Population Accounts

Chapter · February 2011

DOI: 10.1007/978-90-481-8930-4_2

Frans Willekens
Netherlands Interdisciplinary Demographic Institute

Metadata of the chapter that will be visualized online

ChapterTitle	Population Accounts	
Chapter Sub-Title		
Chapter CopyRight - Year	Springer Science+Business Media B.V. 2011 (This will be the copyright line in the final PDF)	
Book Name	Population Dynamics and Projection Methods	
Corresponding Author	Family Name	**Willekens**
	Particle	
	Given Name	**Frans**
	Suffix	
	Division	
	Organization	Netherlands Interdisciplinary Demographic Institute
	Address	The Hague, The Netherlands
	Email	Willekens@nidi.nl

Abstract

Phil Rees pioneered multistate demographic accounts. A population account is a two- or multi-dimensional table that integrates data on population change in a systematic and consistent way. It integrates information on population stocks and flows, demographic events such as births, deaths and migration that cause population change. In a set of population accounts, every person and every relevant demographic event is accounted for. Missing data are estimated from available information, rules of accounting (e.g. no double count) and necessary assumptions. The outcome is a table with consistent data on population size, structure and change. Accounting equations link stocks and flows. Accounts become particularly useful when they serve as the basis for demographic modelling. Account-based models estimate the parameters of the model from the population account.

Chapter 2
Population Accounts

Frans Willekens

Introduction

Phil Rees pioneered multistate demographic accounts. A population account is a two- or multi-dimensional table that integrates data on population change in a systematic and consistent way. It integrates information on population stocks and flows, demographic events such as births, deaths and migration that cause population change. In a set of population accounts, every person and every relevant demographic event is accounted for. Missing data are estimated from available information, rules of accounting (e.g. no double count) and necessary assumptions. The outcome is a table with consistent data on population size, structure and change. Accounting equations link stocks and flows. Accounts become particularly useful when they serve as the basis for demographic modelling. Account-based models estimate the parameters of the model from the population account.

In their book entitled *Spatial Population Analysis*, Phil Rees and Alan Wilson (1977) showed that population accounting is the keystone of demographic modelling. Many modelling problems that previously could be solved on an ad hoc basis only can be solved comprehensively and more effectively if an accounting framework is adopted. That is particularly so when data are incomplete, which most often is the case. Suppose we need information on migration flows by region of origin and region of destination but the available data is limited to (i) arrivals and departures by region and (ii) a selection of migration flows. To obtain an internally consistent set of migration figures, all flows must be determined simultaneously.

In this chapter, the main principles of demographic accounting are reviewed and it is demonstrated that the approach proposed by Rees and Wilson in 1977 is an Expectation-Maximization (EM) algorithm *avant-la-lettre*. What this means is that the accounting method designed by Rees and Wilson incorporates the core features of the EM algorithm, which was not yet available at that time and a dominant method of statistics only later. The EM algorithm is a generic method for estimating model

F. Willekens (✉)
Netherlands Interdisciplinary Demographic Institute, The Hague, The Netherlands
e-mail: Willekens@nidi.nl

J. Stillwell, M. Clarke (eds.), *Population Dynamics and Projection Methods*,
Understanding Population Trends and Processes 4, DOI 10.1007/978-90-481-8930-4_2,

parameters from incomplete data. It identifies two steps. The first step predicts the model outcome using preliminary estimates of the parameters of the model. It is the expectation (E) step. In the second step, the maximization step (M), improved parameter estimates are obtained. The same two steps are distinguished by Rees and Wilson.

The structure of the chapter is as follows. The next section reviews basic features of multistate demographic accounts and distinguishes between accounts on the basis of the measurement of flows of different types. This is followed by a presentation of the algorithm proposed by Phil Rees for estimating the demographic parameters from incomplete data. The next section compares the algorithm proposed by Rees and the EM algorithm. A section of conclusions completes the chapter.

Multistate Demographic Accounts

Demographic accounts are tables that combine data on population stocks and flows. A population consists of people with various attributes such as age, sex, region of residence, marital status, employment status and health status. Attributes change over time and the change can be positioned in one or several time scales. Common time scales include individual time (age) and calendar time. A time scale measures the time elapsed since a reference event such as a census. Age is the duration of life or time elapsed since birth. Calendar time (e.g. t) is the time elapsed since the start of our calendar. Data on attributes and particularly data on attribute changes are generally incomplete. For some people, the attributes may not be known and, even if the attributes are known, changes may remain unnoticed or may be recorded with some delay. For example, whilst we may know the number of people who migrate from one region to another in a time period, it is very likely that we will not know the numbers who migrate and die during the period or those who are born and migrate as infants.

The rules or principles of spatial demographic accounting state how it is possible to organize the available information in an account and how to infer missing attributes and missing information on attribute changes. The missing information may be inferred from the available data using either simple rules or complicated statistical theory. It is not uncommon that rules that have been in existence for quite some time are shown to produce results that are fully consistent with the theories of statistical inference.

The principles of accounting presented in this section are based on Rees and Wilson (1977). First, the population system needs to be defined and the boundaries of the account need to be determined. The population system for which the account is developed is generally not the world population since the origin of humankind, but a well-defined population situated in time and space. The population is divided into subpopulations on the basis of attributes of members of the population. A person with a given attribute is said to occupy a given state. The population can therefore be referred to as a multistate population. A change in attribute implies a move to a different state. Thus a resident of a given region who at time t is aged x and not employed can get a job in another region and migrate during the period between t and t+1. As a result, both the employment status and the region of residence change during the period of 1 year. Note that a person's change of state may remain unregistered for some time of even forever. A distinction is made between the occurrence of an event, which depends on the definition of the event, and the measurement of the occurrence. For instance, if a person relocates to a new address, an event occurs. However, the geographical relocation will not show up in a statistical database unless it is reported or registered in another way, and the information is used in the compilation of statistics.

A multistate population is embedded in a larger population and in history. If no exchange with the larger population is possible, the population is a closed population. In general, the population system for which a population account is prepared is an open system, unless the account refers to the whole of the world. If the population account is for one country/nation, persons may enter by birth or immigration and leave the population by death or emigration. If the population covers a particular age range, a person enters by reaching the lowest age of the age range or immigration and leaves by reaching the highest age or emigration. The account may cover a subpopulation only, e.g. students, health workers, retirees or the residents of a city or region. Moves across boundaries of an open population are sometimes termed external events as opposed to internal events that do not involve the crossing of the boundary of the population system. External events comprise exits and entries. Exits may be subdivided by destination, entries by origin.

Persons are located in space and time. Measurement methods differ substantially in the way they register the location (in space) of a person and the date (location in time) of an event. The location of a person may be measured precisely and displayed in a Cartesian coordinate system or approximately by using a grid system of a system of regions. In a grid system, an area is divided in squares of, say, one kilometre by one kilometre. In a multiregional system, an area is divided in regions that may differ in shape and size. In multiregional population accounts, the location is given in terms of the region of residence. The date of a change in attribute or event may be measured precisely and reported in day, month and year, or approximately and reported in month or year of occurrence. Many demographic surveys report dates of events in terms of month and/or year. Relocations and other events are measured directly or indirectly. The indirect measurement is obtained by comparing regions of residence and other personal attributes at two points in time.

Direct and indirect measurements result in different data types. Data resulting from the direct measurement of events during a given interval have been referred to as movement data (*moves*). Data resulting from the indirect measurement of events have been referred to as transition data (*migrants*); they are essentially a cross-tabulation of final states (state occupied at the end of an interval) and initial states (state occupied at the beginning of an observation interval). Different data types require different methods for estimating the parameters of demographic models (see Ledent, 1980, p. 558).[1] They also lead to different accounts: *movement accounts* and *transition accounts* (Rees & Willekens, 1986).

---

[1]Rogers (1975) refers to the method that uses movement data as the 'option 1' method and the method that uses transition data as the 'option 2' method.

The direct measurement of events is illustrated in population registers. A population register is designed to be a continuous surveillance system and to record events as they occur. A person experiencing an event is expected to report the event right away or within a brief period of time. The date of the event (or date of registration) is recorded. The indirect measurement is illustrated in a population census that records the place of residence at two points in time (e.g. place of residence at the census and at previous census or 1 or 5 years prior to the current census). If the places of residence differ, the person has changed residence at least once. If the places are the same, a person may have left his residence and returned during the observation interval. The census may also register retrospectively the date of the change of address or the duration of stay at the current address. Demographic accounts should be able to handle different types of data and to harmonize data of different type. In developing a typology of data, it is important to distinguish between moves (also referred to as direct transitions) and transitions.

Construction of Population Accounts

Rees and Wilson (1977) consider a mixture of movement and transition data. Births and deaths are movement data and migrations are measured indirectly by comparing the region of residence at the start of a time interval and the region of residence at the end of the interval. The discrete time interval is from t to t+T, with T the interval length. The population at time t is the initial population and the population at t+T is the final population.

The principles of demographic accounting proposed by Rees and Wilson are now briefly outlined. Let $K_i(t)$ denote the population in region i at time t and let K_{ij} denote the number of persons in region i at time t and region j at time t+T ($i = 1, 2, 3, \ldots, N$ and $j = 1, 2, 3, \ldots, N$). K_{ii} denotes the number of people in region i at the beginning of the interval (at time t) who are also in i at the end of the interval. The number of residents of region i at t who die in region j during the interval is denoted by $K_{i\delta(j)}$. If j differs from i, it represents a migration followed by a death. The number of children born in region i during the interval who survive and live in region j at the end of the interval is denoted by $K_{\beta(i)j}$. Some children born during the interval die before the end of the interval. That number is denoted by $K_{\beta(i)\delta(j)}$. Rees and Wilson (1977, pp. 22 ff.) distinguish major flows and minor flows. Major flows involve one demographic event only (in addition to survival). They are K_{ij}, $K_{\beta(i)i}$ and $K_{i\delta(i)}$. The others are minor flows.

The development of the accounts table consists of five steps:

1. Assemble the available data in the account.
2. Obtain preliminary/improved estimates of the population at risk.
3. Estimate birth and death rates for each region.
4. Obtain estimates of the minor flows.
5. Go to step 2.

The population at risk in a given region measures the total time spent in that region by all persons combined included in the account. The exposure time (population at risk) is determined for the events of birth and death, but not for migration. Since migration is measured by comparing the initial region of residence (at t) and the final region of residence (at time $t+T$), the base population is considered a suitable approximation of the population at risk (Rees & Wilson, 1977, p. 34).

The estimation of the population at risk of an event (birth or death) from the available data requires assumptions about the timing of the events in the interval from t to $t+T$. It is assumed that the K_{ij} persons who are in i at t and in j at $t+T$ and who experience a single event during the interval, experience the event in the middle of the interval, hence at time $t+0.5T$. As a consequence, a surviving migrant is exposed half a period in the region of origin i and half a period in the region of destination j. The $K_{i\delta(i)}$ persons who die in the region in which they reside at time t die in the middle of the interval. They also contribute half a period to the total exposure in region i. The $K_{i\delta(j)}$ persons who migrate to another region and subsequently die in that region contribute $0.25T$ to the total exposure in region i and $0.25T$ to the total exposure in region j. It implies the assumption that death occurs in the middle of the period and the migration halfway the first subperiod. The $K_{\beta(i)i}$ children who are born in region i and are in the same region at the end of the interval are assumed to be born in the middle of the interval. They contribute $0.5T$ to the exposure time. The $K_{\beta(i)j}$ newborns who migrate to another region contribute $0.25T$ to the region of birth and $0.25T$ to the region of destination. The $K_{\beta(i)\delta(i)}$ newborns who die in their region of birth contribute $0.25T$ to the exposure time in the region of birth, implying the assumption that they are born in the middle of the interval and die in the middle of the second sub-period. The $K_{\beta(i)\delta(j)}$ newborns who die in a different region are assumed to contribute $0.125T$ to the total exposure in the region of birth and $0.125T$ to the exposure in the region of destination (and death). This implies the assumption that the child is born in the middle of the interval of length T, dies in middle of the second subinterval of length $0.5T$, and migrates in the middle of the first part of the second interval, which has a length of $0.125T$.[2]

The person-years of exposure in region i to the event of birth is:

$$\begin{aligned} PY_i^B = T\,K_{ii} + 0.5T\sum_{j\neq i} K_{ij} + 0.5T\sum_{j\neq i} K_{ji} + 0.5T\,K_{i\delta(i)} \\ +0.25T\sum_{j\neq i} K_{i\delta(j)} + 0.25T\sum_{j\neq i} K_{j\delta(i)} \end{aligned} \tag{2.1}$$

[2]In the context of event history analysis, Yamaguchi (1991) uses a similar reasoning to arrive at estimates of exposure time.

And the person-years of exposure in region i to the event of death is:

$$\begin{aligned} PY_i^D = PY_i^B &+ 0.5T\,K_{\beta(i)i} + 0.25T\,K_{\beta(i)\delta(i)} + 0.25T\sum_{j\neq i} K_{\beta(i)j} + 0.25T\sum_{j\neq i} K_{\beta(j)i} \\ &+\ 0.125T\sum_{j\neq i} K_{\beta(i)\delta(j)} + 0.125T\sum_{j\neq i} K_{\beta(j)\delta(i)} \end{aligned} \tag{2.2}$$

Several of these quantities are not available. The equations are therefore rewritten to express the person-years in terms of observed quantities and unknown quantities. The following terms are usually available directly from the data (Rees & Wilson, 1977, p. 23): the initial population in region i K_{i*}, the total number of births, $K_{\beta(i)*}$, the total number of deaths, $K_{*\delta(i)}$, the numbers of surviving migrants, K_{ij} $(i\neq j)$, and migrating infants, $K_{\beta(i)j}$ $(i\neq j)$. The approximate person-years expressed in these terms are:

$$PY_i^B = T\,K_{i*} - 0.5T\,K_{*\delta(i)} + 0.5T\sum_{j\neq i}(K_{ji} - K_{ij}) \tag{2.3}$$

for births and

$$PY_i^D = PY_i^B + 0.5T\,K_{\beta(i)*} \tag{2.4}$$

The birth rate in region i is obtained by dividing the number of births by the person-years of exposure:

$$b_i = K_{\beta(i)*}\big/PY_i^B \tag{2.5}$$

The rate is an occurrence-exposure rate because it is obtained by dividing the occurrences (number of births) by the exposure time. Since exposure time is measured in person-years, the rate reflects the intensity of the event in a year. In most applications the time unit is a year, but in some applications of multistate models it is a month or even a day. The death rate is obtained in similar way as:

$$d_i = K_{*\delta(i)}\big/PY_i^D \tag{2.6}$$

To illustrate the development of a population account, I use the example provided by Rees and Wilson. Three regions are considered: the West Riding of Yorkshire in the UK (region 1), the rest of England and Wales (region 2), and the rest of the world (region 3). The initial step in developing the population account is to fill the accounts table with available data. Table 2.1 shows the known population stocks and flows.

The approximate person-years based on the available data and the associated birth and death rates are estimated using the equations shown above. For the West Riding, the average annual birth rate is 18.9 per thousand and the death rate 12.3 per thousand. For the rest of England and Wales, the annual birth rate is 19 per

Table 2.1 Population account table for the West Riding of Yorkshire, 1961–1966, known flows

		Population 1966			Deaths 1961–1966			
		1	2	3	1	2	3	Total
Population 1961	1	–	168,207	45,089	–	–	–	3,650,586
	2	137,183	–	841,621	–	–	–	42,453,962
	3	58,804	1,026,175	–	–	–	–	–
Births in 1961–1966	1	–	7,492	3,808	–	–	–	333,135
	2	7,003	–	57,890	–	–	–	3,924,421
	3	2,617	44,250	–	–	–	–	–
Total		–	–	–	226,694	2,537,636	–	–

Source: Rees and Wilson (1977, p. 57).

thousand and the death rate is 11.7 per thousand. The death rates are employed to calculate the following minor flows: the $K_{i\delta(j)}$ migrants who die before the end of the interval, the $K_{\beta(i)\delta(j)}$ children born in region i who migrate to region j and die before the end of the interval, and the $K_{\beta(i)\delta(i)}$ children born in region i who die in that region before the end of the interval. To determine how many migrants from i to j die in j before the end of the interval, the number of migrants, K_{ij}, is multiplied by a function of the rate of dying during a period of $0.25T$:

$$K_{i\delta(j)} = \frac{0.5T\,{}_1d_j}{1 - 0.25T\,{}_1d_j} K_{ij} \tag{2.7}$$

where ${}_1d_j$ is the annual death rate in region j. The number of children born in i who die in j is:

$$K_{\beta(i)\delta(j)} = \frac{0.25T\,{}_1d_j}{1 - 0.125T\,{}_1d_j} K_{\beta(i)j} \tag{2.8}$$

The number of children born in i who die in i is:

$$K_{\beta(i)\delta(i)} = \frac{0.25T\,{}_1d_i}{1 - 0.125T\,{}_1d_i} K_{\beta(i)i} \tag{2.9}$$

where:

$$K_{\beta(i)i} = K_{\beta(i)*} - \sum_{j \neq i} K_{\beta(i)j} - K_{\beta(i)\delta(i)} - \sum_{j \neq i} K_{\beta(i)\delta(j)} \tag{2.10}$$

Rearrangement gives:

$$K_{\beta(i)\delta(i)} = \frac{0.25T\,{}_1d_i}{1 + 0.125T\,{}_1d_i} \left[K_{\beta(i)*} - \sum_{j \neq i} K_{\beta(i)j} - \sum_{j \neq i} K_{\beta(i)\delta(j)} \right] \tag{2.11}$$

Once these minor flows are estimated, better estimates can be obtained of the person-years at risk and the birth and death rates. Iteration results in final estimates of the missing flows in the population account. Table 2.2 shows the final account:

Demographic Accounting Method and EM Algorithm Compared

The final account is obtained by an iterative procedure involving three basic steps. The first is to obtain the population at risk (person-years of exposure) using a demographic accounting equation. In the second step, the births and death rates are obtained by dividing the numbers of events by the populations at risk. In the third step, these updated rates are entered into a model derived from the demographic accounting equation that predict the missing data. The rates that are produced in the

Table 2.2 Population account table for the West Riding of Yorkshire, 1961–1966, final estimates

		Population 1966			Deaths 1961–1966			
		1	2	3	1	2	3	Total
Population 1961	1	3,220,324	168,207	45,089	210,558	5,004	1,404	3,650,586
	2	137,183	39,056,927	841,621	4,271	2,388,922	25,038	42,453,962
	3	58,804	1,026,175		1,831	30,528		–
Births in 1961–1966	1	311,781	7,492	3,808	9,885	111	59	333,135
	2	7,003	3,746,147	57,890	108	112,418	855	3,924,421
	3	2,617	44,250		40	653		–
Total		3,737,712	44,049,198		226,694	25,37,636	–	–

Source: Rees and Wilson (1977, p. 72).

second step are occurrence-exposure rates. An occurrence-exposure rate is a ratio of the number of events experienced by a group of people during a given period and the exposure time during the same period.

The EM algorithm is also based on a model of the data and is an iterative procedure too. The algorithm was developed by Dempster, Laird, and Rubin (1977) and is currently the main method for maximum likelihood estimation in the presence of missing data. The EM algorithm uses a probability model to describe the event of interest. The type of model depends on the type of event. Births, deaths and migrations (moves) are described by Poisson models or by an extension of the Poisson model such as the negative binomial model. The selection of a model determines the probability distribution of the expected number of events. If the Poisson model is used, it is assumed that the number of events during a time interval follows a Poisson distribution. Transitions in discrete time (e.g. migrants, see before) are described by multinomial probability models. The models listed here are standard models in probability theory and they are described in any introductory textbook. For a thorough review of the EM algorithm, see McLachlan and Krishnan (1997). For a discussion of the EM algorithm in the context of migration analysis, see Willekens (1999). The demographic accounting method is not based explicitly on a probability model. The model is implicit, however, in the estimation of the person-years of exposure and, more particularly, in the assumption that events occur in the middle of the risk period. That assumption is consistent with the assumption that events are uniformly distributed during the risk period. The uniform distribution is one of the common probability distributions documented in probability theory. When the occurrence of events follows a uniform distribution, then the expected waiting time to an event, provided it occurs during a given time interval, is half the interval. Traditionally the uniform distribution is (implicitly) assumed in demographic models of event sequences. The Poisson distribution, which implies an exponential distribution of events during the risk period, is (implicitly) assumed in most statistical models of event sequences. The distribution assumes that the rate at which events occur is constant.

In the EM algorithm, the parameters of the distribution are estimated from the data using the maximum-likelihood method. The method maximizes the likelihood of the data given the model. Maximum-likelihood estimates have a number of statistical properties. In the demographic accounting method, the statistical properties of the parameter estimation method are not studied. The properties can be inferred, however. Andersen and Keiding (2002), Andersen and Perme (2008) and others show that the occurrence-exposure rate estimated by dividing an event count by an exposure time agrees well with the Nelson-Aalen estimator *provided the rate at which events occur is constant*. The Nelson-Aalen estimator is a non-parametric estimator of the cumulative rate of the transition/hazard rate. The estimator is part of the toolkit of statistical survival analysis and event history analysis and is well-documented in the literature. Note that the occurrence-exposure rate estimated by the demographic accounting method is not consistent with the Nelson-Aalen estimator because of the assumption of uniform distribution of events.

The EM algorithm is a generic method that offers maximum likelihood solutions when data are incomplete. It reformulates the incomplete-data problem as a complete-data problem and estimates the parameters of the probability model in two steps. The first step of the algorithm infers the missing data using preliminary estimates of the parameters of the model. It is the expectation (E) step. In the second step, the 'complete' data and the maximum likelihood method are used to improve on the parameter estimates. It is the maximization (M) step. These two steps, that are the main characteristic of the algorithm, are also present in the Rees-Wilson account-based method. The first step predicts event occurrences (event counts) on the basis of available data and preliminary estimates of exposure time. The second step estimates event rates (fertility, mortality and migration rates) by dividing event occurrences by exposure time. The third step repeats the first step to produce improved predictions of event counts. The similarity between these two characteristic steps is the reason why the Rees-Wilson method is referred to as 'EM algorithm *avant-la-lettre*'.

Conclusion

The demographic accounting method, pioneered by Phil Rees, is an EM algorithm *avant-la-lettre*. It distinguishes the two basic steps in model development with incomplete data: the prediction of the model outcome from preliminary parameter estimates and the improved estimation of the parameters from the 'complete' data. It results in parameter estimates that are optimal in some way given the data. The parameters have interesting statistical properties that became known only much later after the development of the statistical theory of counting processes.

Phil Rees showed convincingly that a demographic account represents an appropriate basis for demographic modelling (see e.g. Rees, 1979). Since accounts must balance, the accounting framework assures consistency between the flows and the stocks from which the model parameters are derived. The power of account-based models has often been demonstrated, recently in a major report on ethnic population projections for the UK (Rees, Norman, Wohland, & Boden, 2010). The observation that account-based models and modes that are based on statistical inference share important features may lead to renewed research that further strengthen the position of the accounting method in modelling.

References

Andersen, P. K., & Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research, 11*, 91–115.

Andersen, P. K., & Perme, M. P. (2008). Inference for outcome probabilities in multi-state models. *Lifetime Data Analysis, 14*, 405–431.

Dempster, A. P., Laird, M. M., & Rubin, D. B. (1977). Maximum likelihood fromincomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B, 39*, 1–38.

Ledent, J. (1980). Multistate life tables: Movement versus transition perspectives. *Environment and Planning A, 12*(5), 533–562.

McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.

Rees, P. H. (1979). Regional population projection models and accounting methods. *Journal of the Royal Statistical Society A, 142*, 223–255.

Rees, P., Norman, P., Wohland, P., & Boden, P. (2010). *Report on ethnic population projections for the UK*. School of Geography, University of Leeds, Leeds.

Rees, P. H., & Willekens, F. (1986). Data and accounts. In A. Rogers & F. Willekens (Eds.), *Migration and settlement: A multiregional comparative study* (pp. 19–58). Reidel Press: Dordrecht.

Rees, P., & Wilson, A. (1977). *Spatial population analysis.* London: Edward Arnold.

Rogers, A. (1975). *Introduction to multiregional mathematical demography*. New York: Wiley.

Willekens, F. J. (1999). Modeling approaches to the indirect estimation of migration flows: From entropy to EM. *Mathematical Population Studies, 7*(3), 239–278.

Yamaguchi, K. (1991). *Event history analysis*. Newbury Park, CA: Sage Publications.