see the explanation after (2.7).) These sums are used as elements in $2 \cdot \omega$ -order matrices (two sexes and ω age-groups) formed for each municipality and for each year. One separate matrix is made for each of the three demografic events.

Block 5: This sorting-routine is not essential, but to reduce the number of files the death-, the out- and the in-migration-matrices are sorted into three files. For each municipality we find the matrices for the t years from which we use data in a sequence. The main sorting parameter is thus the municipality-number, the second parameter is the calendar-year the data come from.

Block 6: To calculate the denominators in (3.1) and (3.2), sums of the population at the beginning of each year similar to those which have been counted up for deaths etc., are needed. The procedure is similar to that used in block 4.

Block 7: This block is similar to block 5, and any further comments should not be necessary.

Block 8: The data can now be used for calculating estimates of death- and migration-probabilities. The formulae used in this block are (3. 1-7).

Block 9: This routine differs from that in block 5 by processing individual birth-records instead of the more compact data in block 5 and 7.

Block 10: In contrast to the estimation of death- and migration probabilities we found it was more convenient to estimate expected number of children born, directly from the individual birth-records, i.e. the counting of children born in groups according to the mothers' age and municipality of residence, is not done in a separate block, but in direct connection with the calculation of (3.8).

Block 11: This is identical to block 6. The last year from which stock data of the population can be obtained, is used as the initial year for projections. This will generally, but not necessarily be the year after the last year used in block 6.

Block 12: The model structure estimation is finished at this stage of processing. The application of the model for calculating projections by the formulae (2. 1-7) is represented by this block. The calculating procedure is controlled by certain parameters which give information of the number of municipalities at the beginning of the initial year and how many years we want projections for.

The primary output is of course the projected population, but as a by-product projected number of deaths, children born and migrations for each year and municipality are obtained.

Block 13: The detailed results are printed out in a large table. This table is, however, too large to be suitable for publishing. Therefore several other programs for condensing the results have been made.

3. Programs for estimation

The machine used is an IBM 360/40 model with 65 K, five 9-chanels and one 7-chanels tape-drives, two 2311 disk-drives of which one is reserved for the operating system, one card-reader and one printer.

The choice of programming language was based on the following considerations. As we wanted an operative first version of the model as fast as possible, FORTRAN seemed to be the best choice. Another argument for FORTRAN was the mathematics involved and the extensive use of matrices, both aspects more easily programmed in FORTRAN than in any other language available for the computer to be used.

On the other hand, we had to consider the large amounts of input and output data. Using FORTRAN for input/output, would mean a time consumption on the computer far beyond any reasonable limits. COBOL was therefore a better choice for input/output routines. ASSEMBLER was not considered because of the limited programmer capacity available in this language.

A program written in one of these languages can, as a subroutine, easily be linked together with a program written in the other language. Consequently, all input/output routines and programs containing few or no calculations written in COBOL were linked to the rest which were written in FORTRAN with one exception. Block 6 in fig. (4.1) (and therefore also block 11) already existed as an ASSAMBLER-program. We had only to adjust it to our special purpose. An example on how FORTRAN and COBOL are used together, is given in fig. (4.2) and in the text in explaining this figure.

To indicate the amount of data handled, it may be mentioned that the population register contains about 4 mill. individual records of stock data for each year. We have about 180 000 inter-municipally migrations, 70 000 births and 35 000 deaths; annually. After performing the summations in blocks 4, 6 and 11, we are left with 4 · 451 records (there are 451 municipalities in Norway at the beginning of 1969) for each year containing death-migration- and stock-data.

In these programs the data are also transformed into the floatingpoint form as they are going to be used in FORTRAN programs. From this stage all data and results are kept in this form until the final printout.

4. Projection program

We mentioned in the beginning of this chapter that we wanted a modular system. This is especially important in the projection program in which the mathematical model has been implemented. All changes in the model will also lead to changes in this program.

As seen from fig. (4.2) the program consists of a main program and 10 subroutines. The subroutines are linked together by the main program. This way subroutines can easily be replaced or removed even though constants and arrays may be common for several subroutines.

In order to compute a projection, an initial specification of the population has to be placed on the disc store.

Block 1: The control parameters are read. The first of these gives information on the no. of regions for which the program is supposed to make projections.

We want to be able to stop the calculations on certain points and then to restart. This may be done by establishing regular checkpoints, but this requires a spare file. Instead, we now read two parameters. The former tells the program which initial year the current run is using. This need not be the year from which the projections originally started, as we may have stopped and restarted the calculations several timed by now. It is the programmer's responsibility to see that the same initial stock of population is placed on disk. This is easily done by a utility-program copying the last projections (or the "genuine" initial stock) from tape to disk. The latter of these parameters tell the program when to stop the current run.

Finally two parameters are needed to tell the program which proportion of a birth cohort is boys and which is girls. These parameters are denoted c in formula (2.7).

Block 2: The projections for each municipality are calculated in two steps, and therefore the data is also read in two steps. First the population at the beginning of the year, estimated for expected number of children born, out-migration-, and death-probabilities are needed.

Block 3: The total number of children born is calculated according to (2.6). Then this number is split into boys and girls using the parameters ${}^{M}c$ and ${}^{F}c$.

Block 4: The number of deaths in each age-group is calculated according to (2.2).

Block 5: After finding the number of out-migrants in each age-group according to (2.3), these are cummulated with the number of out-migrants from other municipalities. The resulting matrix of the total number of out-migrants in each age-group, is saved for later use. (Formula (2.4).)

Block 6 & 7: By skipping the last number of (2.1), a preliminary projection is calculated and then written on a disc work-area. Before proceeding to next municipality, projected number of deaths, out-migrants and children born are written on an out-put tape. When the calculations are completed up to this point for the last municipality, the input tapes are rewound in order to make the data available for calculations of next years projections.

In the second step the number of in-migrants is added to the preliminary projections developed in the first step.

Block 8: For each municipality the corresponding estimates for the probabilities of in-migration and the preliminary stock are read.

Block 9 & 10: The matrix built up in the machine by adding the out-migrants over all municipalities is by now complete and (2.5) can be used.

Then the last member of (2.1) is ready, and by adding this to the rest, the final projection is also ready.

Block 11: This projection we write both on an output tape and on disc to replace the former input stock of population. The projected number of in-migrants is also saved on a special tape.

These in-migration routines are repeated until the last municipality has been processed. At this stage the tape with estimates for in-migration probabilities is rewound.

If this also is the last projection year in the run, the processing is stopped. If not, the whole projection procedure is repeated.

V. Operational experiences

The system has been used in three extensive applications so far, and after some minor changes in the first version, it now seems to work according to the planned specifications. Some of our experiences may be of general interest.

The projection program needs approximately 40 minutes to calculate and write projections on tape for one year with full specification for 451 municipalities. The two programs for structure estimation need approximately 42 minutes together to make all the estimates of expected numbers of children born, death- and migration-probabilities.

There is, however, a rather time consuming program early in the process. That is the program extracting the necessary statistics from the population register. Running this program through the approximate 4 million records requires from 4 to 9 hours, depending on how many data controls we find necessary for screening the individual records.

We mentioned earlier that we chose to use the floatingpoint dataform through all programs, including that which writes the final results on tape. This is of course an inefficient way for printing out results, and the data are therefore converted before being listed. A special tabel program performs the conversions and takes care of writing detailed results in a more suitable form both on paper and on tape. To write the original stock of population and 22 years projections for 451 municipalities the computer requires approximately 11 hours.

Three hours out of the 11 are used for printing operations. The remaining 8 hours are mainly used for conversion of data and indexing in (23·2·100)-order matrices. This part of the procedure can probably be made more efficient by another system solution. The results are printed out in a format texted and ready for offset printing in a special publication for population projections.

The projection program's adaptability to alternative assumptions has also been tested. This way done by making an alternative projection assuming each municipality to be a closed comunity; i.e. no migration across municipality borders was allowed. This was easily done eliminating subroutines concerning migration. Then those routines affected by this were appropriately altered.

VI. Plans for further development

For some time theoretical model work has been carried out in connection with the socio-demographic model system. Most of this work has been based on the theory of stochastic processes.

The above described demographic model will be improved in two ways. First, the present constant out-migration probabilities may be replaced by migration relations in which a number of influencing factors will be introduced. When introduced in the operative model, the effects of alternative patterns for regional economic development on the population distributes may be studied.

The second improvement will be to introduce cross-classification by marital status in the model. The current data registration necessary for this extension has already been available for several years, but so far the status for the total population has been missing. When the 1970 Population Census has been carried out, marital status will be currently maintained for the whole population, and this characteristic can be introduced in the model. The demographic model will then be more suited for studying the variations for new housing units, better explanation of the number of births can be developed and a more satisfactory basis for labour force availability studies can be obtained.

From 1971 current individual recording of completed education specified for any cources with a duration of more than 5 months will be introduced. The data on education from the 1970 Population Census and this current registration will be an important source for developing an educational model compatible with the demographic model. The first step in developing such a model for explaining the population distribution by education will be to construct a model which explains an age and sex specified population's transitions to different educations on a national level. The main condition of the model system on this first stage model will be that it is compatible with the demographic model as to age specifications. From an initial population specified by sex, age and education, it may then be possible to develop projections for the population distribution by education which will be consistent with the demographic projections.

The ultimate aim is, however, also in this model to explain how the transitions are affected by external factors such as the capacity of the school system, employment prospects, etc.

A model for explaining the labour force potentials subject to given conditions is considered based on the current population registration, the 1970 Population Census and a Quarterly Labour Force Survey.

These and other compatible models will explain the population distributions by different classifications. Quantitative measures of the social welfare may be computed as transformations of these distributions. The construction of this transformation has to be based on an explicit theory of welfare which is a task outside the scope of the socio-demographic model system.

VII. References

(1) Berger, A. (1939). Mathematik der Lebensversicherung. Verlag von Julius Springer, Wien.

(2) Gilje, E. (1969). Model for Population Projections for Norwegian Regions. Yearbook of Population Research in Finland **XI** and Scandinavian Population Studies **1** : 22-32.

(3) Selmer, E.S. (1967). Registration Numbers in Norway: Some Applied Number Theory and Psychology, The Journal of the Royal Statistical Society, Series A **130** (2) : 225-231.

(4) Skaug, Helge (1967). The Norwegian Population Register System. Proc. Int. Sympos. on Autom. of Popul. Register Systems, Jerusalem 1967, **1** : 3-39.

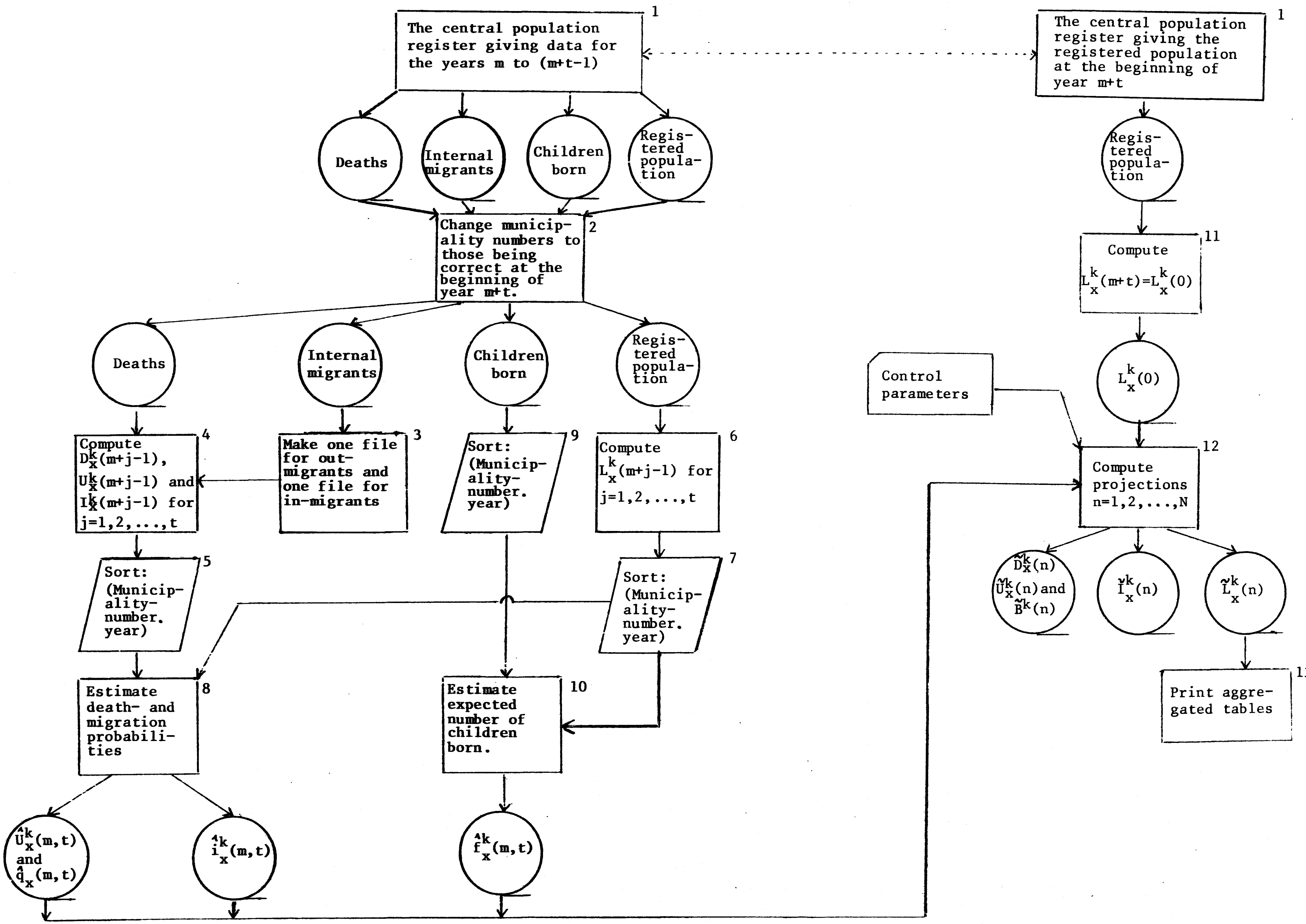

Fig. (4.1): Flow-chart of the regional projection system

Fig. (4.2): Flow-chart of the projection program