Revision 1, 14 August 2005

THE DEMOGRAPHIC MEASUREMENT OF MIGRATION AND ITS ADJUSTMENT FOR UNDERENUMERATION

DRS WILLIAM L J XU-DOEVE

2005

Paper presented at the XXV International Population Conference of the International Union for the Scientific Study of Population (IUSSP), 18-23 July 2005, Tours, France

E-mail: w.l.j.xu-doeve@uwnet.nl

Up-to-date contact details can always be obtained from the IUSSP membership directory at <http://www.iussp.org/>

ABSTRACT

In this paper, we focus on the measurement of migration and its adjustment for underenumeration.

We set out by briefly sketching current worldwide trends in *internal migration*, with an emphasis on migration and urbanization in the Third World, and in *international migration*. This provides the backdrop and a motivation for the subsequent work.

Next, we briefly assess the status quo of demography in the area of relevant theory development and in the area of the measurement of migration.

We then develop an elementary but carefully-argued rigorous axiomatic-deductive theoretical mathematical framework centring on the instantaneous rates at which individual demographic events occur in continuous time. It is a *general demographic framework*, applicable equally to the study of, for example, mortality, fertility and migration.

We continue by investigating how to establish relationships between abstract theory and empirically-observable events. The mathematical theory is shown to lead to valid *universal demographic measurement methods* in an unambiguous and simple manner.

Finally, we demonstrate the power of this method in the study of migration by using empirical migration data for Bangkok, data which are not necessarily fully complete nor fully without error. We show how such data defects can be corrected in a theoretically-justifiable manner in the general case where one is dealing with deficient data sets.

This approach to measurement is the first truly demographic method of measuring and adjusting migration data in populations for which the data are incomplete and defective, such as in statistically less-developed countries or in the case of illegal migration.

KEYWORDS

internal migration, international migration, mathematical demography, demographic theory, multistate demography, demographic accounts, Poisson process, hazard function, methods of measurement, methods of estimation and adjustment

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure 1 Urban Expansion on the Fringes of Bangkok to Accommodate for Natural Population Growth and Migration at the Start of the 21st Century

These aerial photographs show the recent expansion at the urban fringes of Bangkok, the capital city of Thailand. They express the growth of the metropolitan area to accommodate for the city's steadily rising population as a consequence of natural growth and migration.

Data for Bangkok are used to illustrate the theory and methodology developed in this paper.

Bangkok is located close to the sea in the low-lying and flat area of the delta of the Menam (river) Chao Phraya. On the photographs one clearly sees recent high-rise and low-rise housing development, building structures for small and lager scale economic activities and services, and recently developed infrastructure. The regular lay-out of the road network suggests a planned approach. In the bottom two photographs, on can also see informal settlement along the khlongs (the small rivers and canals) in the delta of the Chao Phraya. The silver-coloured pipe next to the double bridge across the khlong in the bottom-centre of the second image is a mains water supply pipe. Particularly interesting as well is, that in all photographs one can clearly see that the land is not yet fully occupied, a typical phenomenon of urban expansion in the major cities of the Third World where urban development is a combination of public and private initiative.

These photographs are courtesy of Drs Paul Hofstee, International Institute for Geo-Information Science and Earth Observation (ITC), Enschede, the Netherlands, who kindly searched his vast library of aerial photographs for suitable images.

1 THE CONTEXT: MIGRATION AND URBANIZATION IN THE THIRD WORLD AND WORLDWIDE INTERNATIONAL MIGRATION

While in the 1950s one third of the world's population lived in urban areas, towns and cities now house half the world's population. In developing countries, many have been and continue to be pushed off the land and out of agricultural employment by factors such as

- population growth in rural areas and a shortage of arable land, resulting in population pressure and the associated
  - subdivision of agricultural land holdings
  - cultivation of more and more marginal land
  - low agricultural productivity

and sometimes also by

- agricultural modernization, resulting in a reduced demand for labour

Often, few alternative employment opportunities are available in rural areas. Typically, one sees

- insufficiently-developed non-agricultural economic activities
- a relatively insignificant cash economy

In some countries and regions also, regional conflict constitutes a powerful force driving rural residents from their homes.

Towns and cities, and in particular the key economic and administrative centres, focal points for investment, production, commerce, communication and consumption, are widely perceived as offering better opportunities and services. They attract the rural landless and unemployed. Table 1 illustrates the growth of the urban and rural populations for the major areas of the world.

Table 1 Total, Urban and Rural Populations by Major Area, selected periods: 1950-2030

Major area	Population (millions)					Average annual rate of change (per cent)		
	1950	1975	2000	2003	2030	1950-1975	1975-2000	2000-2030
Total population								
Africa	221	408	796	851	1,398	2.45	2.67	1.88
Asia	1,398	2,398	3,680	3,823	4,887	2.16	1.71	0.95
Europe	547	676	728	726	685	0.84	0.30	-0.20
Latin America [1]	167	322	520	543	711	2.62	1.92	1.04
Northern America	172	243	316	326	408	1.40	1.04	0.85
Oceania	13	22	31	32	41	2.08	1.46	0.97
Urban population								
Africa	33	103	295	329	748	4.57	4.21	3.10
Asia	232	575	1,367	1,483	2,664	3.63	3.47	2.22
Europe	280	446	529	530	545	1.86	0.68	0.10
Latin America [1]	70	197	393	417	602	4.14	2.76	1.42
Northern America	110	180	250	261	354	1.98	1.32	1.16
Oceania	8	15	23	24	31	2.75	1.51	1.07
Rural population								
Africa	188	305	500	521	650	1.93	1.98	0.87
Asia	1,166	1,823	2,313	2,341	2,222	1.79	0.95	-0.13
Europe	267	230	199	196	140	-0.61	-0.57	-1.17
Latin America [1]	97	125	127	126	109	1.00	0.08	-0.51
Northern America	62	64	66	65	53	0.11	0.14	-0.70
Oceania	5	6	8	9	10	0.76	1.31	0.68

[1] including the Caribbean

Source: United Nations (2004a)

Without exception, the urban population growth rates exceed the rates for the total populations, a phenomenon which is expected to continue unchanged for the period up to 2030 (United Nations, 2004a). The difference between rural and urban growth rates is an expression of a large-scale population redistribution from rural areas to towns and cities through internal migration.

If we ignore the effects of international migration and of the reclassification of rural populations as urban, then we can obtain an indication of the magnitude of this internal redistribution by letting the urban populations increase at the rates for the total populations.

For example, between 1975 and 2000, the urban populations of Africa, Asia and Latin America in table 1 -- the regions approximately corresponding to the developing countries --increased from 875 million to 2055 million. Had these urban populations grown at their respective rates for the total population, then the urban population of these regions would have been only some 1401 million by 2000.

The difference between this figure and the actual number of 2055 million in 2000, some 654 million urban inhabitants, is the result of population redistribution from rural areas to towns and cities through internal migration. Over the period, this net effect of internal migration therefore explained over 55% of the urban growth, amounting to some 26 million people on average each year. While in the developing world natural growth generally tends to be higher in rural areas than it is in urban areas, the net effect of internal migration is in fact greater than this.

The trend to urbanization in much the developing world since the middle of the 20^{th} century is unprecedented in terms of its magnitude and its pace, surpassing the historical experience of developed countries. For example, Africa's and Asia's urban populations stood at only some 15% in 1950. Today they already stand at around 40%.

Urban growth in the Third World is far from uniform. For example, countries including Oman, Botswana, Tanzania, Kenya, Mozambique and Rwanda experienced average annual urban growth rates in excess of 7% between 1975 and 2000 (United Nations, 2004a). Such growth rates imply a doubling of the urban population in under 10 years.

Also, among the urban agglomerations of 10 million inhabitants or more in 2003, Dhaka and Lagos grew at over 6%, Delhi grew at over 4%, Mumbai, Jakarta and Karachi each grew at over 3% on average per year between 1975 and 2000. However, generally it is the medium-sized and smaller towns and cities where the highest population growth rates are found (United Nations, 2004a).

Further, as regards internal migration, there are important differences between sexes and age groups. Males and the younger economically active age groups tend to exhibit the highest propensity to move. In addition, the above figures on net migration conceal circulatory and return movements. Such movements are, for example important among the labour migration in China.

Tens and tens of thousands in the developing world leave rural districts to move to towns and cities every day, day after day, mostly in search of better opportunities or to join relatives. It is a trend towards urbanization which is expected to continue for many decades to come (United Nations, 2003).

The accommodation of such rapid urban growth poses enormous challenges. The urban economies in the developing world are not usually able to provide jobs for many of the migrants. For many, informal sector activities with associated low earnings offer the only opportunity. In Tanzania, for instance, 43.1% of the population aged 15-24 year old in urban areas are unemployed by the national definition (which includes those with a marginal attachment to the labour force). For males, the proportion is slightly better at 36.9%, while for females it is 48.2% or nearly one out of every two (RAWG, 2003).

For many migrants in the Third World, poverty results. Often, first generation migrants end up in shanty towns, squatter settlements and other substandard housing and informal forms of urban shelter in or near existing towns and cities. Frequently, such settlements are densely built and in areas prone to natural hazards, such as in flood plains or on steep hillsides, at risk of flooding, land and mud slides. The construction materials used for dwellings make the settlements highly vulnerable to other hazards as well, such as earthquakes, storms and fires (UN-HABITAT, 2003).

By the definition and estimates of UN-HABITAT (2003), the total urban slum population worldwide in 2001 stood at about 924 million people. It was forecast to exceed one billion by 2005. This means that slightly under 1 out of every 6 of the world's population live in urban dwellings classified as slums. Over 94% of this urban slum population resides in developing countries. Here, over 40% of urban dwellers live in slums. The highest proportions are in Sub-Saharan Africa (over 70%) and in South-Asia (nearly 60%) (UN-HABITAT, 2003).

Urban administrations are confronted with unplanned and uncoordinated urban expansion, land degradation and environmental problems and hazards, including pollution and deforestation. Housing and basic facilities and services, such as access to safe drinking water, sanitation, including waste and sewage disposal, health care and education, and the supply of electricity, are frequently inadequate. Building structures are sometimes unsafe, and overcrowding is common. Sometimes, the land is polluted, or exposed to industrial effluent and noxious waste. Often, residents are faced with land and tenure rights issues and have little security of tenure. Transportation and infrastructure provisions tend to be minimal or absent. The food intake among residents in poor urban areas is frequently insufficient and unbalanced. Political, social and economic inter-community discrimination, exclusion, tension and strife, particularly between migrant and non migrant communities, may occur. Safety issues, such as fires, water-borne and other communicable diseases and crime, need addressing. Risk and disaster management, such as the prevention of flooding and landslides, are pressing concerns (UN-HABITAT, 2002, UN-HABITAT, 2003).

Equally, there are major implications of such large scale population movements for the rural areas. Rapidly expanding urban areas absorb large quantities of agricultural land, often land of good quality and accessibility. In the rural areas, it

is usually the young and strong with most initiative who leave for the cities. The population staying behind is characterized by a skewed age and sex distribution and by high dependency ratios. If only the breadwinner migrates, then family cohesion is stretched. While successful migrants may generate a transfer of wealth to the rural areas, they also bring elements of an urban culture and life style which may clash with the more traditional rural ones. Similarly, they bring urban health problems such as HIV/AIDS to rural areas.

Additionally, on a national scale, there is the movement of people across borders. In many parts of the developing world, international borders have been put in place with little regard for traditional economies or for ethno-tribal and community considerations. Also, traditional nomadic or semi-sedentary life styles, such as shifting cultivation, and modern political entities may not correspond well. Even if economies change, traditional perceptions in people's minds of what constitutes one's homeland, often endure.

Further, the search for better opportunities has led to significant international flows of economic migrants from poorer nations to the developed countries, in particular to the United States and the European Union, and to middle-income countries such as the Gulf States. Some of this migration is legal, some takes place in disregard of official policy and legislation.

Also, regional conflicts continue to lead to refugee flows, frequently across international borders. Many eventually return, but many others are ultimately absorbed and integrated, mostly in towns and cities.

Table 2 gives some indication of the magnitude of international migration and of the recent change in the numbers involved. The data are based on the comparison of country of usual residence and country of birth. These data are therefore the net resultant of processes of international migration flows, processes whose actual size, direction and timing remain hidden. Also, according to United Nations (2002a), the break-up of the former USSR into a number of independent countries accounts for some 27 million persons who in 2000 are classified as international migrants, while formerly being classified as internal migrants within the USSR. Further, international migrants without a legal status in the country of arrival will at best be partially recorded, and estimates of their number vary widely.

Table 2 Size and Annual Growth Rate of Migrant Stock by Major Area, 1990-2000

Major area	1990 Number (thousands)	2000 Number (thousands)	Average annual change 1990-2000 Number (thousands)	Per cent
World	153,956	174,781	2,083	1.27
More developed regions	81,424	104,119	2,270	2.46
Less developed regions	72,531	70,662	-187	-0.26

Source: Absolute numbers for 1990 and 2000 from United Nations (2002a)

According the definition used in table 1, there were some 175 million international migrants in 2000, amounting to nearly 3 per cent of the world's population. Between 1990 and 2000, the number of international migrants grew by some 1.3% per year. Six out of every ten of these migrants reside in the more developed regions, making up some 10% of the population here. For the less developed regions, at about 1.4%, the proportion of international migrants is much lower.

According to United Nations (2002a), 56 million of the world's migrants live in Europe, 50 million in Asia and 41 million in Northern America.

In 2000, about 9 per cent of the international migrants are refugees. Of these, 3 million reside in developed countries and 13 million in developing countries (United Nations, 2002a).

International migration entails the loss of human resources for many countries of origin and may give rise to political, economic, social and cultural tensions in countries of destination. At the same time it may generate valuable contributions to the economies of the receiving countries while generating a transfer of wealth through remittances sent back to the countries of origin. Especially also, successful returning migrants may bring valuable capital, skills and experience back to their original home countries. Examples of the latter process can now be seen in China and India.

While international migration may lead to pressing concerns in the countries and regions affected, the significance cannot be compared with the pace and magnitude of the phenomenon of internal migration and urbanization in the Third World. Recall the net effect of internal migration worldwide, estimated above at an additional 654 million urban inhabitants between 1975 and 2000, or on average some 26 million people per year. This compares to an estimated net effect of

international migration worldwide of some 2 million persons a year on average between 1990 and 2000.

It is useful to note, here, that the data presented in tables 1 and 2 are indicative only. They may be criticized from various angles.

For example, in a global comparative perspective Zlotnik (2002) critically assesses the operational definitions of urban and rural which are used to classify population data in the countries and areas for which the United Nations has evidence. She observes an enduring lack of international comparability, and for many countries even a considerable amount of ambiguity and uncertainty as to the definitions actually used.

As a comment on Zlotnik (2002), we observe that this issue will resolve itself only when census, survey and registration data are properly *geo-referenced using GPS (Global Positioning System) data*, allowing the data user to compile tailor-made regional classifications depending on the specific research objective.

Bocquier (2005) critically examines the method used by the United Nations Population Division in preparing its estimates and projections of the world's urban populations. He tentatively proposes an alternative method as more appropriate, and he shows that this alternative method systematically leads to lower estimates for the world's future urban populations, particularly in the less-developed nations. He also reiterates the request that the United Nations make its original database publicly available for scrutiny and improved projections.

Also, United Nations (2002a) comments on several occasions that the United Nation's data on international migration suffer from serious defects. The available information remains incomplete and often inaccurate, and there are inconsistencies hampering international comparability. One cannot but expect that, for example, data on recent and illegal migrants will be particularly defective.

In the present context, however, the sole purpose of the data presented above is to underline the importance, and to give at least some indication of the order of magnitude, of the two phenomena of internal and international migration. A further in-depth critical assessment of the data presented in tables 1 and 2 is beyond the objective of this paper.

The massive growth of towns and cities in the developing world and the international flows of particularly economic migrants engender wide-ranging issues and challenges which need to be addressed in the interest of harmonious and sustainable urban and rural growth and economic development.

Various countries have attempted to control and stem urbanization through policy measures restricting the free flow of the population. By and large, such policies have proved unsuccessful.

A case in point is China. At the time of the policy of forced agricultural collectivization around the 1960s, the country introduced a system of registered residence, the hukou system. It severely limited the rights to housing, employment, education and health services outside one's local area of registration. In practice, the hukou system was an attempt to tie the rural population to the land.

However, with the country's economic development since the 1980s, the hukou system has broken down completely. Policy changes lag behind reality. Estimates vary, but it is generally assumed that at present, some 200 to 300 million rural people, or about one in every five of all Chinese citizens, are living and working in cities and towns, often illegally and in violation of official hukou policy. They are called China's "floating population".

Official Chinese statistics on the urban-rural population distribution do not reflect these mass movements. All official statistics record a person's place of residence on a de iure basis only. They fail to record the de facto population distribution. Officially, the floating population does not exist. However, it is these internal migrants who are a principal source of the cheap labour which is so fundamental to eastern China's economic miracle. Equally, through the transfer of earnings, this floating population constitutes one of the main channels through which China's seaboard economic success and wealth trickle down to the country's poor rural interior.

Generally, the only realistic alternative for developing countries is the accommodation of the current massive and uninterrupted flows of millions into towns and cities.

It requires the elimination of any unnecessary impediments to, and the encouragement of, enterprise in the formal and informal sectors. Enterprise constitutes the foundation of job and wealth creation, and it is the cornerstone of urban economic development and agricultural modernization.

At the same time, it requires good, effective, and efficient public governance. In towns and cities it calls for the provision of housing, basic infrastructure and facilities and services such as health care and education for the urban immigrants. Given the size of these population movements, this is a truly immense task.

To ensure the effectiveness and the efficiency of policies and programmes dealing with urban growth and development, it is essential, first and foremost, to have timely, valid and reliable empirical information. Policies and programmes

without a solid basis in empirical reality can all too easily be misdirected and wasteful, and scientifically they are of questionable legitimacy.

Given the nature of Third World urbanization, a monitoring system accurately measuring the underlying dynamics of population growth and migration on an ongoing basis or at least at regular intervals is a core ingredient of information-based policy making and programme design, implementation and evaluation.

While demography offers proven instruments to estimate mortality and fertility in developing countries with incomplete and defective data, this is much less the case with migration. In this paper we shall address this issue of the measurement of migration flows and of migration information system establishment. In the following section, we shall briefly assess the status quo in this subdiscipline of demography.

2 THE DEMOGRAPHIC MEASUREMENT OF MIGRATION: A REVIEW OF CURRENT PRACTICE

Until the 1970s, demographic analysis was largely concerned with the study of the structure and change over time of closed national populations (Smith and Keyfitz, 1977; Coale, 1972; Henry, 1972; Keyfitz, 1977; Pressat, 1983). Migration was rarely regarded as a key variable in the core demographic paradigm. Mortality and fertility were the principal variables of change considered. In mathematical terms, theory and models were scalar.

This placed serious limits on the usefulness of demography and demographic analysis in policy making and planning. Population change as a result of the forces of mortality and fertility is a relatively slow process with predominantly long-term implications. For the short- and medium-term planning and provision of, for example, housing, health care and other services within an urban or regional context, such information on long-term population change is useful but insufficient. As seen in section 1, in many parts of the world, short- and medium-term population change on a local scale is highly dependent on migration. Often, migrants constitute a very dynamic element of the population, with a specific age and sex distribution and with specific social and economic characteristics and behaviour.

The key breakthrough in the traditional demographic paradigm came with the increased emphasis on the analysis of migration from the late 1960s. The principal development was the attempt by Rogers (1975) to generalize by analogy elements of the work of Keyfitz (1968b). It was based on the recognition of the formal similarity between mortality and outmigration in respect of their effects on a population. Essentially, the approach taken by Rogers (1975) was the substitution of suitable formulations from linear algebra for the scalars used by Keyfitz (1968b).

It led to successful generalizations of theory which were first denoted by the terms of spatial population analysis and multiregional demography. When the generality was better understood (as, for example, in Willekens, 1980; Willekens et al, 1982; Al Mamun, 2003), the more general terms of multidimensional demography and, now more commonly, multistate demography came to be used.

Yet, a major problem remained, namely the demographically consistent measurement of migration. Proper measurement is based on the establishment of unambiguous, valid and reliable relationships between theoretical concepts and empirical reality.

However, demography has deep roots in applied research. Consequently, there is a strong tradition to take data as a point of departure, and to develop concepts, analytical tools and applications in a bottom-up procedure from there. Burch (2003) discusses this issue at some depth, exploring the role of theory formulation in demography.

Matching existing data types and theoretical strands in a data-oriented and application-driven approach can easily lead to arguments which are unnecessarily approximate or unnecessarily complex. Building on earlier work by Keyfitz (1966, 1968a, 1968b, 1970), Keyfitz and Flieger (1971), for example, resort to an iterative numerical algorithm to reconcile empirical mortality rates from which a life table is derived on the one hand and the resulting life table mortality rates on the other. This is a device which, at least from a theoretical point of view, is a less than optimal. And, for instance, the ubiquitous Lexis diagram as a tool to map data on concepts so as to reconcile a cohort and a period perspective -- a device which in the multistate case can all too easily become quite tedious -- is generally unnecessary if theory is carefully constructed first. This is because properly formulated theory implies data definition.

In addition, migration data recorded in censuses, surveys and administrative systems usually suffer from serious age- and sex-specific underenumeration, often even in statistically developed countries. To gain reliable information, estimation and adjustment procedures are needed in order to correct for such deficiencies.

For countries with defective statistics on age distributions, mortality and fertility, estimation and adjustment procedures were first developed at the United Nations in the 1950s. And since, they were expanded and improved to considerable maturity; see, for example, Lederman (1969), Coale and Demeny (1966) and the improved Coale et al (1983), United Nations (1967), the much enhanced United Nations (1983), United Nations (2002b) and United Nations (2004b).

Migration is a demographic event which shares many of its characteristics with mortality. Importantly, however, in a person's life history, migration must by definition be regarded as a potentially-recurring rather than a once-only event, and this characteristic complicates the measurement of migration as an event.

Also, there is the issue of competing events. So, for example a person subject to forces of mortality and outmigration cannot experience the event of mortality in the region of interest when the event of outmigration has already occurred. And conversely, when the event of mortality has occurred first, the person can no longer experience the event of outmigration. The events of mortality and outmigration mutually compete.

As a consequence, historically, approaches to the measurement of migration have lagged behind in sophistication in comparison with approaches to the

measurement of mortality and fertility (United Nations, 1970; Courgeau, 1980; Courgeau, 1988; United Nations, 1998).

One special issue which plays a role is the fact that *public accessibility of data* on internal and international migration is, at least in practice, much more restricted that is the case for data on mortality and fertility. This seriously hampers creative research in the field of migration. Even meta-information, such as a comprehensive insight into which types of data have actually been collected in the various countries of the world, is difficult to come by, let alone the data themselves.

One example of a resource outlining available data is Nam et al (1990). This is an edited volume providing a systematic and relatively comprehensive review of national sources and data on internal migration as available in the 1980s for a selection of 21 countries around the world. The data for each country are also analysed using a standard analytical format.

Rees and Kupiszewski (1996) detail the sources and data on internal migration which are available for 28 countries which are member states of the Council of Europe.

Bell (2005) describes a more recent and ongoing effort to compile a comprehensive database specifying the data on internal migration collected by United Nations member countries worldwide, and detailing the sources used. In doing so, the resulting compendium on sources and data usefully fills a void left in recent years by international agencies such as the United Nations. Unfortunately, metadata on international migration remain excluded. As of the time of writing, the database can be consulted at <http://www.geosp.uq.edu.au/qcpr/database/IMdata/Imdata.htm>. It is updated from time to time as more information becomes available.

As a comment on Bell (2005) we make two notes, here. The principal motivation for establishing the repository is the desire to enable meaningful cross-national comparisons of the internal migration as experienced by different countries. In this context, Bell observes that approaches used in different countries vary widely, hampering international comparison. While we respect the academic significance of cross-national comparisons, the prime motivation for data collection and for the study of migration has to be local to each country, region, city or town. From its perspective, section 1, above, underlines this priority as well.

Second, anticipating our subsequent discussion, we have to differ with Bell (2005) in one important respect. At least implicitly the paper bears witness to an emphasis on census and survey questions on the place of usual residence some fixed number of years in the past. This is a generic type of question which we shall demonstrate to be of inferior theoretical and methodological value.

Next, let us briefly review the established methods of measurement.

As to internal migration, indirect balancing methods are often advocated for statistically less-developed nations. Here, migration is *defined as* the error or deviation between closed population projections in the absence of migration on the one hand and corresponding empirically-observed population data on the other. Such methods result only in net migration estimates, and circumvent the need to measure migration itself. Data on population stock, fertility and mortality suffice. Magnitude, direction and timing of the underlying migratory processes are not revealed, however.

Indirect estimates of net migration based on such residual analysis also suffer from the major drawback that they render the explanation of migration an elusive affair, both theoretically and methodologically. Net migration is not in itself an empirical phenomenon. It is an abstract concept defined in terms of differences between empirical migratory flows. As mentioned, indirect estimates of net migration reveal the values of these differences only. The magnitudes, directions and timings of the flows themselves remain unobserved and hidden from view. Associating explanatory covariates with residual or net flows can lead to severe difficulties in interpretation.

Methodologically, it is more common to define residuals as unexplained variation by associating them with all relevant variables which have been left out of the analysis. Of course, it is possible to measure the association between the net flows and explanatory covariates. However, when doing so, it is generally problematic to identify what empirical phenomenon has actually been explained other than the balance between unobserved variables which themselves may have taken an unlimited set of values in terms of magnitude. The problem is, of course, exacerbated both if the number of migration defining areas is allowed to increase and if time-varying covariates are considered relevant in the explanation.

In order to obtain an insight into actual population movements, that is, into both migration as an event and numbers of migrants, direct measurement is required. Here, an historically insufficient integration of the concept of migration within mainstream mathematical demographic thought has allowed for considerable ambiguity in respect of what constitute proper approaches to measurement.

For example, the principles and recommendations for population and housing censuses of the United Nations (United Nations, 1997) suggest four different questions for the direct measurement of migration. In addition to the place of usual residence, they are: (1) the place of birth; (2) the duration of residence in the current place of usual residence; (3) the place of previous usual residence;

and (4) the place of usual residence at a specified date in the past, in most cases one year or five years preceding the census (United Nations, 1997).

For the measurement of international migration, the United Nations suggest three additional questions, namely (5) the country of birth; (6) the country of citizenship; and (7) the year or period of arrival in the present country (United Nations, 1997).

Clearly, questions (1) and (5) are similar, while questions (2) and (7) adopt a similar approach, albeit not necessarily with respect to the same migratory event.

In the recommendations it is recognized that, for example, a question on duration of residence is only of limited value in itself because it does not provide information on the place of origin of inmigrants. However, a rigorous systematic and theoretically justified comparative analysis of the merits and demerits of the various possible methods of measuring migration is not presented.

In terms of question choice and formulation, an important objective of the recommendations is the obtaining of *reliable* answers from respondents.

A key issue in formulating questions then is if people can remember accurately what happened, and, if appropriate, when and where. This explains, for example, the choice of question (5), country of birth, for the measurement of international migration. In most circumstances, answers to this question are likely to be accurate to a high degree.

Further, *practical* considerations play a role in the choice and formulation of questions on migration. For example, in several countries the theme of foreigners is an important policy issue. The recommendation of question (6) on citizenship follows principally from this consideration, rather than from its demographic merits as an instrument for the measurement international migration.

The issue of *validity*, that is, the issue of the concept of migration which one desires actually to measure, plays much less of a role in the recommendations. In order to be able to apply demographic theory in analysis and forecasting, this issue of validity is a central one as well, however.

As a consequence, while at least there appears to be some element of redundancy in the questions recommended, to national statistical offices and others involved in the census process it will not immediately be obvious which of the approaches to direct measurement should best be chosen.

For instance, UNECE (2005) itemizes a number of uncertainties in the field of the measurement of migration. Further, it states "The place of usual residence one year prior to the census … is well suited for internal migration." (UNECE, 2005,

p 5) As we shall see in this paper, both from a theoretical point of view and in terms of informative value, this statement is erroneous.

And so, as to the direct measurement of migration, that is, the measurement of gross migration flows, the debate on the issue of the preferred migration questions in population censuses and surveys remains inconclusive. As we shall argue, this is primarily due to this lack of understanding of the fundamental relationship between theory and measurement.

Further, the direct measurement of migration flows is notoriously subject to errors of incompleteness. This is an important issue which is not explored in any great depth in the seminal methodical manuals of the United Nations (United Nations, 1970; United Nations, 1998). United Nations (1998), for example, outlining methods of measuring international migration, limits itself principally to definitions, data sources and tabulations. While there is occasional reference to completeness of the data, the matter of how incomplete and defective data might be corrected is not addressed.

There is yet another issue that constitutes an obstacle in the development of demographically sound methods of measuring migration. It is the selection of descriptive concepts of weak analytical power as the basis of some multistate theory development, in particular the concept of demographic accounts (Rees and Wilson, 1977). This has led to unnecessarily complex measurement arguments.

Demographic accounts find their origins in a predominant focus on empirical data as the point of departure for theory construction. They constitute a basically non-mathematical arithmetic and descriptive framework which essentially focuses on system states in terms of enumerated population numbers by age and sex at distinct places and instants of time. It is a framework which puts a heavy emphasis on the aggregate net transitions from place to place between such discrete instants.

Rees and Wilson (1977) represented their demographic accounting approach as a paradigm shift away from the thinking based on mathematical concepts such as rates and probabilities as represented by Keyfitz (1977).

This emphasis on demographic accounting has brought many analysts to lend strong support to population census and survey migration questions of the type "place of usual residence at $t - \tilde{a}$", where t represents the instant of measurement and $\tilde{a}$ represents a fixed number of years, usually either 1 or 5 (Rees, 1984; UNECE, 2005). Unfortunately, however, this is a measurement instrument of comparatively limited analytical scope and of poor informative value both from a mathematical and from a demographic standpoint.

This is inherent in the fact that this question measures aggregate *net transitions* in discrete time. Mathematically, time-continuous analysis is more powerful than discrete analysis. Demographically, net transitions do not describe migration itself (the events or movements), but only the resultant net effect of migration on population structure and development over a given time interval.

As a consequence, multiple moves within the $\tilde{a}$-year time interval are not recorded. This includes step or staged migration, where a migrant moves from origin to destination via one or more intermediate destinations, as well as return migration. Return migration is ignored altogether, while in the case of step migration, the origins of migrants are misrepresented which makes a proper interpretation of migration impossible.

These are drawbacks not dissimilar to the analytical downside of the use of place of birth data, country of birth data and citizenship data for the measurement of migration. Such data, too, essentially measure net transitions only. They do not reveal the actual population dynamics in time and space.

Rogers (1973, 1975) touched on the measurement issue by attempting to develop by analogy model multiregional life tables on a par with the by then well-established classical model life tables of Coale and Demeny (1966). Subsequently, under Rogers at the International Institute for Applied Systems Analysis (IIASA), considerable effort was devoted to the theme of measurement.

Here, attempts were made primarily to generalize existing and highly-successful approaches developed for mortality and fertility, in particular the work of Coale and Demeny (1966) and Coale and Trussell (1974). Considerable progress was made in the investigation of migration schedules by age and sex (Rogers and Castro 1981) and in the study of procedures to estimate detailed distributions from aggregated marginal distributions (Willekens, 1999; Schoen and Jonsson 2003).

However, emulating the achievements made in the areas of mortality and fertility at Princeton under Coale by developing model schedules and derived methods allowing the indirect estimation of migration events and gross numbers of migrants proved elusive (Rogers, 1973; Rogers, 1975; Rogers and Castro, 1976; Rogers and Castro, 1981; Rogers, 1999). This is not surprising, since the mechanisms underlying the schedules as well as the causes of data incompleteness and errors vary considerably between mortality, fertility and migration. The stability and hence predictability of, for example, mortality schedules is due to the fact that their general *shape* is principally governed by biological factors and medical technology, with schedule *levels* primarily determined by levels of economic development and social equality.

Migration lacks such a stable and predictable biological and medical basis. Here, economic and social factors constitute the principal forces determining patterns

and levels. Rogers and Castro (1981) describe broadly-general migration rate schedules by age. As it appeared, these are bi- or trimodal distributions. They are characterized by an absolute maximum at the age of higher education and early labour market participation; a derived maximum at the youngest ages, associated with child birth among migratory early labour market participants; and sometimes a modest peak at the ages at which the taking of retirement is common.

Even though such broadly-general migration schedules by age may be recognized, variability over place and time is very much higher than it is in the case of mortality. And so, for example, in order to obtain an adequate fit for their migration models to empirically-observed data, Rogers and Castro (1981) required, what Rogers (1982) called, some prior "data massage". An additional disturbing cause here may well have been the fact that, as we shall see later, empirical migration data tend to be quite defective.

In addition, for an adequate fit Rogers and Castro (1981) required mathematically complex model specifications, specifications in which the number of parameters to be estimated approached the number of five-year age groups commonly used in statistically less-developed countries.

This variability and mathematical complexity effectively preclude the application of general and robust calibration parameters and model schedules to estimate or adjust gross migration data in the case of severe data deficiencies. And so it proved to be fundamentally flawed to take a methodological analogy with mortality and fertility modelling and data estimation and adjustment as an approach to the estimation and adjustment of migration data.

The matter of estimation and adjustment was left unresolved. In fact, this was one reason why the demographically-alien net transition approach from the demographic accountants received relatively broad acceptance.

Building on the contribution of the present author to UNESCAP (1982) and on Doeve (1987), in this paper we aim to contribute to the two fields of mathematical theory construction and of measurement methods, since the two are intimately related.

In the next section, we shall develop an elementary but carefully-argued rigorous axiomatic-deductive theoretical mathematical framework for the study of migration. It is a framework which ties in fully with the modern standard demographic paradigm focusing on the instantaneous rates at which individual demographic events occur in continuous time.

It is a formulation of established approaches in demography and in formally-related fields which is developed with a view to being able to derive valuable and general insights and results. Additionally, the formulation is general in the sense

that it applies not only to the study of internal and international migration, but equally to the study of mortality and fertility.

Next, once the development of the elementary theory is complete, we shall investigate how to establish relationships between abstract theory and empirically-observable events. In fact, the mathematical theory will be shown to lead to valid measurement methods in an unambiguous and surprisingly simple fashion. It provides clear guidelines on appropriate methods of measurement, resolving the issue of which questions are both theoretically best justified and of the highest informative value.

We shall demonstrate the power of this method by using empirical migration data for Bangkok which are not necessarily fully complete nor fully without error. And we shall show how such deficiencies can be corrected in a theoretically-justifiable manner in the general case where one is dealing with deficient data sets.

In fact, this approach to measurement is the first truly demographic method of measuring and adjusting migration data in populations for which data are incomplete and defective, such as in statistically less-developed countries or in the case of illegal migration.

3 AN ELEMENTARY THEORY OF COHORT BEHAVIOUR IN CONTINUOUS TIME

3.1 INTRODUCTION

Demography is the social science which describes and explains the generation and the behaviour over time and age of human cohorts.

As a social science, in its description and explanation demography focuses on cohorts as groups, not on the behaviour of individual members of a cohort.

Explanation belongs to the realm of subdisciplines such as economic demography and social demography. Here, we shall concern ourselves mainly with description. We note that all subdisciplines may be involved with forecasting or projections, extrapolating past experience under sets of well-defined assumptions or scenarios. Since the development of theory from first principles is uncommon in demography, we aim to be rather more explicit than usual.

Apart from age, cohort members may be defined as having other attributes, such as sex status, alive status, marital status, birth status, parity status, migration status (usual residence status), migration frequency status, health status, employment status, and so on.

Each attribute takes well-defined values. So, for example, the sex status might take the two values male and female; the alive status might take the two values alive and not alive; the usual residence status might take region 1, region 2, and region rest-of-the-country as its values; and so on. In practice, generally, the allowable values will be determined by the empirical context and by the perspective taken on that context.

In any given context, some of these attributes may be or may be taken as constant, other ones as variable.

In line with demographic tradition, we shall use the term status for a cohort attribute. Further, in abstract formal approaches, attribute values are more usually called states. We limit ourselves to statuses whose values are either finite or countably infinite. Given a set of statuses under study, then the collection of allowable values of these statuses make up the state space of the cohort. In the above example of the alive status, the state space is called binary since it can take only two values.

We note that the term state is also sometimes used to describe the distribution of the cohort over the state space at any one given point in time.

In the present discussion, the term event is reserved for the experience by a cohort member of a change in a status value.

Consider a cohort of a given exact age. Initially, we shall focus on a single status, omitting consideration of all other statuses. For this discussion, the demographic nature of the status is not material, although in the context of the present paper one might reasonably consider it to be the migratory status.

Further, initially, we shall focus on the values of frequency statuses, that is, on the number of events experienced by a cohort member. So, for example, in the case of migration, one may alternatively consider the migration frequency status and the migration status. In the first case, status values are the number of migratory events experienced by a cohort member, while in the second case, status values are admissible places of usual residence.

This initial viewpoint will prove convenient since it provides a number of useful insights of a more general nature and leads to valuable instruments. Further, changing perspective from the number of events to the events themselves is straightforward. So, in the case of migration, we shall discuss the event of place change from one specified place of usual residence to another specified place of usual residence once we have completed our discussion of the migration frequency status. Finally, we shall briefly consider multiple statuses, status values and competing events.

We shall adopt a stochastic approach. We do not assign motives or behaviour to individual cohort members in any deterministic manner. Instead, we map a probability measure on each of the members of the cohort. In our initial approach, this measure represents the magnitude of the risk at which cohort members are of experiencing any given number of occurrences of the event. Thus, it is a measure of exposure to risk. While this does not allow us to say much about the behaviour of any given individual cohort members, it does allow us to arrive at powerful expressions representing aggregate cohort behaviour.

Let $\mathbf{N}$ denote the set of natural numbers and $\mathbf{R}$ denote the set of real numbers. We define continuous variable time as the set $\{t\} = \mathbf{R}\backslash\mathbf{R}^{-}$. Let the number of individuals within the cohort who have experienced the event under consideration exactly n times, $n \in \mathbf{N}$, during some time interval $[0, t)$, $\forall\, t \in \mathbf{R}\backslash\mathbf{R}^{-}$, be denoted by integer function $K_n(t)$. So, $\forall\, n, t$: $K_n(t) \in \mathbf{N}$. Note that, of course, $\{0\} \subset \mathbf{N}$.

Further, let $\mu(t)$ denote the instantaneous rate, defined as a continuous function of t, at which the event occurs. Some alternative terms to denote $\mu(t)$ are the propensity to experience the event, the force or intensity with which the event is experienced, and the hazard rate or the hazard function. $\mu(t)$ is a measure of the intensity at which the event occurs at instant t. Since the number of events occurring at any instant t is at least zero, we have that $\forall\, t:\ \mu(t) \in \mathbf{R}\backslash\mathbf{R}^{-}$.

Finally, let $P_n(t, t + a)$, defined as a continuous function of t, for all $a \in \mathbf{R}^{+}$ denote the probability that the event in question occurs exactly n times, $n \in \mathbf{N}$, to members of the cohort during a time interval $[t, t + a)$. Hence, here n is the discrete stochastic (or, random) variable of interest. If we are referring to a small time interval, we frequently use the notation Δt for a. Further, an integer inequality in the index, such as $n > r$, n, $r \in \mathbf{N}$, in $P_{n>r}(t, t + a)$, refers to the occurrence of the event exactly $(r + 1)$ or more times.

We note that, in general, a probability such as $P_n(t, t + a)$ is a conditional measure, namely upon having reached instant t. Only if $t = 0$ then the measure is unconditional. In the case of unconditional probabilities, we shall omit the first argument. So, for example, $P_n(a)$ would be understood to mean $P_n(0, a)$. Additionally, we shall denote the limiting value of $P_n(a)$ as $a \to 0$ by $P_n(0)$. Stated rather informally, the notation $P_n(0)$ refers to P_n at instant $t = 0$, that is, over the zero-length time "interval" [0,0].

Next, we formulate three postulates as the axiomatic framework of the theory of cohort behaviour. In principle, the formulation of axiomatic postulates is an arbitrary matter. Unless they can be proven to be equivalent, alternative sets of postulates will lead to a different theory. The postulates below have been chosen because they are a smallest set of necessary and sufficient postulates which lead to theory which will be recognized as embracing the standard demographic paradigm.

This paradigm has well-proven empirical validity and applied value. Aside from the formal sciences, empirical validity is an essential criterion in theory development. This is not to say, however, that an alternative formulation of postulates might not lead to theory with an even greater empirical validity. By formulating the postulates explicitly, we at the same time establish a benchmark for comparative validity testing of any such alternative sets of postulates.

3.2 THE BASIC POSTULATES

Postulate 1

$$P_1(t, t + \Delta t) - \mu(t)\cdot\Delta t = o(\Delta t), \tag{1}$$

or, $P_1(t, t + \Delta t) = \mu(t)\cdot\Delta t + o(\Delta t)$, where $o(\Delta t)$ is some continuous function of Δt defined by

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0 .$$

The definition of $o(\Delta t)$ is a formal way of stating that the numerator in the ratio $o(\Delta t) / \Delta t$ is of a smaller order of magnitude than the denominator as $\Delta t \to 0$.

Thus, this postulate states that the difference between probability $P_1(t, t + \Delta t)$ and $\mu(t)\cdot\Delta t$, a quantity proportional to the duration of the exposure, becomes negligible relative to the size of Δt as $\Delta t \to 0$. More precisely, this postulate describes that probability $P_1(t, t + \Delta t)$ approaches the product $\mu(t)\cdot\Delta t$ asymptotically as $\Delta t \to 0$.

Postulate 2

$$P_{n>1}(t, t + \Delta t) = o(\Delta t). \tag{2}$$

This postulate effectively implies that events are mutually exclusive as $\Delta t \to 0$.

Postulate 3

$$\forall\, n_1, n_2 \in \mathbf{N},\ \forall\, a_1, a_2 \in \mathbf{R}^+,\ \forall\, t_1, t_2 \colon t_2 > t_1 + a_1,\ t_1, t_2 \in \mathbf{R}\backslash\mathbf{R}^- :$$
$$P_{n_2 \mid n_1}(t_2, t_2 + a_2;\ t_1, t_1 + a_1) = P_{n_2}(t_2, t_2 + a_2), \tag{3}$$

where $P_{n_2 \mid n_1}(t_2, t_2 + a_2;\ t_1, t_1 + a_1)$ denotes the conditional probability of experiencing the event exactly n_2 times during a time interval $[t_2, t_2 + a_2)$, given the experiencing of the event exactly n_1 times during a time interval $[t_1, t_1 + a_1)$.

And thus, this postulate states that the numbers of occurrences of events in any of two non-overlapping or disjoint time intervals are mutually stochastically independent.

These three postulates cast the frequency of occurrence over time of the event among cohort members as a stochastic Poisson process, one of the well-known counting processes of numbers of events occurring over time.

Probabilities such as P_n which are a function of some $\mu(t)$ are called non-stationary probabilities unless $\mu(t)$ is time-invariant. If $\mu(t) = \mu$, that is, if $\mu(t)$ is constant, then a probability P_n is called stationary. A stochastic process based on non-stationary probabilities is said to be a non-stationary stochastic process. Sometimes the terms inhomogeneous or non-homogeneous stochastic process are used here.

Further, a stochastic process embodying postulate 2, excluding the simultaneity of events, is called ordinary or orderly.

Finally, a stochastic process incorporating postulate 3 is called non-hereditary, memoryless or without after-effect; the process is said to have the Markov property. Postulate 3 implies, for example, that information on cohort behaviour on past intervals does not contribute to improving predictions made about behaviour on any subsequent intervals.

From this axiomatic framework, we shall now derive a number of fundamental theoretical results.

3.3 SELECTED FUNDAMENTAL RESULTS

Theorem 1

Recall that $P_0(t)$ is the probability of zero events occurring during the time interval $[0, t)$. Then

$$P_0(t) = \exp(-\int_0^t \mu(u)du), \tag{4}$$

where $t \to \exp(t)$ denotes the exponential function $t \to e^t$.

Proof. First, consider the probability $P_0(t + \Delta t)$ that zero events occur during the slightly longer time interval $[0, t + \Delta t)$. This number of occurrences of the event can come about in only one way, namely, if we first have that zero events occur during $[0, t)$, followed by zero event occurring during $[t, t + \Delta t)$. Using postulate 3 stating that the numbers of occurrences of events in any of two non-overlapping time intervals are mutually stochastically independent, we have that

$$P_0(t + \Delta t) = P_0(t) \cdot P_0(t, t + \Delta t). \tag{5}$$

Next, also consider the occurrence of one or more events $(n = 1, 2, \ldots)$ during $[t, t + \Delta t)$. If $n = n_1$ events occur during $[t, t + \Delta t)$, $n \in \mathbf{N}$, then $n \neq n_1$ events will not occur. So, the realizations of each of the possible status values of $n = 0$ events, $n = 1$ event, $n = 2$ events, …, during $[t, t + \Delta t)$ are mutually exclusive. Further, the denumeration for all $n \in \mathbf{N}$ is exhaustive.

Therefore, using postulates 1 and 2, we have that

$$P_1(t, t + \Delta t) - \mu(t) \cdot \Delta t + P_{n>1}(t, t + \Delta t) = 2o(\Delta t) = o(\Delta t), \tag{6}$$

and, using the addition axiom from probability theory, that the sum of the probabilities of 0, 1, 2, 3, … events during $[t, t + \Delta t)$ adds to unity, that is,

$$P_0(t, t + \Delta t) + P_1(t, t + \Delta t) + P_{n>1}(t, t + \Delta t) = 1. \tag{7}$$

Hence, using (6),

$$\begin{aligned} P_0(t, t + \Delta t) &= 1 - P_1(t, t + \Delta t) - P_{n>1}(t, t + \Delta t) \\ &= 1 - \mu(t) \cdot \Delta t + o(\Delta t). \end{aligned} \tag{8}$$

Substitution of (8) in (5) then yields

$$\begin{aligned} P_0(t + \Delta t) &= P_0(t)\cdot\{1 - \mu(t)\cdot\Delta t + o(\Delta t)\} + o(\Delta t) \\ &= P_0(t) - P_0(t)\cdot\mu(t)\cdot\Delta t + o(\Delta t). \end{aligned} \tag{9}$$

Rearranging (9) and dividing by Δt, we obtain

$$\frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = -\mu(t)P_0(t) + \frac{o(\Delta t)}{\Delta t}. \tag{10}$$

Passing to the limit $\Delta t \to 0$, we obtain the important ordinary first-order linear homogeneous differential equation

$$\frac{d}{dt}P_0(t) = -\mu(t)P_0(t). \tag{11}$$

Further, we have

$$\forall\ n \in \mathbf{N}^+ :\ \lim P_n(t, t + a) = 0 \ \text{ as } \ a \to 0, \tag{12a}$$

$$\lim P_0(t, t + a) = 1 \ \text{ as } \ a \to 0. \tag{12b}$$

For $n = 1$, (12a) follows directly from postulate 1, and for $n > 1$ (12a) follows directly from postulate 2. And, using the addition axiom from probability theory, we have (12b) by implication. This is, of course, a result well known to all who are familiar with probability density functions.

So, for differential equation (11) we have the initial or boundary condition

$$P_0(0) = 1, \tag{13}$$

and hence equation (4) as its solution. This completes the proof of theorem 1.

Theorem 2

Recall that $P_{n>0}(t)$ is the probability of one or more events occurring during the time interval $[0, t)$. Then

$$P_{n>0}(t) = 1 - \exp(-\int_0^t \mu(u)du). \tag{14}$$

Proof. $P_{n>0}(t)$ is the complement of $P_0(t)$.

We note that, for instance, in the analysis of mortality, theorems 1 and 2 are applied as a matter of course in classical life table construction. Theorems 1 and 2 are also well-known results from survival analysis and event history analysis, analytical approaches which have emerged in areas such as the biomedical sciences and engineering and which are, at least formally, intimately related to demography.

Examples are the study of the effects of medical intervention, failure analysis of mechanical devices, and so on.

There is an equally close relationship with the field of operations research, for instance, in waiting time analysis.

More recently, economics and econometrics have developed an interest in this area, particularly in labour force analysis, such as in the study of unemployment duration.

However, in all these fields, the initial approach tends to differ. While our initial focus is on stochastic variable n, the number of events experienced by individual cohort members, the point of departure in, for example, survival analysis is on another stochastic variable, namely τ, the *time interval* between successive events. Specifically, most commonly the approach is narrower than this, in that the focus is on stochastic variable τ_1 as the time interval until the first event (which corresponds to event count value $n = 1$).

Clearly, n and τ are closely related: they are merely another perspective on the same process. The knowledge of P_n as a continuous function of time fully determines $P\tau$, and, conversely, the knowledge of $P\tau$ fully determines P_n. We shall consider stochastic variable τ in theorem 5, below. We shall return to the focus on the restricted state space $\{0, 1\}$ for stochastic variable n, later, as well.

In order to facilitate cross-disciplinary work, we give some standard terms, notations and interpretations.

In survival analysis, common alternative notations are $F(t)$ for $P_{n>0}(t)$, $S(t)$ for $P_0(t)$, and $\lambda(t)$ and $h(t)$ for $\mu(t)$. Consequently also, in survival analysis, $-\frac{d}{dt}P_0(t)$, the derivative of the complement $(1 - P_0(t))$ of the survivor function, is commonly denoted by $f(t)$. Further, here, $P_0(t)$ is called the survivor function or survival function, and hazard function is the common term for $\mu(t)$. The integral $\int_0^t \mu(u)du$ is called the cumulative hazard, the cumulative risk or the integrated hazard. It is properly denoted by the capitalized version of the symbol used for the hazard function, such as $M(t)$, $\Lambda(t)$, or $H(t)$.

Theorem 1 and its proof lead to some familiar notions. For example, observe that from (4) we have

$$\mu(t) = -\frac{d}{dt} \ln P_0(t) \tag{15}$$

or

$$\mu(t) = -\frac{\frac{d}{dt} P_0(t)}{P_0(t)}, \tag{16}$$

results often encountered in survival analysis, as well.

From (11) or (16) we see that $f(t)$ may simply be obtained by multiplying $\mu(t)$ and the survivor function $P_0(t)$. Since in survival analysis $P_{n>0}(t)$ is a distribution function, $f(t)$ is a probability density function, specifying the density of events at instant t. So, from (16) we have that $\mu(t)$ equals the density of events at instant t, conditional on not having experienced any event during interval $[0, t)$.

After this brief cross-disciplinary review, let us now return to our own line of theory development for the general case where the state space is not restricted to the limited set $n \in \{0, 1\}$, but instead unrestricted, that is, $n \in \mathbf{N}$. Then we next have the following important theorem.

Theorem 3 (the general theorem)

Recall that $P_n(t)$ is the probability of exactly n events occurring during the time interval $[0, t)$. Then

$$P_n(t) = \frac{(\int_0^t \mu(u)du)^n}{n!} \exp(-\int_0^t \mu(u)du) \tag{17}$$

A proof by mathematical induction for the sequence $n = 0, 1, 2, \ldots$ using the standard approach employed in the proof of theorem 1 is elementary and left to the reader. Here we only note that setting $n = 0$ results in theorem 1, thus proving theorem 3 for $n = 0$.

Theorem 3 is an important result where we have events which are potentially of a repetitive nature. Examples are births and migratory moves. This theorem then allows us to formulate the probability that a cohort member will experience 0, 1, 2, … such events on any given interval $[0, t)$. And by using conditional probabilities in a way similar to our approach in the proof of theorem 1, this can then be extended to such probabilities any given interval $[t, t+a)$.

Next, consider any interval $[0, t)$ of interest and let us postulate stationary probabilities on this interval. By definition, this equates to postulating that

$$\forall\, t \in \mathbf{R}\backslash\mathbf{R}^{-}: \qquad \mu(t) = \mu, \tag{18}$$

that is, to postulating a constant or time-invariant hazard rate $\mu(t)$. In this case simple and familiar results emerge. We then have

Theorem 1A

$$P_0(t) = e^{-\mu t}, \tag{19}$$

Theorem 2A

$$P_{n>0}(t) = 1 - e^{-\mu t}, \tag{20}$$

Theorem 3A

$$P_n(t) = \frac{(\mu t)^n}{n!} e^{-\mu t}. \tag{21}$$

Theorem 3A describes the distribution of stochastic variable n as the well-known Poisson distribution or Poisson probability mass function (pmf) with parameter $\mu \cdot t$.

The general equation (17) is the formulation of a non-stationary Poisson pmf. It is much less well known but by its generality it is considerably more powerful in theory construction and applied analysis. It is, for example, briefly mentioned by Courgeau (1980), formulated in terms of integrated hazards. However, in this work Courgeau makes no further use of this important result.

While stationary probabilities may not seem empirically valuable, they are at least analytically useful, for example, if it is analytically convenient to break down longer time intervals into smaller subintervals, each so small that it is reasonable to assume piecewise stationary probabilities on each individual subinterval.

Note that equation(18) is equivalent to postulating that $\frac{d}{dt}\mu(t) = 0$. If, on the other hand, for any value of t, $t \in \mathbf{R}\backslash\mathbf{R}^{-}$ we have that $\frac{d}{dt}\mu(t) > 0$ or $\frac{d}{dt}\mu(t) < 0$, then we speak of an increasing and decreasing hazard at t, respectively.

Next, we turn to a new concept, namely, that of cohort mass.

3.4 COHORT MASS

Let us next define a continuous function $Q_n(t)$ of t, such that $\forall\, t \in \mathbf{R}\backslash\mathbf{R}^- : Q_n(t) \in \mathbf{R}\backslash\mathbf{R}^-$, representing a measure of the mass of the cohort at instant t. Specifically, $Q_n(t)$ denotes that part of the total mass of the cohort which has experienced the event under consideration n times over a time interval $[0, t)$.

Of course, $\forall\, n_1, n_2 \in \mathbf{N} : n_1 \neq n_2$, a part of the mass of the cohort that has experienced the event n_1 times on a given interval $[0, t)$, cannot also have experienced the event n_2 times on that same interval. So, the parts of the mass of the cohort as defined are disjoint or non-overlapping. And hence, summing exhaustively over all parts gives us the total mass of the cohort at instant t. Denoting this total mass by $Q(t)$, we thus have

$$Q(t) = \sum_{n=0}^{\infty} Q_n(t) \tag{22}$$

Equation (22) describes the partitioning of the total cohort mass according to the number of events experienced. Having defined the mass measure $Q_n(t)$, we can now state

Theorem 4

$$\forall\, n \in \mathbf{N} : \quad Q_n(t) = Q_0(0) \cdot P_n(t). \tag{23}$$

Proof. The part of the total mass of the cohort that has experienced the event under consideration exactly n times during a time interval $[0, t)$ is proportional to the probability to experience the event n of times during that time interval, that is,

$$Q_n(t) = \kappa \cdot P_n(t). \tag{24}$$

Recalling (22) and summing over all n, we have

$$Q(t) = \sum_{n=0}^{\infty} Q_n(t) = \kappa \sum_{n=0}^{\infty} P_n(t) = \kappa\,. \tag{25}$$

We can break down the sum total of the mass of the cohort on the LHS into two components

$$Q_0(t) = \kappa P_0(t)\,, \tag{26a}$$

$$\sum_{n=1}^{\infty} Q_n(t) = \kappa \sum_{n=1}^{\infty} P_n(t) \,. \tag{26b}$$

Clearly, equations (26a) and (26b) must hold for any value of t. In order to establish the constant of proportionality κ; without loss of generality we set $t = a$, and pass to the limit using (12a) and (12b)

$$\lim_{a \to 0} Q_0(a) = \kappa \lim_{a \to 0} P_0(a) = \kappa \,, \tag{27a}$$

$$\sum_{n=1}^{\infty} \lim_{a \to 0} Q_n(a) = \kappa \sum_{n=1}^{\infty} \lim_{a \to 0} P_n(a) = 0 \,. \tag{27b}$$

Of course, $\forall\, n \in \mathbf{N}$ the limiting value of $Q_n(a)$ as $a \to 0$ equals $Q_n(0)$, the initial or boundary condition of the cohort mass for each value of n. Substituting (27a) in (24) therefore completes the proof.

Note that, since, by definition, $\forall\, n, t:\ Q_n(t) \geq 0$, it follows from (27b) that $\forall\, n \in \mathbf{N}^+:\ Q_n(0) = 0$. Hence the initial condition of the cohort mass is described completely by $Q_0(0)$.

Further, note from (25) and (27a) that

$$\forall\, t \in \mathbf{R} \backslash \mathbf{R}^-:\ Q(t) = \sum_{n=0}^{\infty} Q_n(t) = Q_0(0) \,, \tag{28}$$

that is, the total mass of the cohort remains constant over time. This is the law of the conservation of cohort mass.

Next, recall that non-negative integer function $K_n(t)$ denotes the number of individuals within the cohort who have experienced the event under consideration n times, $n \in \mathbf{N}$, on the time interval $[0, t)$. We can now give a precise formal definition of $K_n(t)$ by relating it to the cohort mass measure $Q_n(t)$. Specifically, the relationship between cohort mass Q and number of individuals K is defined by

$$\forall\, n, t:\quad K_n(t) = \lfloor Q_n(t) + \tfrac{1}{2} \rfloor \,, \tag{29}$$

where, in general, $\forall\, u \in \mathbf{R}:\ \lfloor u \rfloor$, the floor of u, is defined as the greatest integer less than or equal to u.

The distinction between the real-valued function cohort mass Q and the integer function number of cohort members K is merely a matter of mathematical

principle. While theoretical results will always be real valued, empirical results will always be integer valued. Therefore, expression (29) serves as a formal link between theory and empirical reality.

However, if in practice one tacitly agrees that values of Q will always be rounded to the nearest integer, then it is not necessary to make a notational distinction between Q and K. Henceforth, we shall adopt this convention.

Before proceeding, we note the following. The law of the conservation of cohort mass might at first sight be somewhat counter-intuitive. For, major analytical approaches in demography, such as life tables, cohort survival projection models and stable population analysis, all seem to suggest otherwise.

Life tables, however, traditionally largely ignore those members of the cohort who have experienced the event of dying and who subsequent to this event have taken alive status value not alive. But, of course, they remain cohort members, albeit with status value not alive.

Population projections come in many forms. They may, for example, be formulated in the continuous terms of some renewal equation or in the discrete terms of a cohort survival model such as the one based on a Leslie matrix. By definition, however, they all deal with populations which are, in principle, self-renewing. They deal not only with survival but also with the generation of new cohorts by existing cohorts.

Clearly, it is, therefore, important to distinguish between the concept of a *cohort* as defined above, on the one hand, and that of a *population*, on the other.

A population and a cohort are not normally identical. A population is usually composed of a sequence of cohorts of successive ages. Also, populations are not normally defined so as to include cohort members who have experienced events such as death or outmigration. Common definitions do include inmigrants, on the other hand.

In the population renewal process, no existing cohort changes in mass, but the population mass may well vary over time. Through the event of giving birth, an existing cohort merely creates new cohorts, not necessarily of identical mass. Population projections trace the mass composition of the sequence of existing and new cohorts as time and age progress. In the process, any deaths and outmigrants are removed from the population, and any inmigrants are added.

Stable population analysis can essentially be conceived as a long-term population projection process where lifetime hazard functions $\mu(t)$ are identical for all cohorts and remain unchanged. Under mild assumptions, such a projection process leads to a population mass composition which remains constant over time

in relative terms: a stable population. If the population mass composition remains constant over time in absolute terms as well, then the population is called stationary.

Next, we shall continue by exploring the interpretation of the hazard function $\mu(t)$ in some depth.

3.5 A PHYSICAL INTERPRETATION OF $\mu(t)$

Earlier we have referred to the hazard function $\mu(t)$ as an intensity and as an instantaneous rate, $\forall\, t \in \mathbf{R}\backslash\mathbf{R}^{-}$. However, we have so far not demonstrated that such descriptions of $\mu(t)$ are legitimate. While results such as (15) and (16) are of course valid, they might perhaps not be as intuitive or as easy to interpret as the terms instantaneous rate or intensity. Therefore we shall next demonstrate the legitimacy of these terms.

In order to be able to do so, it is convenient to shift our perspective from event counts in continuous time to the time interval between successive events. We then have

Theorem 5
$\forall\, t \in \mathbf{R}\backslash\mathbf{R}^{-}$: if $\mu(t)$ is constant (independent of t) on any time interval $[0, t)$, that is, if $\mu(t) = \mu$, then μ equals the inverse of the average time interval between successive events.

Proof. Let continuous stochastic variable τ, $\tau \in \mathbf{R}\backslash\mathbf{R}^{-}$ be the time interval between two successive events experienced by a member of the cohort. $\forall\, t \in \mathbf{R}\backslash\mathbf{R}^{-}$ the probability of a first event occurring at or after some instant t is equivalent to the probability of zero events occurring during $[0, t)$. Thus

$$P(\tau \geq t) = P_0(t), \tag{30}$$

so, by theorem 1 we have

$$P(\tau < t) = 1 - P(\tau \geq t) = 1 - P_0(t) = 1 - \exp(-\int_0^t \mu(u)du)\,. \tag{31}$$

The probability density function of random variable τ, $pdf_\tau(t)$, is

$$pdf_\tau(t) = \frac{d}{dt} P(\tau < t) = \mu(t) \cdot \exp(-\int_0^t \mu(u)du)\,, \tag{32}$$

and so, the average value $\bar{\tau}$ of τ is given by

$$\bar{\tau} = \int_0^\infty t \cdot pdf_\tau(t)\, dt = \int_0^\infty t \cdot \mu(t) \cdot \exp(-\int_0^t \mu(u)du)\, dt\,. \tag{33}$$

Clearly, to develop the RHS, it is necessary to specify $\mu(t)$. Now if we specify that $\mu(t) = \mu$ holds true, then (33) simplifies to

$$\bar{\tau} = \mu \int_0^{\infty} t \cdot e^{-\mu t}\, dt\,. \tag{34}$$

Evaluating (34), recalling that $\mu \in \mathbf{R}\backslash\mathbf{R}^{-}$, we obtain

$$\begin{aligned} \bar{\tau} &= \mu \lim_{\zeta\to\infty} \int_0^{\zeta} t \cdot e^{-\mu t}\, dt = \mu \lim_{\zeta\to\infty} -\frac{(1+\mu t)\, e^{-\mu t}}{\mu^2}\Bigg|_0^{\zeta} \\ &= \frac{1}{\mu}\{(-\lim_{\zeta\to\infty}(1+\mu\zeta)\, e^{-\mu\zeta}) - (-1)\} = \frac{1}{\mu}, \end{aligned} \tag{35}$$

proving theorem 5 for the time interval until the first event.

Next, we set $t = 0$ at the time point at which the first event ($n = 1$) occurs, and we observe the time interval between this and the second event ($n = 2$). Clearly, if $\mu(t) = \mu$ holds true, then expression (35) again results.

Since, if $\mu(t) = \mu$, we can carry such a translation of time through for all values of stochastic variable n, this completes the proof.

From theorem 5, it follows, of course, that the hazard is a *rate*, and not a probability.

Now consider $\mu(t)$ some interval $[t, t+a)$, $t \in \mathbf{R}\backslash\mathbf{R}^{-}$, $a \in \mathbf{R}^{+}$. If we let $a \to 0$, then, by the definition of $\mu(t)$, we have that $\mu(t) \to \mu$. Therefore, theorem 5 also proves that the expressions instantaneous rate and intensity at t are justified.

Theorem 5 is also important since it gives a direct physical meaning to the hazard function $\mu(t)$; it allows us to visualize and interpret $\mu(t)$ empirically. An empirical interpretation of the hazard function is, of course, fundamental, because it is this function which provides the link between individual *demographic events* on the one hand and *cohort behaviour* as completely expressed by theorem 4 on the other.

The condition that $\mu(t)$ be constant for such an empirical interpretation to be valid, is less restrictive than it may seem. We can simply approximate continuous function $\mu(t)$ on the interval $[0, t)$ of interest by a step function, and translate each step interval $[t_i, t_i + a)$, $0 \le t_i \le t - a$, over a distance of $-t_i$. Since we are free to choose the number of steps within any given interval $[0, t)$, the condition

of a piecewise constant hazard can always lead to an arbitrarily close approximation to $\mu(t)$ by letting the number of steps increase without bound.

Before we proceed, let us briefly recapitulate some important results so far. We have formulated *a single general theory* to describe the occurrence of demographic events, irrespective of whether we are, for example dealing with the event of giving birth, the event of migrating, the event of dying, or any other formally similar event.

Further, as we have seen, demographic analysis is essentially *cohort analysis*. A particular cohort under consideration is traced over time as it is exposed to the risk of experiencing demographic events. This risk is *governed exclusively by the hazard function*.

Finally, in our formulation, *time*, denoted by the variable t, is defined as a *continuous variable*: $\{t\} = \mathbf{R}\backslash\mathbf{R}^{-}$.

Of course, as time progresses, the cohort in question ages. It is common to use the continuous variable x to denote the *exact age* of the cohort, generally defined as $\{x\} = \mathbf{R}\backslash\mathbf{R}^{-}$. Clearly, therefore, variables t and x are interchangeable. If, for example, a cohort is aged x_1 at $t = 0$, then all that is necessary is a translation over a distance of $|x_1|$. So, $t = x - x_1$, and $x = t + x_1$.

Next, we shall explore how important more traditional approaches in demography tie in with the theory developed thus far.

3.6 FROM EVENT COUNTS TO EVENTS AND COMPETING EVENTS

Traditionally, the demographic paradigm centres on events rather than on event counts. The major exceptions are in the study of parity and in the study of repeat migration behaviour (the study of multiple moves by cohort members). Usually, also, the discipline adopts a less general perspective. We shall explore these issues next.

To build a link between our general theory of demographic behaviour and the traditional paradigm, we now broaden our perspective from event counts to events themselves. This can easily be achieved by extending the state space to include the appropriate status values. At the same time, however, it requires that we restrict our earlier state space by limiting the range of random variable n to the set $\{0, 1\}$.

For example, in the case of mortality analysis, we add the values of the alive status to the state space. Let us denote status value alive by λ and status value not alive by δ. The state space then becomes the collection $\{\{0, 1\}, \{\lambda, \delta\}\}$. We shall refer to the two subsets of such a state space as the event count (or event frequency) state space and the event state space, respectively. Note that both subsets are ordered.

In the case of fertility analysis, the additional status values might be not having given birth, say, ψ and having given birth, say, φ, with state space $\{\{0, 1\}, \{\psi, \varphi\}\}$. A more sophisticated and practically more useful fertility analysis recognizes parity status. Then the status values might be having given k births, $k \in \mathbf{N}$, denoted by φ_k, and having given k+1 births, φ_{k+1}. Now the state space is $\{\{0, 1\}, \{\varphi_k, \varphi_{k+1}\}\}$. The analysis is then carried out separately for all values of k.

As in the case of fertility analysis, in the analysis of migration there are also various options to define the status values. The simplest option (option 1) is to define i as the current place of residence and j as the next place of residence, $j \neq i$. (For the sake of convenience., we shall drop the adjective "usual" and assume that reference to any place of residence is always understood as a place of usual residence.) The state space is now $\{\{0, 1\}, \{i, j\}\}$. The analysis is carried out separately for all permutations of the admissible values of categorical variables i and j.

An alternative option (option 2) in the analysis of migration is quite different from the above approaches. We now define the event state space as the set of admissible places of residence. Clearly, in general, this is a categorical, and hence unordered, set. The difference with the approaches discussed above is twofold. We allow multiple events on any given time interval $[t, t+a)$, $\forall\, t \in \mathbf{R}\backslash\mathbf{R}^{-}$, $\forall\, a \in$

$\mathbf{R}^+$. In addition, we only specify the status values at the two time points t and $t+a$.

This approach leaves the sequence of migratory events for individual members of the cohort unspecified, provided only that this sequence is compatible with these two status values at t and $t+a$, respectively. Such sequences are properly called *endpoints-only-specified event sequences*. Consequently, except -- by postulate 2 -- in the case where one lets $a \to 0$, migratory events on a time interval *cannot even be counted at all*.

Now, the analysis is carried out simultaneously for all permutations of the admissible state space values. A well-known example of this approach in discrete rather than continuous time is Markov chain analysis. For the moment, when discussing migration we shall not refer to this option 2; we shall return to it later.

In all these cases, then, the event count state space $\{0, 1\}$ is matched by what is or what may be considered to be a categorical event state space. Remember here that the two subsets of the state space are defined as ordered. More formally, there is always a one-to-one correspondence between the first element of the first subset and the first element of the second subset of the state space; and similarly there is always a one-to-one correspondence between each of the second elements of the two state space subsets. An event experienced by a cohort member, that is, a value change of the event frequency status n from 0 to 1 is uniquely associated with a value change of the event status from λ to δ, from ψ to φ, and so on.

These correspondences allow for a simpler notation of the state space by omitting the first subset. However, as we shall see, the existence of these correspondences is a valuable notion in theory construction.

Further, to be explicit, the hazard function $\mu(t)$ should be properly specified so as to define the hazard in question: $\mu_{\lambda\delta}(t)$, $\mu_{\delta\lambda}(t)$, $\mu_{\psi\varphi}(t)$, $\mu_{\varphi\psi}(t)$, and so on.

As defined above, once a cohort member has taken fertility status values φ or φ_{k+1}, respectively, then for logical reasons it is no longer possible to take status values ψ or φ_k, respectively. So, formally, $\forall\, t \in \mathbf{R}\backslash\mathbf{R}^-$, hazard functions such as $\mu_{\varphi\psi}(t)$ and $\mu_{\varphi_{k+1}\varphi_k}(t)$ are identically zero.

State φ is called an absorbing state: once a cohort member has given birth, then this remains true forever. On the other hand, once, in the second case, value φ_{k+1} has been taken, then a value of φ_{k+2} remains possible, of course. However, this has to be considered separately if the event count space has been restricted to the set $\{0, 1\}$.

In the case of mortality, a cohort member cannot take status value λ once status value δ has been taken. Here, the reason is empirical rather than logical. Thus, here the value δ is an absorbing state. Theoretically, there are two approaches, however. One is to define δ as an absorbing state and to define $\forall\, t \in \mathbf{R}\backslash\mathbf{R}^{-}$: $\mu_{\delta\lambda}(t) = 0$. The alternative approach is to defer this issue to the measurement stage, likely to obtain a measured value for $\mu_{\delta\lambda}(t)$, at least within acceptable measurement error bounds, identically zero for all admissible values of t.

In the case of migration, there is no such a priori logical or empirical objection to a status value change from j to i after having experienced an earlier status value change from i to j, and, in general of course, neither should there be. But again, this has to be considered separately if the event count space has been restricted to the set $\{0, 1\}$.

We note that normally all analysis will always start from $t = 0$. This is without loss of generality since through a simple linear translation any non-zero time and corresponding age point can be translated to $t = 0$ and its corresponding exact age. Observe that this is a standard device in the application of life table construction where the perspective is shifted step by step from one exact age to the next, conditional by state at that first exact age.

In all these examples of demographic analysis, on some time interval $[0, t)$ a cohort member can experience either 0 or 1 events, $\forall\, t \in \mathbf{R}\backslash\mathbf{R}^{-}$. The probability of not having experienced an event on $[0, t)$ is, of course, given by theorem 1. Further, the set $\{0, 1\}$ is an exhaustive denumeration -- there are no other alternatives --, and its elements are non-overlapping (mutually exclusive). So, the probability of having experienced the event is given by theorem 2.

Finally, to complete the link with the traditional paradigm, it is necessary to *partition the cohort* $K(t)$ by event status value analogous to the way we encountered this in theorem 4 and its proof.

For example, in the case of mortality analysis we would have the partitioning $K_{\lambda}(t)$ and $K_{\delta}(t)$ such that $K_{\lambda}(t) + K_{\delta}(t) = K(t)$, $\forall\, t \in \mathbf{R}\backslash\mathbf{R}^{-}$. The occurrence of an event to a cohort member at time point t leads to $K_{\lambda}(t)$ being reduced by 1 and $K_{\delta}(t)$ being increased by 1. Informally, then, such an event results in this individual being transferred from cohort part K_{λ} to cohort part K_{δ}. Clearly, recalling the correspondences between the event count state space and the event state space, we have from theorem 4 the initial condition that $K_{\lambda}(0) = K(0)$, and $K_{\delta}(0) = 0$.

Consequently, if we have, in general, ordered state space $\{(0, 1), (\eta, \vartheta)\}$, then the life history of a cohort K is governed by

$$K_\eta(t) \;=\; K_\eta(0)\,\exp(-\int_0^t \mu_{\eta\vartheta}(u)du) \tag{36a}$$

$$K_\vartheta(t) \;=\; K_\eta(0)\,\{1-\exp(-\int_0^t \mu_{\eta\vartheta}(u)du)\} \;=\; K_\eta(0)-K_\eta(t) \tag{36b}$$

These equations are, for example, the basis of classical life table construction in the analysis of mortality. Here, $K_\eta(0)$ (that is, $K_\lambda(t)$) is called the radix. Its value is often set arbitrarily at 100,000. Equation (36a) then is applied step by step to subsequent exact ages conditional on survival to that exact age. Traditionally, the life history of the cohort part $K_\vartheta(t)$ (that is, $K_\delta(t)$) tends to receive scant attention in mortality analysis. However, by the law of the conservation of cohort mass we have, of course, that these individuals remain cohort members.

We note that in practical empirical applications of classical life table construction, often the hazard functions of a sequence of distinct observed cohorts are applied to a synthetic cohort. The period life table is a prime example of this procedure. Clearly, without strong additional assumptions such an approach has no theoretical validity.

While equations (36a) and (36b) are essentially simple in recognizing only a single hazard and a restricted event count state space, they are of considerable value in demographic measurement. This is easy to see. All that is required is the tracing of a cohort's mass (that is, size) in terms of either $K_\eta(t)$ or $K_\vartheta(t)$ over time within this simple analytical state space framework. The only unknown remaining then is the hazard function $\mu_{\eta\vartheta}(t)$, for which it is now easy to solve.

For example, if $K_\eta(t)$ is observed, then we have

$$\mu_{\eta\vartheta}(t) \;=\; \frac{d}{dt}(-\ln\frac{K_\eta(t)}{K_\eta(0)}) \;=\; \frac{d}{dt}(\ln\frac{K_\eta(0)}{K_\eta(t)}) \tag{37}$$

If we remember that the hazard function is the exclusive governor of a cohort's demographic behaviour, then equation (37) is a *measurement instrument which completely measures this behaviour*. It is worth noting that this is true irrespective of whether the behaviour in question concerns mortality, fertility, migration, or any other formally similar behaviour.

If $K_\vartheta(t)$ is observed, then, using $K_\eta(t) = K_\eta(0) - K_\vartheta(t)$, the same measurement instrument can be employed.

Thus far, we have centred on theory development focused on what is sometimes called a single decrement. That is to say, there is only one single force which determines cohort behaviour, identically for all cohort members, and this force leads to a one-way state change.

However, empirically, there are many fruitful lines of thinking where the force may be thought of as composite, leading to *multiple* decrement analysis; or where it is useful to *relax the one-way constraint*, leading to increment-decrement analysis. This can then be taken one step further yet by combining these two extensions, leading to what is now called multistate analysis.

Multiple decrement analysis merely extends the state space to $\{(0, 1), (\eta, \{\vartheta\})\}$, where $\{\vartheta\}$ is some (generally unordered) set of status values $(\vartheta_1, \vartheta_2, \vartheta_3, \ldots)$. Essentially, all that is required is the *partitioning of the hazard rate* $\mu_{\eta\vartheta}(t)$ into $\mu_{\eta\vartheta_1}(t)$, $\mu_{\eta\vartheta_2}(t)$, $\mu_{\eta\vartheta_3}(t)$, … such that

$$\mu_{\eta\vartheta}(t) = \mu_{\eta\vartheta_1}(t) + \mu_{\eta\vartheta_2}(t) + \mu_{\eta\vartheta_3}(t) + \cdots . \tag{38}$$

Each of these partial forces acts in competition, leading to the description as a competing events framework. A well-known example is the analysis of mortality by cause of death. Here, the set $\{\vartheta\}$ of status values is defined as the set death by cause 1, death by cause 2, and so on. Recalling postulate 2, a cohort member can die by one of these causes only, illustrating the competing nature of the partial forces.

While multiple decrement analysis leads to a measurement instrument similar to that of (37), we now have the sum of the partial forces on the LHS of the equation. Consequently, it is not possible to solve for the partial forces without *independent* additional information.

In the common formulation of multistate analysis incorporating mortality and migration, the general state space is the union of an ordered state space as encountered above in the case of the alive status, and an unordered set of all admissible places of residence as discussed above under option 2. In other contexts, any other formally similar status may, of course, be substituted for the migration status.

Following the approach adopted in the proof of theorem 1 and using the above concept of endpoints-only-specified event sequences in the case of migratory events as appropriate, it is elementary, albeit somewhat tedious, to prove for the multistate case that

$$\forall\, t \in \mathbf{R}\backslash\mathbf{R}^- : \qquad \frac{d}{dt}\boldsymbol{P}(t) = -\boldsymbol{P}(t)\boldsymbol{\mu}(t), \tag{39}$$

the equivalent of differential equation (11) above, with initial condition

$$\boldsymbol{P}(0) \;=\; \boldsymbol{I}\,, \tag{40}$$

where matrix $\boldsymbol{I}$ is a suitably dimensioned identity matrix, the multistate equivalent of (13).

The solution of (39) with (40) is formally given by

$$\forall\, t \in \mathbf{R}\backslash\mathbf{R}^{-}: \qquad \boldsymbol{P}(t) \;=\; \exp(-\int_{0}^{t} \boldsymbol{\mu}(u)du)\,, \tag{41}$$

the Taylor expansion of which can easily be seen to satisfy (39) with (40). Equation (41) is the multistate formulation of theorem 1. We leave the proof to the reader.

This result is general, in that it allows for various detailed formulations of the competing events and the associated state space.

In the case where the admissible events are mortality and migration, one common and simple formulation is that $\boldsymbol{P}(t)$ is a matrix structured as follows

$$\begin{bmatrix} P_{\lambda 1 \lambda 1}(t) & P_{\lambda 1 \lambda 2}(t) & \cdots \\ P_{\lambda 2 \lambda 1}(t) & P_{\lambda 2 \lambda 2}(t) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}, \tag{42}$$

and $\boldsymbol{\mu}(t)$ a matrix structured as

$$\begin{bmatrix} \{\sum_{j:j\neq 1} \mu_{1\delta}(t) + \mu_{1j}(t)\} & -\mu_{12}(t) & \cdots \\ -\mu_{21}(t) & \{\sum_{j:j\neq 2} \mu_{2\delta}(t) + \mu_{2j}(t)\} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}. \tag{43}$$

The admissible alive status values here are as before: alive and not alive. We again use the symbols λ and δ, respectively. The migratory status values are the categorical set (1, 2, 3, …), denoting the admissible places of residence. We again use indices i and j to denote elements of this set.

The elements of $\boldsymbol{P}(t)$ here are defined as follows: $P_{\lambda i \lambda j}(t)$ is the probability to be alive and in place of residence j at time point t, conditional on being alive and in place of residence i at time point 0.

And here, the elements of $\boldsymbol{\mu}(t)$ are defined as follows: $\mu_{i\delta}(t)$ is the instantaneous rate of death in place of residence i; and $\mu_{ij}(t)$ is the instantaneous rate of moving from place of residence i to place of residence j.

From the definition of $P_{\lambda i \lambda j}(t)$ it will be clear that (41) traces the endpoints-only-specified life history of individuals $\forall\, t \in \mathbf{R}\backslash\mathbf{R}^{-}$ specific by place of residence at $t = 0$. It is therefore appropriate to consider the sets of individuals with an identical place of residence at $t = 0$ as distinct cohorts.

Equation (41) forms the basis of multistate life table construction. Since it deals with distinct cohorts simultaneously, the life table has multiple radices.

As explained above, option 1 in the analysis of migration considers all permutations of admissible status value changes separately, and therefore it does not formally differ from the analysis of mortality and fertility described earlier. Option 2, however, considers these permutations simultaneously, as endpoints-only-specified event sequences. In addition, the multistate formulation of theory allows for competing events. This necessarily leads to the development of sets of simultaneous equations in the proof of theorem 1 for the multistate case. This also explains why the use of linear algebra has become an obvious analytical tool in the study of migration.

However, even in the simpler case of a multiple decrement framework, we already noted that that framework does not readily lead to simple measurement instruments. This applies all the more so to the multistate framework. However, in the multistate case, there is an additional reason. As noted, endpoints-only-specified event sequences are considered in the case of migratory events, rather than full event histories.

By implication, therefore, the multistate framework as set out above does not enable one to measure migratory event intensities. Only theorem 3 (or theorem 1 as a special case of theorem 3) directly allows the measurement of these hazard functions.

Consequently, when it comes to the development of measurement instruments, the elementary single decrement framework still wins the day.

We note that the focus on the multistate framework in its original formulation based on endpoints-only-specified migration event sequences, has contributed significantly to the confusion as to the best approach to the measurement of migration. Put in slightly different terms, this framework describes net transitions over time, that is, the balance of the effect of migratory events, rather than the migratory events, or the moves, themselves.

It comes as no surprise, therefore, that advocates of this framework are easily satisfied with data on place of residence at some fixed time point in the past, usually one or five years prior to the time point of measurement. For, such migration data also measure net transitions -- in discrete time -- rather than the events themselves in continuous time.

As explained, only event counts in continuous time allow for the measurement of event intensities, that is, the measurement of the hazard function. So, such transition data inhibit all attempts to recover the elementary function which completely governs a cohort's demographic behaviour.

However, equally important, as we shall see later, such migration data do not allow either for the adjustment of measured data for migration-specific incompleteness (underenumeration).

Finally, we note that the emphasis on multistate Lexis diagrams to fill in some of the gaps in demographic knowledge necessarily left by transition data, is at least in part a consequence of the focus on such net transition data in discrete time. Our approach, focusing on the events themselves as experienced by cohort members in continuous time, entirely removes the need for any such devices.

This completes our review of event counts, events and competing events. Next we shall briefly explore some of the limitations of the theory developed, and indicate some principal alternatives.

3.7 POSTULATES AND A PRIORI DEFINITIONS REVISITED

All theory developed thus far is based on our a priori definitions and our three postulates. It is useful to review to which extent this places limitations on the theory, and briefly to point to how such limitations might be relaxed, thus further enhancing the generality of the theory.

Postulates 1 and 2 are what is called weak. That is, their formulation is quite general, placing no fundamental constraints on the theory, neither in formal terms nor in terms of major implied restrictions on empirical applicability in the field of demography.

For example, postulate 2 would be problematic only in such extreme cases as where a cohort member dies en route in the removal van.

The same cannot be said of postulate 3, however. Postulate 3 is quite strong. As a consequence, it allows powerful and highly-transparent theoretical results in conjunction with postulates 1 and 2. At the same time, however, depending on the empirical context, it may well run counter to empirical evidence, thus restricting the applicability of theory.

For example, it is not uncommon for an individual that later migration behaviour is related to earlier migration behaviour. By postulate 3, the theory cannot recognize this. Let us give another example, this time from labour force analysis. If employers consider the duration in the state of unemployment an indicator of the quality of job seekers, then the intensity at which applicants join the labour force is dependent on their duration in that state of unemployment. Again, by postulate 3, the theory does not identify, and cannot reckon with, such a dependency.

One approach to incorporating some memory into the system is by postulating a continuous-time semi-Markov process. The hazard function can then be defined as depending on the length of the time interval since the last event, as well as, of course, on the status value change involved at that event. Thus, following this approach, *dependence* on earlier behaviour is *built in into formal theory*.

While dependence only on the time elapsed since the last event is an improvement, it is quite restrictive still. We note that carrying such and similar dependence further back, that is, conditioning on serial dependence, unavoidably leads to complex formalizations, however.

Let us return to our first example: individuals with a different migration history may exhibit different migration behaviour. For example, those cohort members who have experienced one or more migratory events in the past might be subject to a lower or higher propensity to move later in life.

For example Courgeau (1980) develops some extensions and alternatives of the theory so as to avoid the restrictive nature of our postulate 3. They are based on geometric and negative binomial formulations allowing the incorporation of dependency on the rank of the move and on the age at the previous move. In an illustrative worked example, he obtains the most satisfactory results using the negative binomial formulation.

A note of caution is in place, here, however. The weakening of postulate 3, making current migration behaviour dependent on, for example, earlier migration behaviour, does not only lead to more complex theoretical formulations. It necessarily leads to more complex measurement instruments, as well. For, the probability to experience an event at time point t then is formulated as no longer solely dependent on the hazard rate at that time point but also on earlier experience. This has major implications for data collection, too. It requires that more complete migration histories of individual cohort members be recorded.

While this may be possible when using population registration data or data from special migration surveys, it is unlikely that this will be given adequate priority in population censuses. In the majority of countries, population censuses are, and will likely continue to be, the principal or even the only source of migration data with national coverage. Population censuses face demands from users with widely differing interests, and the inclusion of additional questions on past migration behaviour is on the basis of competition with potential questions on other issues vying for inclusion.

Also, the recording of individual migration event histories is a topic which requires considerable skill and time on the part of the census field staff if accurate answers are to be obtained in terms of the individual(s) involved, the timing of the events and the associated previous places of residence. In actual operational practice, censuses are not particularly suitable for this.

This calls for alternative approaches. Now quite another perspective on this issue is the matter of *heterogeneity*.

In the development of theory thus far, we have assumed that individual cohort members are identical, and that each is subject to the specified hazard(s) in the same way. However, let us assume that the cohort comprises two distinct and independent subgroups, namely frequent and infrequent movers. Then in option 1 of section 3.6, we have that frequent movers rapidly move to the next place of

residence, leaving most of the infrequent movers behind. Thus, the cohort de-mixes.

Measurement of the hazard function will then record that, after high initial intensities, the intensity will drop sharply as time progresses.

While this is indeed the observed behaviour of the hazard function over time for the aggregate cohort, this behaviour of the hazard neither applies to the frequent movers nor to the infrequent movers. It is not even impossible in this scenario that each of these subcohorts independently experiences an increasing hazard as time progresses. In epistemology, such a phenomenon is well-known and there it is referred to as an ecological fallacy.

Formulated differently, in our case of frequent and infrequent movers with uncontrolled heterogeneity, measurement suggests a time-varying nature of the hazard which is at least in part a function of the heterogeneous nature of the cohort.

Heterogeneity may be broken down in two categories, *observed* and *unobserved* heterogeneity. In the case of observed heterogeneity, demography traditionally treats independent subgroups separately (stratification). Familiar examples are disaggregation of cohorts into subcohorts by sex and age. Thus, any heterogeneity associated with these status variables is eliminated.

Unobserved heterogeneity, however, does not in itself facilitate such treatment. If it is suspected that unobserved heterogeneity might play a role in co-explaining the observed hazard function, then the only approach is to measure additional relevant time-invariant and/or time-varying covariates (or explanatory variables), and to explore any dependencies by controlling for these covariates.

A special case of extreme unobserved heterogeneity, namely where members of one subgroup never experience the event of a move, has received considerable attention. It was originally developed in the field of labour force analysis by Blumen et al (1955) and further developed by Goodman (1961). It is called the mover-stayer model, and it is an example an approach to deal with time-invariant unobserved characteristics of workers affecting mobility.

The mover-stayer model is a mixed Markov model designed to capture this dependence. Stayers are assumed to have a zero propensity to make transitions during their observed life time, while it is assumed that the transition behaviour of movers can be described by a first order Markov chain model.

Differentiating between movers and stayers is an extreme position, however, which rather simplifies the heterogeneity which may be present in cohorts. Usually a somewhat more finely discriminating approach will be called for.

Generally in the case of migration it is therefore recommended to make any suspected relevant heterogeneity observable by measuring one or more covariates which allow one to differentiate between the distinct subgroups which together make up the cohort. This requires, of course, that one has an insight into the factors associated with, or into the causes underlying, the heterogeneity. In the study of migration, one principal source of heterogeneity among cohort members is usually the frequency of migrating: the mobile status. It has often been observed -- see for example Courgeau (1980) -- that the most mobile are a distinct subgroup amongst migrants, with specific migration behaviour. It is precisely this form of heterogeneity which also provides the rationale for, and which is taken to the extreme in, the mover-stayer model. The question then is, how to differentiate between this subgroup and other members within any one given cohort.

Such differentiation is most easily achieved by including an additional short and simple question in the census or survey which briefly *summarizes migration behaviour over a slightly longer period*, such as over the past five years. The common question on usual place of residence five years prior to the enumeration would already suffice. Then, if that place of residence differs from the previous place of residence, the individual concerned would be classified as mobile. In fact, only in such a supplementary discriminating role does this question on place of residence five years prior to the enumeration have any relative theoretical and methodological merit.

An alternative question might be on the number of migratory events experienced in the immediate past, say, five years, allowing for a slightly more subtle differentiation between mobile and not-so-mobile cohort members. This is a question included in the most recent Japanese population census. Such a fixed historical time interval could also be defined as back in time starting at the time point of the most recent migratory event, rather than at the time point of the enumeration. In that case the most recent event would not be included in the answer. However, clearly, the more probing the summary question, the less suitable it will be for a population census.

Obviously, any such question on summary migration behaviour need be asked of recent migrants only, say, those who arrived within the last two or three years at most. Other cohort members by definition do not qualify as mobile.

The heterogeneous cohort can then be disaggregated into two internally more homogeneous subcohorts on the basis of the value on this dichotomous mobile status, and the two subcohorts can now be analysed separately using the theory developed thus far. In practice, this approach is usually preferable to substituting one or more weaker postulates for postulate 3, except perhaps in the case of specialist migration surveys and complete population registration data accurately recording full migration event histories of cohort members.

Further refinements of this approach are easily conceivable. For example, the disaggregation of the cohort can be extended into three subcohorts by also treating the non-movers as separate, and into more subcohorts by treating all frequency values obtained in a summary question on the number of events experienced as separate. In such cases, the mobile status will correspondingly be defined as multivalued instead of as dichotomous.

Two interesting further possible amendments of the theory briefly deserve mentioning here. They relate to the a priori definitions.

The hazard function itself may be defined as a random variable varying independently of the history of the process. The theory then belongs to the domain of what are called doubly-stochastic processes. Clearly, it leads to more complex formulations.

Secondly, the context of application might suggest that we define the state space as continuous. This too will then lead to alternative specifications of theory. One avenue, for example, is the type of specification encountered in economics and econometrics in the field of the analysis of time series.

Finally, as set out to do, we have limited theory development essentially to description. However, the extension to explanation is, of course, straightforward, namely by introducing additional time-invariant and/or time-varying explanatory variables which are external to the process.

After this brief review of postulates and definitions, we shall next compare and contrast the approach taken thus far with traditional approaches in demography using a familiar example. This serves to highlight some major instances where both approaches concur and where they differ.

A well-established approach to formalizing cohort behaviour in demography is to start by measuring rates as stationary, that is, time-invariant, rates for individual age groups.

We shall take the case of mortality analysis as our example here, since it will be familiar to all demographers. However, mutatis mutandis, the discussion below similarly applies to the cases of fertility, migration, and other formally similar areas of study within the field of demography.

The preferred choice of information system for the measurement of mortality rates is a complete civil registration system so that events and exposure can best be determined. The rates are measured on the assumption of stationarity within each age group. In other words, within each age interval, the hazard is assumed to be uniform.

Next, probabilities of dying are constructed from these empirically observed rates. A well-known standard method is as follows. Let K_x be the mid-year population count for some one-year wide age group $[x, x+1)$ and let D_x be the number of observed deaths experienced by that population during the year. Then, assuming a linear distribution of deaths over the year, the start-of-year population can be approximated as $K_x + \frac{1}{2}D_x$. The probability of dying during the year conditional on survival to the beginning of the year, q_x, is then approximated as

$$q_x = \frac{D_x}{K_x + \frac{1}{2}D_x} = \frac{\frac{D_x}{K_x}}{1+\frac{1}{2}\frac{D_x}{K_x}} = \frac{m_x}{1+\frac{1}{2}m_x}, \qquad (44)$$

where m_x is the observed mortality rate for the age group. Under the same assumption of a linear distribution of deaths but for five-year wide age groups, the analogous result becomes

$${}_5q_x = \frac{5D_x}{K_x + \frac{1}{2}\cdot 5D_x} = \frac{\frac{5D_x}{K_x}}{1+\frac{5}{2}\frac{D_x}{K_x}} = \frac{5m_x}{1+\frac{5}{2}m_x}. \qquad (45)$$

Clearly, these results are problematic, not only because of the approximating nature of the assumption of linearity.

A second difficulty, namely, is that a stationary hazard leads to an exponential distribution of survivors, as can be seen from theorem 1. Thus, there exists a

contradiction between the assumption of a linear distribution of deaths on the one hand and the assumption of a stationary hazard on the other.

Yet another problem is that by mixing a period and a cohort approach, events and exposure are not properly reconciled. To improve the latter condition, an appeal to approximating interpolation using Lexis diagrams is common. Such interpolation is usually linear. However, when such interpolation is not exponential, then we clearly have another inconsistency.

However, even with such an improvement to reconcile a period and a cohort approach by interpolation, the result is still not perfect. This is easy to see if one realizes that the proper definition of the stationary probability in question for the cohort is given by theorem 2A. Let us here denote that probability for the case of mortality by $P_{\lambda\delta}(t)$ where λ represents the status value alive at the start of the analysis, and δ the status value not alive. We then have

$$P_{\lambda\delta}(t) = 1 - e^{-\mu t} \tag{46}$$

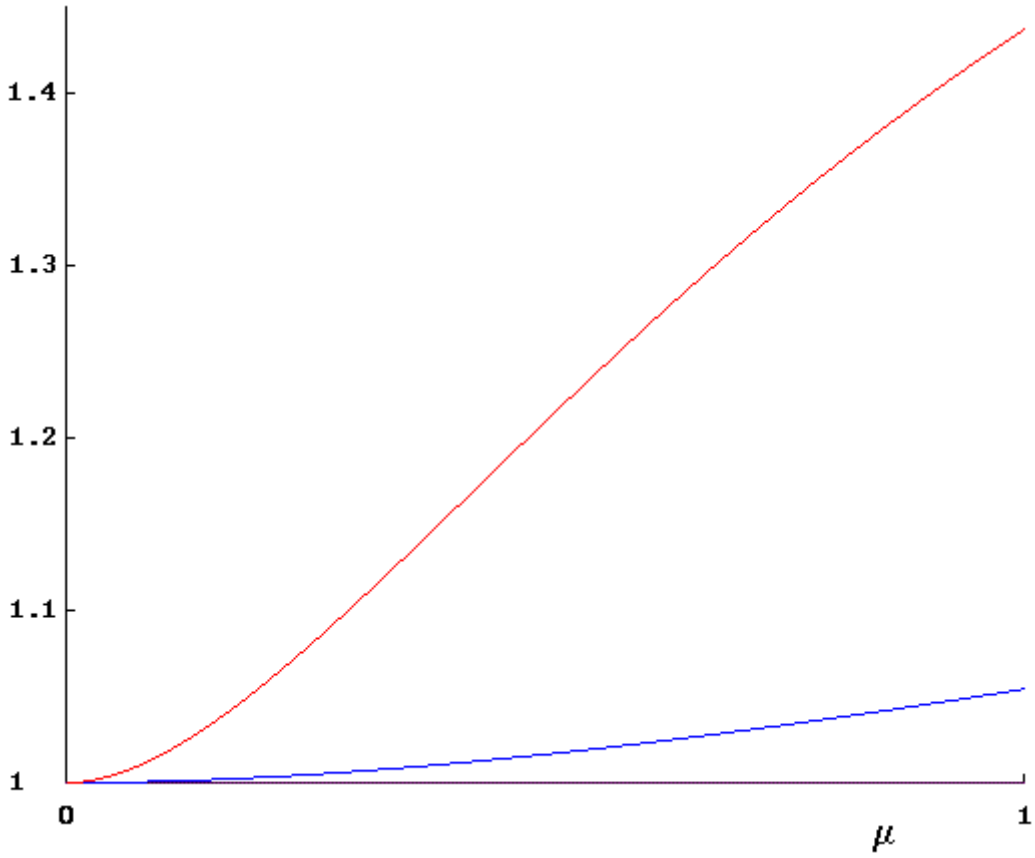
Comparing expressions (44) and (45) with this exponential function is made easier if we expand the latter using a Taylor expansion, allowing us to write

$$\begin{aligned} P_{\lambda\delta}(t) &= 1 - \{1 - \mu t + \mu^2 \frac{t^2}{2!} - \mu^3 \frac{t^3}{3!} + \mu^4 \frac{t^4}{4!} \cdots\} = \\ &= \mu t - \mu^2 \frac{t^2}{2!} + \mu^3 \frac{t^3}{3!} - \mu^4 \frac{t^4}{4!} \cdots \end{aligned} \tag{47}$$

with $t = 1$ and $t = 5$ respectively. Clearly, this is quite different from equations (44) and (45), respectively.

Figure 2 further illustrates this difference by showing the ratios $q_x / P_{\lambda\delta}(1)$ (the blue graph) and ${}_5q_x / P_{\lambda\delta}(5)$ (the red graph), setting $m_x = \mu$, as μ ranges from 0 to 1. If all were perfect, then, of course, both these ratios would be identically 1, $\forall\, \mu \in [0, 1)$.

Figure 2 Ratios of the Traditional Approach to Mortality Probability Approximation and the True Probabilities for One- and Five-Year Age Intervals (Blue: one-year age interval; Red: five-year age interval)



Clearly, (44) and (45) overestimate the true probabilities, and the discrepancy increases both with the value of the hazard rate and with the width of the age interval considered.

This example of the analysis of mortality demonstrates the significance of *theory-based measurement* as against the traditional *measurement-based theory*.

We note that (44) has been improved by various authors so as to obtain a better correspondence between empirical evidence and derived probabilities, particularly for ages where the hazard is high. Well-known early examples of such improvements are, of course, Reed and Merrell (1939) and Greville (1943).

However, in both cases, the attempt is to improve this correspondence by relaxing the assumption of the stationarity of the hazard function within age intervals.

Reed and Merrell base their approach on selected empirical data from the early 20th century, while by employing Gompertz' "law of mortality", Greville uses a more formal approach. Such advances are, therefore, not in formal theory construction but merely in improved specifications of the hazard function $\mu(t)$ with a view to obtaining a better model fit. The matter was further discussed by others, including, for example, Keyfitz (1966, 1968a, 1968b, 1970), Keyfitz and Flieger (1971) and Chiang (1968, 1972, 1984). We refer to these sources for further details.

With these observations we conclude our development of theory. Building on these results, we shall next explore how these developments lead to *operational approaches to measurement*.

As we shall see, elementary flaws in standard traditional approaches such as those highlighted above, can be avoided altogether. Given the theory developed thus far, the approach to measurement is an obvious one. And, in addition, it is one which in principle -- that is, in the case of good quality data -- does not require any *fundamental* concessions in terms of approximation or in terms of logical or empirical consistency.

Importantly, also, however, the approach to measurement allows us to deal with *incompleteness of the basic migration data* and to adjust for any such incompleteness in a theoretically justified and transparent manner.

4 THE OPERATIONAL MEASUREMENT AND CORRECTION OF MIGRATION DATA

4.1 INTRODUCTION

Having completed the development of theory, we now turn to measurement. We shall discuss operational approaches in a step-by-step fashion, presenting them in a manual-like form. While we shall focus on migration, the approach is general, *applying equally to the measurement of other formally similar demographic behaviour such as mortality, fertility, and so on.*

One element, however, is particular to migration. This is the estimation of, and adjustment for, any incompleteness which is specific to the process of migration.

Further, we shall use data on *internal migration* However, in terms of procedure, there is no difference between measuring internal and *international migration.*

By way of example we shall use data from the 1970 Population and Housing Census (PHC) of Thailand (NSO, 1972-1977). This is a standard census enumeration, closely following the United Nations' principles and recommendations for such censuses. It is generally considered a successful enumeration, not plagued by the procedural difficulties encountered, for example, in the 1980 census (Wanglee, 1982).

Further, the 1970 census is the first census in the kingdom in which data were collected enabling us to apply the theory developed earlier. Previously, only data on place of birth were recorded. Therefore our findings below can serve as a suitable benchmark for comparison with more recent censuses in the country, allowing the detection of trends and developments.

Finally, following the United Nations' principles and recommendations for the 1970 round of population censuses, available data were not *tabulated* with a degree of detail which allows for the measurement of migration hazard functions for cohorts *disaggregated by age and sex* and *for the regions (migration-defining areas) of interest.*

This required special database queries. Through our work with Thailand's National Statistical Office (NSO) at the time, we have such special tabulations for the regions of interest, although some limitations remain -- we return to this below. With the passing of time, it becomes increasingly unlikely that NSO will still be

able to produce any special tabulations from this census. Therefore, our data set provides valuable insights which in the future might not be available anymore.

We shall begin by presenting, organizing and discussing the data.

4.2 THE DATA: PRESENTATION, ORGANIZATION AND DISCUSSION

From the theory developed earlier, we know that for the measurement of the hazard function we require data which match the format of equation (36a):

$$K_\eta(t) \;=\; K_\eta(0)\,\exp(-\int_0^t \mu_{\eta\vartheta}(u)du)\,, \qquad (36a)$$

or, equivalently, equation (37):

$$\mu_{\eta\vartheta}(t) \;=\; \frac{d}{dt}(-\ln\frac{K_\eta(t)}{K_\eta(0)}) \;=\; \frac{d}{dt}(\ln\frac{K_\eta(0)}{K_\eta(t)})\,. \qquad (37)$$

This means that for any given cohort we must have the *life history of its members at least until the fist event*.

The simplest form in which one can obtain such information is if data on the <u>*duration of residence*</u> are available. Population registers recording residence are a prime source. However, particularly in the developing world, relatively few countries keep such registers.

The second best source is a full population census, provided, of course, that it includes the appropriate question. We note, however, that unfortunately many countries do not record duration of residence data in their population censuses. There are at least two major reasons for this.

Neither the United Nations principles and recommendations for population censuses (United Nations, 1997, and earlier versions) nor the United Nations recommendations on statistics of international migration (United Nations, 1998, and earlier versions) have a sufficiently well-argued basis in formal demographic theory on which a solid case could be made for one or more of the possible questions suggested for the measurement of migration.

The second reason is advocacy by proponents of the concept of demographic accounts (see section 2), who -- mistakenly as we have seen in section 3 -- prefer a census question on the place of residence some fixed numbers of years prior to the enumeration.

Of course, while the duration of residence specifies the *timing* of migratory events, it does not specify the *direction*. Usually in the study of migration, direction will be of interest, as well. In the case of a population census, then, the appropriate direction measurement instrument is a question on the <u>*place of previous residence*</u>. to supplement the question on the duration of residence.

The 1970 PHC of Thailand is the kingdom's first census to include a question on the duration of residence. And it also includes a question on the place of previous residence.

For our analysis, we have divided Thailand into two regions or migration-defining areas, namely the capital Bangkok and the rest of the kingdom. We shall investigate internal migration from the rest of the kingdom to Bangkok.

Given the nature of the data, this means that we must focus on the cohorts whose place of residence at the time of the enumeration was Bangkok. For migrants among the cohort members, the place of previous residence is thus the rest of the kingdom. Clearly, because of the experience of migratory events in the life history of cohort members, the cohorts observed together constitute the population of Bangkok only at the time of the enumeration.

This is an example of the common fact that *cohorts* and *populations* and their respective life histories cannot normally be identified except at a specific instant of time.

Since Bangkok comprises several distinct districts used to measure migration in the enumeration, some cohort members reported a place of previous residence which differed from their place of current residence, with both of these places however being within our definition of Bangkok.

From our point of view, these cohort members are, therefore, not migrants on the basis of the most recently experienced event. We have corrected the data for such intra-Bangkok migrants.

Such a correction on the basis of data on the most recent event only can, of course, not be considered totally adequate. Had we known the life histories of these cohort members further back in time, then some of them might well have proved to be migrants after all. However, given data only on the most recently experienced event, it is impossible to verify this. All we can say is that the number of migrants estimated as a result of our analysis below, will therefore be a lowest estimate; the true number may be higher.

Since migration is often sex specific, we eliminate any possible effects of heterogeneity due to sex on the measured hazard functions by considering both sexes separately. Here we shall report results for males only.

Similarly, any effects of heterogeneity due to age are eliminated to the maximum extent possible by analysing age-specific cohorts separately. Available data at

best allow for the disaggregation by age intervals [5, 10), [10, 15), …, [60, 65), [65, ∞) at the time of the enumeration. This gives us 13 distinct cohorts of which we wish to trace the history until the first event.

In the 1970 Thai PHC, durations of residence were measured in one-year wide intervals, rather than in exact durations. This is common, and, as we shall see, it will not cause any major methodological difficulties in terms of the measurement of the hazard function. Specifically, the intervals used were [0, 1), [1, 2), …, [4, 5), and [5, ∞).

We note, however, that the current principles and recommendations of the United Nations (1997) suggest another duration of residence classification for migrants, namely [0, 1), [1, 5), [5, 10), [10, ∞). This is *unnecessarily crude*, and for the measurement of migration hazard functions, a *classification in one-year intervals is much to be preferred*.

Further to our observations on the 1970 Thai PHC, above, it is useful to mention that, today still, the United Nations recommendations on population and housing censuses (United Nations, 1997) provide guidelines on census tabulations. Unfortunately, however, these do not include the important tabulation of duration of residence by place of current and previous residence both by age and sex.

While this may be understandable in the days of printed tables -- it could potentially become a rather large table if the regional and age resolutions (disaggregations) are fine --, such limitations are no longer applicable with current information systems technology.

One may reasonably assume that the revision of United Nations (1997) for the 2010 and subsequent rounds of censuses will recommend *standard database queries instead of standard tabulations*, supplemented perhaps by some very elementary printed tables.

For the analysis of internal and international migration, finely disaggregated data on duration of residence by place of current and previous residence both by age and sex are elementary. *It is therefore recommended that a database query producing such data as a matter of routine be provided in the coming revision of United Nations (1997).* A similar recommendation applies to the *next revision of United Nations (1998)* on international migration.

As noted earlier, in the case of our data from the 1970 Thai PHC, disaggregation by place of current and previous residence, residence duration, age and sex remained incomplete. Duration of residence data were available by place of residence, age and sex, but not by previous place of residence. Previous place of

residence data were available by place of residence, age and sex, but not by duration of residence.

Generally, if special tabulations are not an option, then the best estimate of the body of the table, given the marginals, can be obtained through straightforward iterative proportional fitting (IPF); see for example Bishop et al (1975). IPF allows such estimates to be further improved if information on the structure of the body of the table is available as well.

Our choice was limited to applying iterative proportional fitting given the data as described above. Clearly, of course, this is a limitation of the available data from the 1970 Thai PHC which does not in any way affect the migration hazard function measurement procedure.

After this preliminary discussion, we can now present the basic data. Note that all cohort mass data are in hundreds, unless indicated otherwise. Table 3 presents the distribution of the mass of the 13 cohorts by duration of residence.

Table 3 Distribution of the Completed Duration of Residence (Years), Bangkok Male Cohorts 1970 (×100)

Cohort	[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, 5)	[5, ∞)	Unkn	Total
[5, 10)	12.0	26.9	24.6	24.5	16.8	1,945.2	0.0	2,050.0
[10, 15)	16.7	34.8	29.2	24.8	17.6	1,889.9	0.0	2,013.0
[15, 20)	47.8	76.9	57.9	46.7	34.5	1,682.2	0.0	1,946.0
[20, 25)	51.6	118.7	83.3	53.2	31.0	1,213.2	0.0	1,551.0
[25, 30)	20.9	38.0	33.0	31.7	22.2	1,023.2	0.0	1,169.0
[30, 35)	14.5	25.6	22.1	21.9	14.1	983.8	1.0	1,083.0
[35, 40)	8.3	15.2	13.7	13.1	9.9	821.9	0.0	882.0
[40, 45)	4.8	9.2	8.2	7.9	5.8	647.1	1.0	684.0
[45, 50)	3.4	6.4	6.1	6.0	3.7	475.4	0.0	501.0
[50, 55)	3.2	5.6	5.5	4.2	3.5	401.0	0.0	423.0
[55, 60)	2.0	3.5	2.9	3.5	2.0	316.1	0.0	330.0
[60, 65)	0.9	2.7	2.9	2.0	1.5	225.7	0.0	235.7
[65, ∞)	1.8	4.1	4.4	4.1	2.6	346.2	0.0	363.3
Unknown	1.3	0.4	0.1	0.2	0.2	10.8	6.0	19.0
Total	189.2	368.0	293.8	243.8	165.6	11,981.6	8.0	13,250.0

Table 3 contains small numbers both of cohort members whose age is stated as unknown and of cohort members whose duration of residence is given as unknown. We distribute the unknowns proportionally. Again, the standard

procedure is iterative proportional fitting given both the marginals and the known data. The results are displayed in table 4.

Table 4 Distribution of the Completed Duration of Residence (Years), Bangkok Male Cohorts 1970, Adjusted for Unknowns (×100)

Cohort	[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, 5)	[5, ∞)	Total
[5, 10)	12.1	26.9	24.6	24.5	16.9	1,947.9	2,052.8
[10, 15)	16.8	34.8	29.2	24.8	17.7	1,892.5	2,015.8
[15, 20)	48.2	77.0	57.9	46.7	34.6	1,684.5	1,948.9
[20, 25)	52.0	118.9	83.3	53.3	31.0	1,214.8	1,553.4
[25, 30)	21.0	38.1	33.0	31.8	22.2	1,024.6	1,170.7
[30, 35)	14.6	25.7	22.1	21.9	14.2	986.1	1,084.5
[35, 40)	8.3	15.2	13.7	13.1	9.9	823.0	883.2
[40, 45)	4.8	9.2	8.2	7.9	5.9	648.9	685.0
[45, 50)	3.5	6.4	6.1	6.0	3.7	476.0	501.7
[50, 55)	3.2	5.6	5.5	4.2	3.5	401.6	423.6
[55, 60)	2.0	3.5	2.9	3.5	2.0	316.5	330.5
[60, 65)	0.9	2.7	2.9	2.0	1.5	226.0	236.0
[65, ∞)	1.8	4.1	4.4	4.1	2.6	346.7	363.8
Total	189.3	368.2	293.9	243.9	165.7	11,988.9	13,250.0

Next, we have to consider the fact that the data as measured are an expression of *two competing risks*, namely the risk of experiencing a change in the alive status value and the risk of experiencing a change in the migration status value. Put in other words, the migration histories of those who have died prior to the enumeration have not been recorded. There are now two alternative approaches.

The first is to proceed without further considering the composite nature of the hazard. The analysis can then proceed without any further processing of the basic data, and the cohort masses can be adjusted for underenumeration as described below.

However, by equation (38), the hazard function $\mu_{\eta\vartheta}(t)$ in equations (36a) and (37) cannot then be interpreted as representing purely the risk of experiencing migratory events. For each individual cohort $\mu_{\eta\vartheta}(t)$ then is a composite measure representing the unpartitioned combined force of mortality and migration. This is clearly undesirable since it seriously limits the value of the analysis The preferred approach is therefore first to eliminate the competition.

As discussed before, it is impossible to separate competing events without *independent additional information*. Within the framework of the analysis based on the 1970 PHC, period life tables were constructed using well-established

indirect demographic estimation procedures (NSO, 1972-1977). We shall use such a life table for Bangkok males. Table 5 presents the life table column giving the number of years lived per person born, traditionally denoted by ${}_5L_x$.

Table 5 Life Table Bangkok Males, 1970: ${}_5L_x$

Age Group	${}_5L_x$
[0, 5)	4.61139
[5, 10)	4.47523
[10, 15)	4.43490
[15, 20)	4.38863
[20, 25)	4.31827
[25, 30)	4.23557
[30, 35)	4.14934
[35, 40)	4.05278
[40, 45)	3.93465
[45, 50)	3.78040
[50, 55)	3.57378
[55, 60)	3.29751
[60, 65)	2.92987
[65, ∞)	6.26908
Total	58.45137

Given the observed cohort masses at the time point of the enumeration, we apply routine backsurvival of all duration of residence classes, properly considering the average length of exposure to the risk of dying for each such class. The latter is achieved by appropriate interpolation of the 5-year inverse survival ratios derived from the life table.

We note that while this procedure is the best possible given the information available, it remains approximate. Strictly one would require cohort life tables for each cohort as of 1970. However, these will not usually be available.

Second, when applying a period life table, one should consider any historical trends in mortality in the backsurvival procedure. In the case of proper cohort life tables, any such trends are of course already accounted for in the life tables, since they trace the cohorts' actual life histories.

Third, while a period life table may be given as of 1970, in reality this time reference may not be accurate. Many indirect estimation methods for life table construction use retrospective data, so that in reality the resulting life table applies to an earlier time point than is indicated by the data collection time point.

Further, as we saw, when using 5-year period life tables, one has to interpolate so as to obtain inverse survival ratios correctly representing the exposure of the single-year duration of residence categories. Any such interpolation is, of course, also approximate.

Finally, the implicit assumption is that there is no heterogeneity between those cohort members who first experienced an event of migration status value change and those who first experienced an event of alive status value change. In other words, suppose that those who have actually died, had not died instead. Then the implicit assumption is that they would have shown the same behaviour as those who actually stayed alive.

Table 6 shows the data after elimination of the competing risk of leaving the alive status for all cohorts. Note that the procedure of applying backsurvival in fact reverses the order of time -- we shall return to this issue below. Therefore, the cohorts at the enumeration time point have not yet been subject to the force of mortality, so that the cohort mass at this time point remains unaffected. Only the mass values at earlier time points shows the effect of the elimination of the force of mortality; and this effect is the stronger the further back in time we go. This can be seen by comparing the data in table 6 with those of table 5.

Table 6 Distribution of the Completed Duration of Residence (Years), Bangkok Male Cohorts 1970, Adjusted for Unknowns and After Elimination of the Risk of Dying (×100)

Cohort	[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, 5)	[5, ∞)	Total
[5, 10)	12.1	27.2	25.0	25.1	17.3	1,946.2	2,052.8
[10, 15)	16.8	34.9	29.4	25.0	17.8	1,891.9	2,015.8
[15, 20)	48.2	77.3	58.2	47.1	34.9	1,683.2	1,948.9
[20, 25)	52.1	119.5	84.0	53.9	31.5	1,212.4	1,553.4
[25, 30)	21.1	38.3	33.4	32.2	22.6	1,023.2	1,170.7
[30, 35)	14.6	25.8	22.3	22.3	14.4	985.0	1,084.5
[35, 40)	8.3	15.3	13.8	13.3	10.1	822.3	883.2
[40, 45)	4.8	9.3	8.4	8.1	6.0	648.4	685.0
[45, 50)	3.5	6.5	6.2	6.2	3.9	475.5	501.7
[50, 55)	3.3	5.7	5.7	4.3	3.7	400.9	423.6
[55, 60)	2.0	3.6	3.0	3.7	2.2	315.9	330.5
[60, 65)	0.9	2.8	3.1	2.2	1.7	225.3	236.0
[65, ∞)	1.9	4.7	5.5	5.5	3.7	342.6	363.8
Total	189.7	370.9	297.9	248.8	169.9	11,972.9	13,250.0

Table 6 is interesting, since it constitutes the best approximation of the data when they would have been derived from a population registration system registering

persons and demographic events. For, in such a system the events of dying and of migrating are normally recorded separately for each person, so that the migration history of those who have not survived until the observation time point is preserved.

Next we cumulate the data. The results are shown in table 7.

Table 7 Cumulative Distribution of the Completed Duration of Residence (At Least X Years), Bangkok Male Cohorts 1970, Adjusted for Unknowns and After Elimination of the Risk of Dying (×100)

Cohort	0	1	2	3	4	5
[5, 10)	2,052.8	2,040.7	2,013.5	1,988.6	1,963.5	1,946.2
[10, 15)	2,015.8	1,999.0	1,964.1	1,934.7	1,909.7	1,891.9
[15, 20)	1,948.9	1,900.7	1,823.4	1,765.2	1,718.1	1,683.2
[20, 25)	1,553.4	1,501.3	1,381.8	1,297.8	1,243.9	1,212.4
[25, 30)	1,170.7	1,149.6	1,111.3	1,078.0	1,045.8	1,023.2
[30, 35)	1,084.5	1,069.9	1,044.1	1,021.7	999.5	985.0
[35, 40)	883.2	874.9	859.6	845.7	832.4	822.3
[40, 45)	685.0	680.1	670.8	662.5	654.4	648.4
[45, 50)	501.7	498.2	491.7	485.5	479.4	475.5
[50, 55)	423.6	420.3	414.7	409.0	404.6	400.9
[55, 60)	330.5	328.4	324.8	321.9	318.1	315.9
[60, 65)	236.0	235.1	232.3	229.2	227.0	225.3
[65, ∞)	363.8	361.9	357.2	351.7	346.2	342.6
Total	13,250.0	13,060.3	12,689.4	12,391.5	12,142.7	11,972.9

Table 7 shows how each of the cohorts experiences *increments as a consequence of inmigration* into Bangkok as time progresses from 1965 and 1970.

However, this is not quite the process described by equation (36a). Equation (36a) formulates the diminishing mass of cohort part $K_\eta(t)$ as a consequence of the experiencing of a first migratory event as time progresses. In other words, this equation describes the effect of *decrements due to outmigration*. So, $\forall\, t \in \mathbf{R}\backslash\mathbf{R}^-$, $\mu_{\eta\vartheta}(t)$ in equation (36a) in this case represents the instantaneous outmigration rate from region η to region ϑ.

It is straightforward, however, to structure our data so that they match our theoretical framework. This we achieve by the simple device of the *reversal of the order of time*. All we need to do is to transform 1970 to 0 (or t_0), 1969 to 1 (or t_1), …, 1965 to 5 (or t_5). This reverses the process of inmigration from t_0 onwards. Now, that is, going back in time, the 1970 Bangkok cohort parts

experience decrements due to the undoing of the inmigration as time progresses. It is as if we play backwards the film-recording the cohorts' life histories. We see cohort members resident in 1970 in Bangkok leaving the city as time progresses to 1965.

This is exactly the same approach as the one which we applied to mortality in the cohort backsurvival procedure, above.

Thus we have achieved a data set which fully matches the theoretical framework developed earlier, and this completes our discussion of the data and our database establishment procedure. It is now an appropriate point to turn to the issue of underenumeration.

4.3 UNDERENUMERATION AND ADJUSTMENT FOR UNDERENUMERATION

Most data sources, be it population registers, censuses or surveys, will suffer from defects resulting in errors and incompleteness, both in the enumerated cohort mass and in the recorded demographic events. There are various approaches to avoid, detect and correct such errors. They all fall under the general heading of quality assurance. Adherence to methodologically sound and a priori established procedures is, of course, a prerequisite. However, once the data collection has been finalized, errors may still be present.

The basic options remaining then include internal consistency checks and external consistency checks. Internal consistency checks might for example consist of attempts to reconcile various sets and subsets of the data.

External consistency checks might for instance consist of attempts to reconcile other sources, such as earlier, contemporaneous, or more recent sources, with the source under investigation. If there exists independence between the sources used, then ceteris paribus the confidence in the results will be enhanced. In addition, theoretical consistency checks are a highly valuable class of independent external checks. Here, the question is, do the data match validated theory.

In the case of population censuses, a specially constructed external source is the post-enumeration survey (PES). As a component of the 1970 PHC, a PES was conducted in Thailand, producing useful insights into the completeness of the census enumeration. We shall return to this information later.

When it comes to enumeration completeness, migrants are a special group of concern: they are disproportionately prone to underenumeration. There are a number of causes which underlie this, and it is useful briefly to review the principal causes in a general context.

First, there are organizational and administrative causes. For example, census mapping and household listing procedures necessarily take place some time before the date of the actual enumeration. Therefore, the resulting maps are liable to exclude recently constructed dwellings -- particularly informal ones -- accommodating new arrivals; and the resulting lists are similarly liable to exclude recent arrivals.

The second group of causes is socio-cultural. Newly-arrived migrants may not initially consider their stay as permanent and/or may regard themselves as foreign, not truly belonging to their new place of residence. Consequently, they may well not yet regard their new place of residence as their real place of residence. Then, when asked for their place of residence, they are more likely to give a family or

parental address in their place of previous residence as their perceived true place of residence. Clearly, once settled this tendency will diminish.

Another often related group of causes is economical. If new arrivals are still looking for opportunities to become economically active, then, similarly, their sense of belonging to their new place of residence will be relatively weak. Once successfully engaged in making a living, this sense of belonging will grow stronger. The tendency to misspecify the de iure place of residence will lessen.

Fourth, there are causes of a socio-organizational character. For example, as we saw in section 1, in the developing world squatter settlements and slums frequently house relatively large proportions of recent arrivals in towns and cities from the rural areas. For social and organizational reasons, the quality of enumeration in such settlements is usually poorer than elsewhere in urban areas.

Fifth, legal causes may play a very significant role. Fear for restrictions on internal and/or international migration may easily lead to the self-perception of being an illegal migrant, whether this is justified de iure or not. Clearly, the result will be a feeling of apprehension when confronted, for whatever reason, with persons perceived to belong to the authorities. Also, if one is economically active without a proper legal employment status, then this will similarly induce such apprehension. As a consequence, any contact with persons such as census takers is likely to be avoided. Again, once properly settled in the new place of residence, maybe with a legal employment status, the grounds for such fears will ease.

Before we proceed, we must mention yet another, sixth, factor which may play a role in producing defective data on migrants. Namely, a special problem arises in census enumerations where a person is defined as being a migrant only if he or she has been usually resident in the current place of residence for a minimum period, such as three or six months. The consequence of such an arbitrary de iure definition of the place of usual residence is that the most recent migrants -- all those with a duration of residence value on the interval [0, 0.25) or [0, 0.5), respectively -- are classified as residents in their respective places of origin.

Thus, cohorts in the places of origin are artificially inflated by de facto non-members who actually belong to cohorts in the places of destination. Cohorts in the places of destination are deprived of these members. Gardiner and Oey-Gardiner (1990), for example, note that such an arbitrary de iure reclassification was applied in the Indonesian census which they studied. There is no obvious theoretical, methodological or empirical rationale for this practice.

The application of such a de iure definition in itself does not lead to any underenumeration, but to a systematic misclassification of recent migrants as non-migrants and to their allocation to cohorts of incorrect migration-defining areas. Any separate specific ex-post correction for such misclassification will usually be an approximate estimate at best, depending on the information available.

However, an approximating ex-post correction is entirely unnecessary if the census bureau involved is explicit in the reallocations made pursuant to the de iure definition adopted. Better still, of course, is avoiding this problem altogether by employing a de facto definition based on the actual true place of usual residence at the time of the enumeration.

Thus, in respect of this sixth factor we are dealing with a problem which can easily be prevented or rectified, and which therefore merits no further special methodological attention. In what follows, we shall omit any further explicit reference to this issue in the interest of methodological clarity.

All these factors cause underenumeration which is specific to migration. Equally, all these causes are of a temporary nature: they are all strongly related to *recentness of arrival*. The more recent the arrival, the stronger will be the effect. Over time, once settled in the new place of residence, a migrant will become more and more similar to persons who have lived much longer in the new place of residence, at least from the point of view of *statistical observability*.

In other words, as a rule the disproportional underenumeration of migrants relative to the non-migrant element of a cohort is directly related to the recentness of arrival.

When, for example, Bell (2005) observes that the most mobile groups are those most likely to be overlooked in enumerations, he merely restates this fact. On any measurement, the most mobile persons will fall in the category of recent arrivals.

This observation leads to an important conclusion. Effectively, here, we have *partitioned underenumeration* into two classes, namely, the *migration-related* (that is, the recentness of arrival related) underenumeration, and the general underenumeration *from all other causes* that applies equally to all cohort members.

Recognizing this form of unobserved heterogeneity due to differential tendencies to be underenumerated, forms the basis of any adjustment of migration data for incompleteness.

For best cohort mass estimates, any underenumeration rate estimated for a cohort therefore first requires partitioning, rather than applying this rate to all cohort members equally, migrants or non-migrants.

The question then reduces to how this unobserved heterogeneity can be brought to the surface. So far, we have two relevant concepts, namely that of recentness of arrival and that of the partitioning of underenumeration. Unobserved heterogeneity can only be captured using additional information. We prefer an

independent external consistency check. And, in line with the approach taken in this paper, we clearly prefer this to be in the form of theoretical consistency.

Such a check can suitably be based on a third concept, namely that of the *expected distribution of the cohort by recentness of the event of migrating*. This expected distribution is, of course, readily available in the form of equation (36a).

There should be a good agreement between the empirically observed duration of residence data and the expected distribution given by equation (36a), except for short durations of residence. The latter discrepancy is the result of the migration-specific part of the overall underenumeration.

Since we are dealing with inmigration into Bangkok, it would appear appropriate here to use the specific index B for Bangkok, instead of the general index η, and R for the rest of the kingdom, instead of ϑ, in equation (36a).

The method to adjust migration data for underenumeration now proceeds in two steps. In the first step, *equation (36a) is estimated* by *fitting it to the observed data* using an appropriate formal specification of $\mu_{RB}(t)$ and using a suitable parameter estimation procedure. Below we shall return to the specification of $\mu_{RB}(t)$ and to parameter estimation procedures. This estimation procedure provides estimates of the parameters of the hazard function $\mu_{RB}(t)$, as well as an initial estimate of the true value of the cohort mass at $t = 0$, say $K_B(0)_{\text{init}}$.

However, in this parameter estimation procedure, *observed data for the shortest durations of residence are excluded*, since it is these and only these data which suffer from the migration specific component of underenumeration. Clearly it would be erroneous to use such defective data for the measurement of the hazard function.

Given the values obtained for $K_B(0)_{\text{init}}$ and for the parameters of $\mu_{RB}(t)$, initial estimates of $K_B(t)$ for any further values of t which may have been left out of the parameter estimation procedure on account of their recentness, can now also be obtained by substituting the appropriate values of t in equation (36a).

The difference between the initial estimate of the cohort mass at $t = 0$, say $K_B(0)_{\text{init}}$, and the observed value, $K_B(0)_{\text{obs}}$, say, then is the *estimator of the migration-specific underenumeration*. So, denoting the latter estimator by θ_m, we have

$$\theta_m = K_B(0)_{\text{init}} - K_B(0)_{\text{obs}} \tag{48}$$

and

$$\theta = \theta_m + \theta_g \tag{49}$$

where θ_g denotes the estimator of the general underenumeration from *all other causes*, and θ represents the estimator of the total underenumeration for the cohort. Such an estimate of θ has to be provided exogenously, for example from a PES.

We note that here all values of the thetas are expressed in *absolute cohort mass* terms, that is, in absolute numbers of cohort members. It is also common to express underenumeration either in terms of an *underenumeration rate* or in terms of an *adjustment multiplier*.

An underenumeration rate is a fraction whose numerator is the estimated deficit and whose denominator is the estimated true cohort mass. An adjustment multiplier is a factor which, when applied to the observed cohort mass, produces the estimated true cohort mass.

It is easily verified that the relationship between an underenumeration rate $\hat{u}$ and an adjustment multiplier $\hat{a}$ is given by

$$\hat{a} = \frac{1}{1-\hat{u}}, \tag{50a}$$

or, alternatively by

$$\hat{u} = 1-\frac{1}{\hat{a}}. \tag{50b}$$

In reports, one often encounters underenumeration rates, frequently multiplied by 100. However, in analysis the use of adjustment multipliers tends to be more convenient.

Next, in step two of the adjustment procedure, the difference $\theta - \theta_m$ is evaluated. In other words, the estimated part of the underenumeration specifically due to migration is removed from the total estimated underenumeration as assessed for the cohort. By equation (49), the result is an estimate of the remaining general underenumeration from all other causes θ_g.

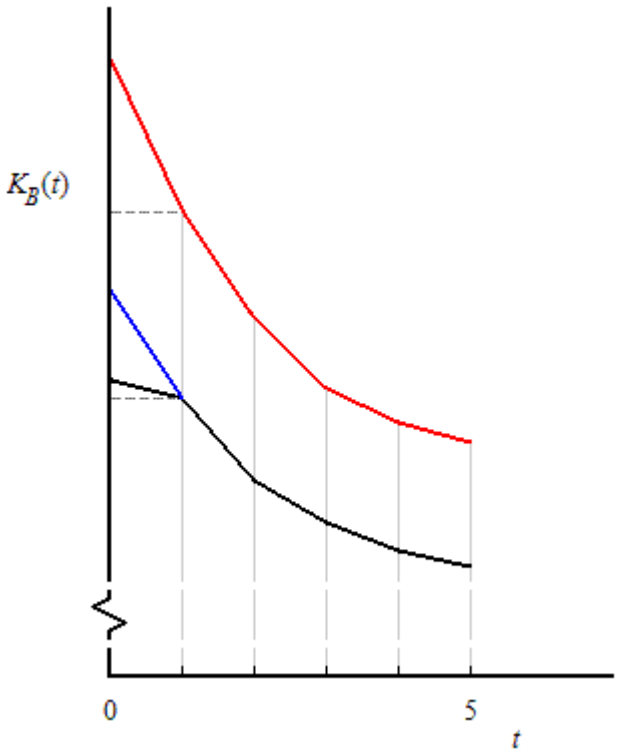
This number θ_g is then added to the entire cohort at time point $t = 0$, irrespective of the duration of residence of cohort members. Each duration of residence category receives its proportional share of θ_g. Thus, the cohort is adjusted for the remaining underenumeration which is due to *general causes other than the recentness of arrival*. This completes the adjustment procedure.

Aspects of the adjustment procedure can usefully be illustrated graphically. Figure 3 schematically shows data for a cohort from table 7 (graph 1) and table 6 (graph 2), respectively, in black. The blue elements represent the effect of step one of the adjustment procedure, and the red elements represent the effect of step two.

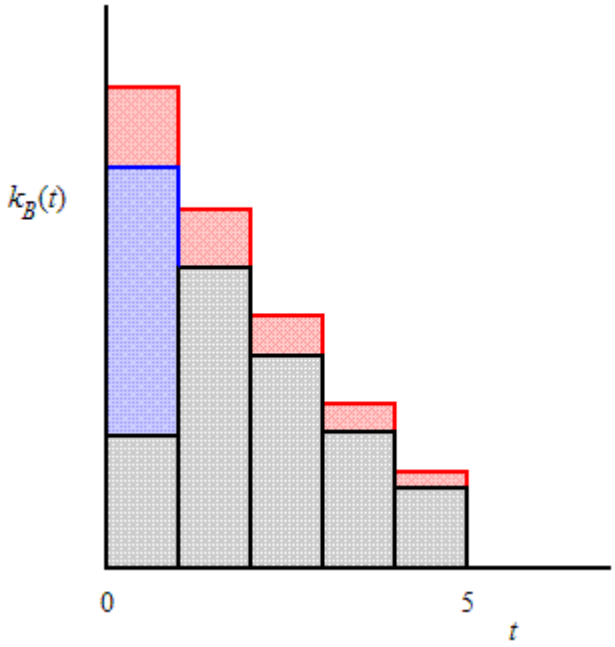
Figure 3 The Method of Adjusting Migration Data For Underenumeration or Incomplete Registration

(*Black: enumerated data adjusted for unknowns and competing risks; Blue: after adjustment step one; Red: after adjustment step two*)

Graph 1 Cumulative Distribution of the Completed Duration of Residence (At Least X Years) in the Current Place of Residence



Graph 2 Distribution of the Completed Duration of Residence (Years) in the Current Place of Residence



We make three observations. First, as noted, the availability of a value for θ, the estimated total underenumeration of the cohort must be provided separately from the adjustment method just described as an exogenous estimate. However, the availability of such a value is not a necessity. If such a value is not available, then it is merely necessary to assume that

$$\theta = \theta_m , \tag{51}$$

so that $\theta_g = 0$. In other words, one only has to assume that the observed data suffer exclusively from migration-specific underenumeration. Step two of the adjustment procedure then reduces to an empty formality.

Second, the relevance and timewise endurance of the causes of migration-specific, that is, duration of residence related, underenumeration will, of course, vary from empirical context to context. It is therefore not possible to make any general statements about the length of the time interval $[0, t)$ on which observed data points should be considered unreliable, and therefore be left out of the hazard function estimation procedure.

Such judgement has to be based first and foremost on external expertise, that is, on a deep insight into and understanding of relevant local conditions on the ground which may affect migration-specific underenumeration.

Second, systematically comparing and contrasting the goodness of fit for a well-selected sequence of values of t can provide additional valuable insights.

Finally, inspection of the data such as those displayed in graph 1 of figure 3, may help to obtain an indication of the point on the graph where the rate of change over time of the curve (that is, the derivative of its formal representation) shows a pronounced discontinuity. However, unless $\mu_{RB}(t)$ is time-invariant, care should always be exercised, as the behaviour of the graph may well also represent true variation over time in the hazard function.

In conclusion, we note that, in the procedure just described, the measurement of the hazard function, in this case $\mu_{RB}(t)$, and the adjustment of enumerated data for incompleteness go hand in hand. However, while the hazard function can be measured without adjusting the enumerated data for incompleteness, the reverse is not true.

If the data are subject to migration-specific incompleteness, then all that is necessary for the measurement of the hazard function is the removal from the analysis of affected data points, that is, of data points representing short durations of residence where migration-specific underenumeration plays a role.

On the other hand, from equation (48) we see that estimates of the parameters of the specified hazard function in equation (36a) are required for the evaluation of θ_m.

This completes our discussion of underenumeration -- or, in the case of a population register, of incomplete registration -- and of the adjustment procedure to correct for such data defects. Before we can now turn to the application of the adjustment procedure to our data, we first have to explore procedures which may be used to estimate the hazard function.

4.4 SPECIFICATION AND ESTIMATION OF THE HAZARD FUNCTION

It is clear from our discussion of Rogers and Castro (1981) that, generally, the search for a single hazard function which characterizes migration behaviour over the complete realized life time of a cohort is an elusive affair. The standard alternative is a *piecewise* approach. That is, we estimate the hazard function over an overlapping or an adjoining sequence of intervals, separately for each such interval. In principle, the sequence of intervals is $[t_0, t_0+a)$, $[t_0+\mathring{a}, t_0+\mathring{a}+a)$, $[t_0+2\mathring{a}, t_0+2\mathring{a}+a)$, …, $\mathring{a} \in \mathbf{R}^+ : \mathring{a} \leq a$. Variable $\mathring{a}$ is called the offset.

If $\mathring{a} = a$, then the intervals are adjoining instead of overlapping. In that case it may be appropriate to set conditions on the first order and, if desired, the second order derivative of the piecewise hazard functions at the juncture of such adjoining intervals. This then ensures a smooth transition of the piecewise hazard functions from one interval to the next.

Which interval width a and which offset $\mathring{a}$ to choose will in practice be determined to a large extent by the length of the observed (part of the) cohort life history and the timewise precision (resolution or fineness) with which observations have been recorded. Another practical consideration is the number of parameters of the hazard function to be estimated, as this sets a minimum on the number of data points required for each interval. Finally, as discussed above, for the first, or as appropriate the first few, intervals it must be considered that data points with a high degree of recentness are to be left out of the hazard function estimation procedure if migration-specific underenumeration plays any role.

The present data set has rather severe limitations, both in terms of the length of the cohort life history which has been observed and in terms of the timewise resolution of the data. An observation length of no more than five years of cohort life is available; and data are recorded at annual time points. This is quite common for reported census data. It is adequate for analysis, but, of course, where such data restrictions are less severe, more, and more detailed, information can be derived from the data.

Below, we shall see that in our case the data pertaining to recentness interval $[0, 1)$ must be left out of the analysis for reasons of migration-specific underenumeration. This leaves at best five annually-spaced data points for each observed cohort. Therefore we have little choice here but to limit our piecewise analysis to no more than a single time interval with $t_0 = 0$ and $a = 5$ for each cohort.

As far as the estimation of the hazard function is concerned, we have to distinguish carefully between the two basic research designs which are possible. They are random sampling from among the cohort members, and a full enumeration of all cohort members. We shall discuss both in some detail so as to provide clear guidelines for the procedure to estimate the hazard function in each case.

Since our data set derives from a population census, which is a full enumeration, we are, with our data, of course, unable to illustrate the procedure of hazard function estimation in a random sampling design.

4.4.1 Estimation in the Case of Random Sampling

In the case of random sampling, one uses the information from the realized sample to make inferences about the parameters of the hazard function of the full cohort. Here, there are a number of criteria by which to judge the quality of an estimator. They relate to the expected value and to the variance of the sampling distribution of the estimator.

Generally, one prefers both the absence of any *bias* and the smallest possible variance. The absence of bias means that, averaged over the universe of realizable samples, the estimator produces the true value of the cohort parameter. A small variance implies that the probability of obtaining an estimate from any one given sample realization that deviates widely from the true value of the cohort parameter is small. An unbiased estimator is said to be the most *efficient* if its variance is the smallest possible variance attainable amongst estimators. If this is the case, then the estimator is called the minimum variance unbiased estimator (MVUE) or best estimator.

When it is not possible to assess the properties of bias or efficiency of an estimator for a finite sample size, then this may sometimes still be possible in the case of sampling with replacement where the sample size is allowed to increase without bound. Such properties are called the asymptotic properties of an estimator.

Thus, an estimator might be shown to be *asymptotically unbiased* (approaching unbiasedness) and *asymptotically efficient* (approaching the MVUE) as the sample size increases without bound. Further, if in this case an estimator's bias and variance both approach zero (they both vanish), then the estimator is said to be *consistent*. Thus, given a consistent estimator, then one can always choose a

sample size such that the estimate lies within an arbitrarily small neighbourhood of the true value of the cohort parameter with a probability that lies within an arbitrarily small neighbourhood of 1.

We note that, while good asymptotic properties of an estimator are interesting, they offer limited certainty in the case of finite sample sizes, and the less so the smaller the sample size. In particular, asymptotic properties do not specify a minimum sample size which will guarantee that the asymptotic properties are approached with a given degree of precision. Further, in practical applications of sampling among cohort members, sampling will never be with replacement, so that a basic assumption underlying the asymptotic properties is violated.

Analysis of the quality of estimators has given rise to the widespread application of *maximum likelihood estimators* (MLEs). The informal intuition which originally led to ML estimation is that, generally, under some imaginable values of the unknown cohort parameters, the observed data are a more probable sample outcome than under other conceivable cohort parameter values. ML estimation then is concerned with establishing the likelihood of alternative conceivable cohort parameter values, given a set of observed data. Specifically, an MLE produces such estimates of the unknown true cohort parameter values that they maximize the probability that the actually observed data are realized in any one given sample of a specified size.

As estimators, MLEs have a number of good properties. They result in estimates which are at least asymptotically unbiased; which are asymptotically efficient; and which are consistent. Further, they are asymptotically normally (Gaussian) distributed. Finally, they are invariant under common transformations: If $\hat{u}$ is the MLE of some random variable u, and $f(u)$ is some continuous function of u, then $f(\hat{u})$ is the MLE of $f(u)$.

Let us assume random sampling with replacement. Then it is, for example, easy to demonstrate that ordinary (linear or non-linear) least squares (OLS) estimates are ML estimates of the true cohort parameters if, for all values of the independent variable, the dependent variable is independently and identically distributed (iid) as a normal (Gaussian) distribution. In this case, the estimates are also unbiased.

However common the application of ML estimation may have become, caution is always appropriate, since the properties of MLEs which hold true in general, are only asymptotical. So, as a consequence of this, it can for example not be excluded a priori that for some given finite sample size and some given cohort parameter, there exists some other estimator which has a smaller variance than the MLE.

The application of maximum likelihood estimation in the case of random sample data similar to our data, is straightforward. Consider our single piecewise interval of observation $[t_0, t_0+a)$ with $t_0 = 0$ and $a = 5$. Suppose that an event occurs to a member of the cohort at some time point t_i on this interval.

Then in ML estimation we evaluate the associated instantaneous probability of this event, that is, the probability density at t_i. This density is, of course, given in general by the derivative of $(1-P_0(t))$. It is useful to remember here that this derivative equals the product of $P_0(t)$ and $\mu(t)$; see for example equation (11) or (16). Of course, this result can also be derived directly from the three postulates: the instantaneous probability in question is given by $P_0(t_i)\cdot P_1(t_i, t_i+\Delta t)$, letting $\Delta t \to 0$.

We repeat this evaluation similarly for all other observed events on the interval of observation. The function to be maximized then is the joint probability for all observed events on the interval. Since by our postulates the occurrences of events are stochastically independent, this joint probability is the product of the individual probabilities. The joint probability is called the likelihood function, and its global maximum is the ML estimator.

For numerical accuracy, it is always recommended first to take the natural logarithm of the likelihood function. Further, since numerical algorithms designed to locate extremes of functions are traditionally formulated as minimization routines, the log-likelihood function is commonly multiplied by -1 and its global minimum is then determined instead.

We note that under certain circumstances the assumption of the independence of events may be violated. If, for example two cohort members belong to a single family, then the occurrence of an event to one family member may well be related to the occurrence of an event to another family member. Within the cohort, such relationships may give rise to stochastic dependence. As discussed earlier, this problem can be mitigated by eliminating heterogeneity to the maximum extent possible.

For example, by considering males and females as separate cohorts and by partitioning cohorts comprising broad age ranges into several distinct subcohorts each more uniform in age, the number of cohort members having some form of mutual association will be much reduced. There will, in that case, still be dependence; but by breaking a heterogeneous cohort up into distinct internally homogeneous subcohorts, this dependence is moved from within the single cohort to between the distinct subcohorts. ML estimation is not prejudiced by the existence of any associations between subcohorts which are analysed separately.

Two further issues must be considered here, however. They are censoring and the parameterization or specification of the hazard function.

Censoring merely means that there is less than full observation. In the present case, this takes two distinct and unrelated forms.

First, it is quite possible that not all cohort members experience the event in question on the observed interval of time $[t_0, t_0+a)$. This eventuality is called right-censoring: no event for these cohort members has been observed on $[t_0, t_0+a)$, and, if such an event does occur, then for these members this will be later, on $[t_0+a, \infty)$.

If these individuals are left out of the likelihood function, then the *exposure* is limited to only those cohort members who do experience the event. This would result in an estimated hazard function which improperly inflates the propensity to experience the event. Therefore, these cohort members are included as an additional multiplicative factor in the joint probability, that is, in the likelihood function, each with the appropriate probability $P_0(t_0+a)$.

Second, in the case of our observations, the exact timing of events is not observed. All we have is that events are recorded as having happened within the last year, within the year before, and so on. This is called interval censoring: full timing detail on events occurring on any one such annual interval has not been recorded. In principle, such interval censoring is easily remedied by making a reasonable distributional assumption about the timing of events within each such annual time interval.

We note that in our case both kinds of censoring are uninformative. There exists no dependence between the occurrence of events on the one hand and the length of observation in the case of the right censoring on the other. Nor does there exist any dependence between the occurrence of events and the length of the intervals in the case of the interval censoring.

A final issue in ML estimation is the *parameterization or specification of the hazard function*. Essentially, here, there are two approaches to ML estimation. They are non-parametric and parametric estimation.

In the case of non-parametric estimation, no functional form for the hazard function is specified, and an appropriate estimator such as the Kaplan-Meier or product limit estimator is used. While circumventing the need to specify some functional form for the hazard function might seem appealing, in the present context this has a number of fundamental drawbacks.

First, we have established earlier that in the measurement of migration the observed data point(s) relating to the most recent events are unreliable. In estimating the hazard function, we ignore any such unreliable data points. However, ultimately of course, it is our goal to use the estimated hazard function to supply us with more reliable estimates for these recent data points. This requires that we must be able to extrapolate from information derived from less recent but more reliable data points. Such extrapolation is not possible without some explicitly parameterized functional form for the hazard function.

Secondly, as we just saw, we have to deal with interval censoring. This requires a distributional assumption about the timing of events within each annual interval. Any such assumption at least implicitly involves some parameterized functional form for the hazard function.

Finally, depending on the data quality, in particular on the degree of heaping and shifting in the reporting of event timings, for example through digit preference or through reference date preference, an element of graduation of the recorded data might be desirable. This, too, requires that a parameterized functional form be specified for the hazard function.

As a consequence, parameterized ML estimation is the indicated procedure when dealing with random sample data. In the exploratory phase in the case of data not previously analysed, at least piecewise constant, piecewise linear, quadratic and cubic polynomials, and piecewise exponential (Weibull) specifications should be evaluated, compared and contrasted.

A piecewise constant hazard function may seem out of order, given the well-established empirical fact that migration intensities clearly vary with time (that is, with age) as one traces the life history of a cohort. However, a piecewise constant hazard provides an excellent *benchmark* against which to assess the performance of alternative functional specifications of the hazard function. It has the benefit of parsimony, and it should only be discarded if it is outperformed by alternatives.

The exponential is indicated in particular, since in the extensive attempts of Rogers and Castro (1981) to find model forms, the only form which provided some degree of fit over the human age range proved to be a linear combination of exponentials. However, their findings were tentative, and by no means do they preclude the exploration of reasonable alternative functional forms.

As to polynomial specifications, it is not recommended to extend the exploratory analysis beyond cubics, particularly if the number of data points for each piecewise interval is limited. While higher degree polynomials can always be fitted satisfactorily from a statistical point of view, even to catch all the data points in a saturated scenario, they oscillate wildly. Consequently, interpolation

and extrapolation are unlikely to produce results which properly represent actual cohort experience.

We note that in ML estimation a so-called semi-parametric approach is also used, as an alternative to parametric and non-parametric strategies. It is an intermediate approach consisting of stepwise estimation on subintervals, and assuming a piecewise constant hazard on each of the smaller subintervals.

Semi-parametric estimation is quite similar to our method of dealing with the entire migration life history of a cohort. A major difference, however, is this: Within each subinterval, we clearly opt for the assessment of alternative formal specifications of the piecewise hazard function in addition to a constant hazard.

In the present context, it may also be interesting to refer to Courgeau and Lelièvre (1992). This publication is entirely devoted to event history analysis within a maximum likelihood estimation framework and from a demographic perspective. It includes multivariate cases, as well.

We conclude the treatment of sample data by noting a special problem. As always, in addition to testing the significance of the parameters and the establishment of appropriate confidence intervals, exploring residuals belongs to the first line approach to assessing the quality of the fit of an estimated hazard function. However, above we have also suggested that one of the tasks of the proposed parameterization of the hazard function might be the graduation of unreasonably irregular data. Assessing the goodness of fit and the graduation of observed data are two tasks which cannot properly be mixed.

Assessing the goodness of fit assumes that, apart from sampling variability, the data are recorded without error. Applying graduation of the recorded data, on the other hand, is based on the assumption that there does exist additional systematic non-sampling error. If the need for graduation has been established, then graduation should in principle be performed before ML estimation.

However, unless one is willing to use a non-informative DIY "file and fill" polishing tool, then properly graduation requires the a priori specification of the hazard function. It is precisely the ML or other estimation procedure which enables one to assess which is the best amongst the various functional forms proposed, and which then provides its parameter values. Clearly, it becomes problematic to assess the goodness of fit if the data also suffer from non-random error. In our analysis of the data from Thailand, below, we shall return to this issue of goodness of fit and graduation from a slightly different perspective.

Next, we briefly set out the principal framework for when we are dealing with a full cohort enumeration, rather than a random sampling research design.

4.4.2 Estimation in the Case of a Full Enumeration

In the case where a cohort has been fully enumerated, all the above considerations concerning random sampling are irrelevant. There is an observation for each individual member of the cohort. Consequently, there is no uncertainty associated with sampling variability, that is, with partial observation based on a random selection of some subset of the cohort.

At first sight, this might perhaps seem odd, since we are dealing with a stochastic process, where the behaviour of individual cohort members is governed by the laws of probability. However, it is important to distinguish between the formulation of the probability laws subject to which cohort members are on the one hand, and realized outcomes on the other. This is, for example, identical to the position taken in quantum mechanics which is essentially based on a rather similar theoretical framework.

To each exposed individual cohort member and on any single interval $[t, t+\Delta t)$ as $\Delta t \rightarrow 0$, $\forall t \in \mathbf{R} \backslash \mathbf{R}^{-}$, either an event does occur or no event occurs. Before t has been reached, there are two conceivable outcomes for this interval for each exposed cohort member. Once the interval has been passed, one can observe *with certainty* which of the two possible outcomes actually did occur. Stated informally, one must distinguish between what, anticipating, might occur and what, as time has passed, in actual fact did occur.

So, in the present context, at the time point of the census enumeration, the stochastic process has run its course, and there is a single realization for the cohort, namely the set of current observations, free from any stochastic uncertainty.

Not only does this imply that techniques such as the construction of confidence intervals and significance tests are out of order, here. It also removes the stochastic raison d'être, expressed in terms of bias, efficiency and consistency, of the concept of maximum likelihood estimation.

This is not to say that there may not be a place for ML estimation. However, its appropriateness will have to be argued on the basis of other than such considerations relating to random sampling. Specifically, the quality of estimators of the parameters of the hazard function in the case of a full cohort enumeration can be assessed *only on the basis of the goodness of fit* of the resulting hazard function.

We recall, here, the reservation made earlier about the possible dual purpose for the hazard function of estimation and graduation. If the hazard function estimated is also used in a graduating role, then goodness of fit measures become difficult to interpret on their own merit. We shall return to this in some more detail, below.

Of course, when dealing with data which have not previously been analysed, the *same functional forms* should be explored, compared and contrasted in the case of a full enumeration as when dealing with a random sampling design. The minimum set comprises piecewise constant, piecewise polynomial up to cubic, and exponential specifications.

We shall now proceed by outlining the basic approach to *hazard function estimation* and to the *adjustment for underenumeration* using our data from the 1970 Thai PHC.

In order to keep the discussion as straightforward and clear as possible, we select a first degree polynomial as our piecewise functional form, and combine this with the ordinary least square (OLS) approach to curve fitting.

Of course, a piecewise constant hazard function would be even more straightforward. However, discussing the case of a piecewise linear hazard is more general, in that it also shows how to deal with higher degree polynomials, as well as how to simplify matters when a constant hazard is assumed.

4.5 Estimation of the Hazard Function and Adjustment for Underenumeration

As set out above we shall now turn to the estimation of the hazard function $\mu_{RB}(t)$ and to the adjustment of the migrants for underenumeration, using the data of table 7. As explained, here we shall use a piecewise first degree polynomial specification of $\mu_{RB}(t)$ individually for each cohort, and we shall use OLS estimation so as to obtain estimates of the parameters of each of these cohort-specific hazard functions.

The results can then be compared and contrasted with outcomes for alternative specifications of the hazard function with a view to exploring which specification is to be preferred in our empirical case. However, such a comparative analysis falls outside the scope of the present paper.

Also, as discussed, we shall limit ourselves to considering one time interval only, namely 1970-1965, in the piecewise estimation procedure for each cohort, and we shall not extend the estimation procedure piecewise further back into the life history of each cohort.

Before we can proceed, we now first have to settle the issue of which data points are considered unreliable as a consequence of migration-specific underenumeration.

Based on local expert knowledge of the 1970 PHC procedures and experience (NSO, 1982; Wanglee, 1982), supplemented by inspection of the data for all individual cohorts with the aid of graphs similar to graph 1 of figure 3, we have sufficient reason to assume that the data may be considered reliable in the above sense from time point 1969 onwards back in time. So, in the hazard function estimation procedure, we shall discard all data relating to more recent time points. In our case, the latter comprise all data for time point 1970.

In other words, we shall consider cohort mass data only where the fist event is at least one year prior to the enumeration. This leaves us with five mass data points for each cohort, corresponding to time points 1969, 1968, …, 1965, respectively.

For convenience, we translate the calendar years to a simpler time scale, setting t_0 to 0, t_1 to 1, t_2 to 2, and so on, with t_0 representing 1970 and t_5 representing 1965.

From equation (36a), we see that when using a piecewise first degree polynomial (that is, a piecewise linear) specification of $\mu_{RB}(t)$, we shall need to estimate 3

parameters for each cohort: a level parameter representing the initial condition of the cohort, and the two hazard function parameters.

Given five data points, this leaves some freedom for the individual data points to deviate from the estimates. Now, recall that in our research design we cannot allow for any variation in the data due to random sampling since we have a full count, free of any sampling error. Then, put in extremes, this deviation can be interpreted in one of two ways.

The first interpretation of any such deviation is as unexplained variance: the estimator does not capture the full information contained in the data. The quality of such capture can then be expressed in terms of goodness of fit statistics.

Alternatively, the data can be interpreted as deviating from the true relationship due to systematic measurement errors as a consequence of shifting and/or heaping in the reporting of event timings. Such errors are routinely corrected by graduation, evening out the data. Some non-saturated model is normally used as a correcting graduator.

As explained, in the second interpretation, goodness of fit statistics cannot then be interpreted as such, since the assumption is that the observed data are in error and that they do not represent the true relationship. It is the graduator, that is, the non-saturated estimator, which is assumed to represent the true cohort behaviour. The obvious graduator is, of course, the hazard function. In other words, given that the graduator is true, a goodness of fit statistic for the graduator now is a measure of the degree of the shifting and/or heaping in the reporting of event timings. Therefore, this interpretation does not allow for a straightforward comparison of alternative specifications of the hazard function, since this would require an a priori explicit formulation of the data errors first.

Let us return to the hazard function. We now have, individually for each cohort,

$$\mu_{RB}(t) = \beta_1 + \beta_2 t \,, \tag{52}$$

where β_1 and β_2 are the unknown parameters of the hazard function. Integrating (52) over the time interval $[0, t)$ and substituting in (36a), we have

$$K_B(t) = K_B(0) \exp\{-(\beta_1 t + \tfrac{1}{2}\beta_2 t^2)\} \,. \tag{53}$$

Let us write $\exp(\beta_0)$ for $K_B(0)$ and $\exp\{y(t)\}$ for $K_B(t)$. Then, taking logarithms, we obtain

$$y(t) = \beta_0 - \beta_1 t - \tfrac{1}{2}\beta_2 t^2 \,, \tag{54a}$$

or, more conveniently,

$$y(t) = \beta_0 + \beta_1(-t) + \beta_2(-\tfrac{1}{2}t^2) \; . \qquad (54b)$$

From table 7 we have, individually for each cohort, five data points, one each for $t = 1$, $t = 2$, … , $t = 5$. Taking the logarithm of each data point, we have the set of $y(t)$ values.

The parameters of equation (54) can now be obtained by OLS estimation, using any well-validated statistical software application.

Some caution is required, here, however. Most quality statistical software is almost exclusively designed for use in random sampling research designs. This means that care must be taken in two ways. First, it should be established that the mathematical statistics underlying the routines used is valid in a full enumeration research design. Second, a careful selection of outputs must be made, discarding any results which are specific to a random sampling design.

The calculations required here are straightforward, and in order to be able to verify the statistical software used, it is useful to perform the calculations independently. We give the necessary equations next.

In matrix notation and using (54b), we have, individually for each cohort,

$$\boldsymbol{y}(t) = \boldsymbol{T}(t)\boldsymbol{\beta} + \boldsymbol{\varepsilon} \; , \qquad (55)$$

where the vector $\boldsymbol{\varepsilon}$ represents the deviations of the individual log data points for a cohort from the regression plane, given the values of the parameters β_0, β_1 and β_2 of that cohort. These three parameters make up the column vector $\boldsymbol{\beta}$. Column vector $\boldsymbol{y}(t)$ has the cohort's log data points (the logarithms of the values of $K_B(t)$) for $t = 1$, $t = 2$, … , $t = 5$; that is, the log data point for $t = 0$ is left out of the estimation procedure. Matrix $\boldsymbol{T}(t)$ is identical for each cohort, since the values of the independent variable (the time points) are the same for each cohort. It is structured as follows:

$$\boldsymbol{T}(t) \; = \; \begin{bmatrix} 1 & -1 & -\tfrac{1}{2} \\ 1 & -2 & -\tfrac{1}{2}2^2 \\ 1 & -3 & -\tfrac{1}{2}3^2 \\ 1 & -3 & -\tfrac{1}{2}4^2 \\ 1 & -5 & -\tfrac{1}{2}5^2 \end{bmatrix} . \qquad (56)$$

The OLS estimator $\hat{\boldsymbol{\beta}}$ of the parameter vector $\boldsymbol{\beta}$ is then given by

$$\hat{\boldsymbol{\beta}} = \{\boldsymbol{T}(t)'\boldsymbol{T}(t)\}^{-1}\boldsymbol{T}(t)'\boldsymbol{y}(t) . \quad (57)$$

Evaluating (57) completes the estimation procedure of the hazard function $\mu_{RB}(t)$ for each cohort. At the same time, it completes step one of the adjustment procedure for underenumeration. Recalling that the observed data point $K_B(0)_{obs}$ was discarded in the estimation procedure for being unreliable, the estimate for β_0 now yields our initial estimate of the true value of $K_B(0)_{init}$ by

$$K_B(0)_{init} = \exp(\beta_0) . \quad (58)$$

The results for the 13 cohorts are given in table 8. The table also includes a column giving the ratio of the explained variance to the total variance (R^2), multiplied by 100. As we saw above, R^2 may alternatively be interpreted as a measure of the goodness of fit, or as a measure of the degree of heaping and/or shifting in the reporting of event timings.

Table 8 Parameter Values of Hazard Function $\mu_{RB}(t)$, Initial Estimate $K_B(0)_{init}$ (×100), and R^2 (×100)

Cohort	β_0	β_1	β_2	$\exp(\beta_0)$	$R^2 \times 100$
[5, 10)	7.63645	0.01581	-0.00127	2,072.365	99.91450
[10, 15)	7.62094	0.02178	-0.00265	2,040.470	99.99054
[15, 20)	7.59683	0.05057	-0.00677	1,991.867	99.98939
[20, 25)	7.41522	0.11106	-0.01927	1,661.077	99.99533
[25, 30)	7.08559	0.03977	-0.00346	1,194.631	99.94821
[30, 35)	7.00347	0.02921	-0.00277	1,100.447	99.91446
[35, 40)	6.79393	0.02047	-0.00162	892.417	99.96858
[40, 45)	6.53774	0.01602	-0.00133	690.721	99.96167
[45, 50)	6.22677	0.01608	-0.00140	506.119	99.86066
[50, 55)	6.05746	0.01702	-0.00171	427.287	99.94484
[55, 60)	5.80621	0.01221	-0.00079	332.358	99.67354
[60, 65)	5.47601	0.01632	-0.00183	238.893	99.82899
[65, ∞)	5.90772	0.01605	-0.00065	367.867	99.68859

From this table we see that all cohorts have a negative value of parameter β_2. This indicates that, as we go back into the life histories of the cohorts, the hazard declines. In other words, as real time progresses, the intensity of migration into Bangkok increases. This finding agrees with observations for the kingdom as a whole on the basis of the census, made in the special analytical subject report on migration published as a part of the census publishing programme (NSO, 1972-1977).

Further, we see very high values of R^2 for all cohorts. This is especially remarkable, since we are only using a linear specification of the cohort hazard functions. One may expect to obtain even better values when using alternative specifications which allow for more flexible variation of the hazard functions over the piecewise life histories of the cohorts.

It is also interesting to note that the values of R^2 are slightly lower for older cohorts, and to some extent for the youngest cohort as well. One explanation may well be that the average quality of the reporting of event timings for these cohorts is somewhat less. One obvious cause of this is that for these cohorts, more so than in the cohorts ranging in age from, say, 15 to 55, other persons will be answering the census questions or completing the census questionnaire on behalf of the cohort member concerned.

An alternative or complementary explanation could be that the linear hazard performs relatively less well for these cohorts. Only a comparative assessment for all cohorts of reasonable alternative specifications for the hazard functions can eliminate this second explanation.

Let us give an example of the findings reported in table 8. For instance, for the cohort [20, 25), we have the hazard function $\mu_{RB}(t) = 0.11106 - 0.01927t$. So, the instantaneous inmigration rate at the time of the enumeration is estimated to be 0.11106.

Note that all data in this table -- as those in all other tables -- have been rounded to the number of decimal places shown. Any manual recalculations for the purpose of verification may therefore show minor discrepancies.

Now, suppose that this instantaneous migration rate is stationary for, say, two years and irrespective of origin and destination. Then we may use theorem 3A to find, for example, that the probabilities to make 0, 1, and 2 moves within these two years equal 80.1%, 17.8%, and 2.0%, respectively. Using theorem 2A, one may also say that under these assumptions one in five in the cohort will make at least one move within two years.

Our two assumptions allowing us to use theorems 2A and 3A may well be unrealistic in this instance. However, when appropriate, such probabilities and ratios give a useful alternative perspective on the migration intensity to which the cohort was subject at the time of the enumeration.

Further, for the same cohort we see in table 8 that $\beta_0 = 7.41522$. Using (58) we therefore have that $K_B(0)_{\text{init}} = \exp(7.41522)$, resulting in an initial estimate of the cohort mass at $t = 0$ of 166,108 persons. (Recall that cohort mass sizes in table

8 are in hundreds.) According to table 7, the value observed in the census, $K_B(0)_{obs}$, was 155,340 persons.

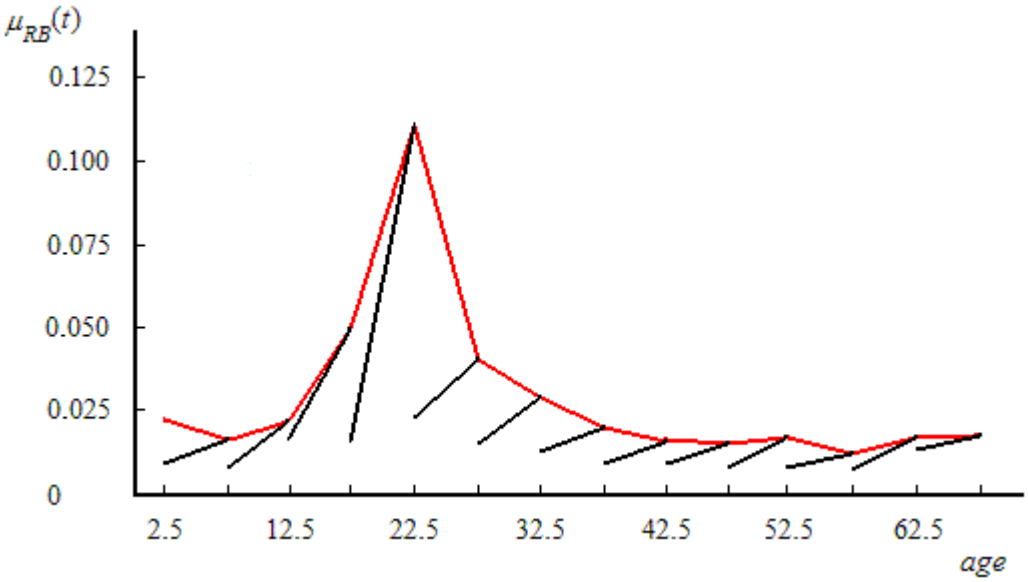
As a consequence, through the first step of our estimation procedure, and using (48), we have recovered θ_m = 166,108 – 155,340 = 10,768 cohort members who were missed by the enumerators. Specifically, of course, these are all migrants who arrived within the year immediately prior to the census. Figure 3 illustrates this procedure graphically.

Figure 4 shows the estimated hazard functions for all 13 cohorts. On the horizontal age axis, the cohort graphs (depicted in black) have each been centred on the midpoint of the age interval defining the respective cohort.

Figure 4 also shows the instantaneous period rates as of the time point of the enumeration (the red curve), linearly interpolated for ages for which there are no period data. The period rates for ages [2.5, 7.5) were derived by extrapolation using the relationship parent-child.

Note that the cohort and period data for age interval [62.5, ∞) have been graphed as if the interval were [62.5, 67.5). This was done so as to maintain the vertical scale, so that rates of change between cohorts, and from period to period, may be compared.

Figure 4 Estimated Hazard Functions $\mu_{RB}(t)$ for All Bangkok Male Cohorts from 1965 to 1970, and Instantaneous Period Inmigration Rates for 1970
(*Black: estimated hazard functions; Red: period inmigration rates*)



Clearly, the general shape of the graphs agrees well with well-known life history patterns of migration as governed by common life-cycle events such as enrolling in higher education, joining the labour force, retiring, and so on.

However, the quality of the results also has to be assessed in view of the performance of reasonable alternative formal specifications of the hazard function as discussed earlier. Obvious alternative specifications to be considered in addition to $\mu_{RB}(t) = \beta_1 + \beta_2 t$ are

$$\mu_{RB}(t) = \beta_1 \,, \tag{59}$$

that is, a piecewise constant hazard, and

$$\mu_{RB}(t) = \beta_1 \exp(\beta_2 t) \,, \tag{60}$$

a piecewise exponential hazard. Specification (59), while empirically unrealistic in most applied contexts, is the most basic specification possible. It can therefore serve as a useful *benchmark* when comparatively assessing the performance of alternative specifications.

Further, one might also consider piecewise second and third degree polynomial specifications:

$$\mu_{RB}(t) = \beta_1 + \beta_2 t + \beta_3 t^2 \,, \tag{61}$$

and the saturated

$$\mu_{RB}(t) = \beta_1 + \beta_2 t + \beta_3 t^2 + \beta_4 t^3 \,. \tag{62}$$

However, as explained earlier, considerable caution should be exercised here. The closer the specification approaches saturation, the more closely any unobserved errors in the distribution of the observed values of $K_B(t)$ due to factors such as shifting and/or heaping in the reporting of event timings, will be *translated directly* into the form of the estimated hazard function $\mu_{RB}(t)$. Also, one has to be able to justify that such higher degree polynomial specifications do reasonably represent the true hazard as time and age vary.

Yet further specifications of the hazard function might be appropriate in addition to the ones mentioned here. Such further comparative analysis using different specifications of $\mu_{RB}(t)$ falls outside the scope of this paper, however.

Note that the estimation of (59) is elementary. All that needs doing is taking away parameter β_2 in vector $\boldsymbol{\beta}$ of equation (55), and taking away the third column in matrix $\boldsymbol{T}(t)$ as defined in equation (56). The estimator (57) then remains valid.

Similarly, for equation (62), parameters β_3 and β_4 need to be included in vector $\boldsymbol{\beta}$, and additional columns $(-\frac{1}{3}t^3)$ and $(-\frac{1}{4}t^4)$ need to be included in matrix $\boldsymbol{T}(t)$. For equation (61), the required additions are parameter β_3 and column $(-\frac{1}{3}t^3)$, respectively, only.

We reiterate, that the hazard functions obtained represent cohort instantaneous rates of *inmigration* into Bangkok. They were merely *derived as* outmigration rates by reversing the order of time. As such, they are proper occurrence / exposure rates.

Further, analysts wishing, for example, to construct a period multistate life table can do so *only for* $t = 0$, since, as discussed, this is the only time point at which the *cohorts* and the male *population* of Bangkok coincide. In line with the development of the theory -- see equation (41) --, such a life table requires period instantaneous *outmigration* rates. Obtaining such rates requires three additional steps.

First, step two of the adjustment procedure for underenumeration has to be completed, resulting in final estimates of the cohort mass $K_B(t)_{\text{fin}}$ for each cohort and for all six observed time points; we shall turn to this shortly. Next, of course, the analysis has to be repeated similarly for inmigration from Bangkok into the rest of the kingdom. Finally, the period inmigration rates have to be converted into period outmigration rates, using

$$\mu_{RB\text{out}}(0) \;=\; \mu_{RB}(0) \cdot K_B(0)_{\text{fin}} \,/\, K_R(0)_{\text{fin}} \,, \qquad (63a)$$

and

$$\mu_{BR\text{out}}(0) \;=\; \mu_{BR}(0) \cdot K_R(0)_{\text{fin}} \,/\, K_B(0)_{\text{fin}} \,, \qquad (63b)$$

where $\mu_{RB\text{out}}(0)$ is the period instantaneous outmigration rate at $t = 0$ from the rest of the kingdom to Bangkok, and $\mu_{BR\text{out}}(0)$ is the similar rate for outmigration from Bangkok to the rest of the kingdom. Equations (63a) and (63b) can easily be seen to represent a straightforward rebasing procedure.

We now turn to *step two of the adjustment procedure* for underenumeration. While step one focused exclusively on *migration-specific* underenumeration, step two centres on general underenumeration due to *all other* causes. In other words,

we shall now focus on underenumeration which is not recentness-related, and which thus affects the cohorts irrespective of the duration of residence.

In order to be able to conduct step two of the adjustment procedure, independent information is required on the total underenumeration θ for each cohort, as discussed; see also equation (49).

As an outcome of the overall 1970 PHC quality assurance process, NSO produced such information for internal use (NSO, 1982). The findings for Bangkok are reproduced (after rounding) as columns 1 and 3 of table 9. We computed the remaining columns.

Table 9 Estimates of the Overall Underenumeration of Bangkok Male Cohorts 1970 (cohort mass data in absolute numbers ×100)

Cohort	Enumerated Cohort Size [1]	Estimated Deficit [2]	Estimated Cohort Size [3]=[1]+[2]	Underenum. Rate (×100) [4]=100×[2]/[3]	Adjustment Multiplier [5]=[3]/[1]
[5, 10)	2,052.6	41.2	2,093.8	1.97	1.0201
[10, 15)	2,015.2	-4.0	2,011.3	-0.20	0.9980
[15, 20)	1,948.7	171.4	2,120.1	8.08	1.0879
[20, 25)	1,553.4	437.9	1,991.3	21.99	1.2819
[25, 30)	1,170.7	186.0	1,356.7	13.71	1.1589
[30, 35)	1,084.0	-69.4	1,014.6	-6.84	0.9360
[35, 40)	883.5	-60.3	823.2	-7.33	0.9317
[40, 45)	684.8	-7.0	677.8	-1.03	0.9898
[45, 50)	501.1	38.8	539.9	7.19	1.0775
[50, 55)	423.4	16.4	439.7	3.72	1.0386
[55, 60)	330.6	10.1	340.8	2.97	1.0306
[60, 65)	244.6	1.0	245.6	0.40	1.0041
[65, ∞)	355.6	1.4	357.0	0.41	1.0041
Total	13,248.2	763.7	14,011.8	5.45	1.0576

Two aspects of table 9 are noteworthy in this context. First, from column 2 we see that four of the cohorts were found to have been "overenumerated". This is a consequence of the effects of age heaping and age shifting in the observed data which has been corrected in column 3.

Second, comparing column 1 with column 1 of table 7, we note several minor discrepancies in the 1970 cohort masses reported as enumerated. Clearly, some undocumented corrections have been applied by the National Statistical Office between the publication of the official census reports (NSO, 1972-1977) and the production of the data reproduced in table 9. However, these discrepancies are of no consequence if we use the underenumeration rates $\hat{u}$ or the adjustment

multipliers $\hat{a}$ computed for each cohort in table 9 as columns 4 and 5, respectively, rather than the absolute cohort mass sizes.

Let us now explain the procedure of step two of the adjustment procedure by means of a worked example for one of the cohorts. It may be useful to refer to figure 3 which displays this procedure graphically.

Consider again the cohort [20, 25). So far for this cohort, we have $K_B(0)_{\text{obs}} = 155{,}340$ and $K_B(0)_{init} = 166{,}108$, and thus $\theta_m = 166{,}108 - 155{,}340 = 10{,}768$.

According to table 9, the true cohort mass at $t = 0$, $K_B(0)_{fin}$, can be computed as $1.2819 \times 155{,}340 = 199{,}130$. So, the total deficit, $\theta = K_B(0)_{fin} - K_B(0)_{\text{obs}}$, is $199{,}130 - 155{,}340 = 43{,}790$.

In other words, we have already recovered 10,768 cohort members out of this total deficit of 43,790 in step one of the adjustment procedure. These 10,768 are cohort members who have been missed in the enumeration on account of the recentness of their move. They belong to duration of residence class [0, 1).

From equation (49) we have that this still leaves $\theta_g = \theta - \theta_m$, or $43{,}790 - 10{,}768 = 33{,}022$ cohort members unaccounted for due to all other causes of underenumeration. Therefore these 33,022 cohort members are distributed proportionally over all duration of residence classes.

Now recall that from table 6 we have for this cohort, by duration of residence, in hundreds:

[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, 5)	[5, ∞)	Total
52.1	119.5	84.0	53.9	31.5	1,212.4	1,553.4

After adjustment step one, we then obtain, in hundreds,

[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, 5)	[5, ∞)	Total
159.8	119.5	84.0	53.9	31.5	1,212.4	*1,661.1*

Here we have added the 107.68 (×100) recovered recent migrants to the 52.1 (×100).

Now, to accommodate step two of the adjustment procedure, we have to distribute the remaining 33,022 cohort members who are still unaccounted for, proportionally over all duration of residence classes.

Therefore we compute (330.22 + 1,661.1) / 1,661.1, giving us an adjustment factor of 1.1988 to be applied to all duration of residence classes as obtained *after step one* of the adjustment procedure. And so, we obtain, again in hundreds,

[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, 5)	[5, ∞)	Total
191.6	*143.3*	*100.7*	*64.6*	*37.8*	*1,453.4*	*1,991.3*

This means that the *overall* adjustment factor $\hat{a}$ for duration of residence class [0, 1) is computed as 191.6 / 52.1 = 3.6775. For all other duration of residence classes of this cohort -- which, as explained, do not suffer from migration-specific underenumeration --, the value of the adjustment factor $\hat{a}$ is uniformly 1.1988.

Using equation (50b) we therefore have that the underenumeration rate $\hat{u}$ for this duration of residence class [0, 1) has a value of 1 – 1/3.6775 = 0.7281.

In other words, *from the cohort aged [20, 25), nearly three-quarters of all recent migrants were missed in the enumeration*.

Referring to table 9, we see that the underenumeration rate for this cohort as a whole is 0.2199, a considerable rate in itself, but very significantly lower than that for recent migrants. This difference underlines the *disproportional propensity of migrants to be incompletely enumerated*.

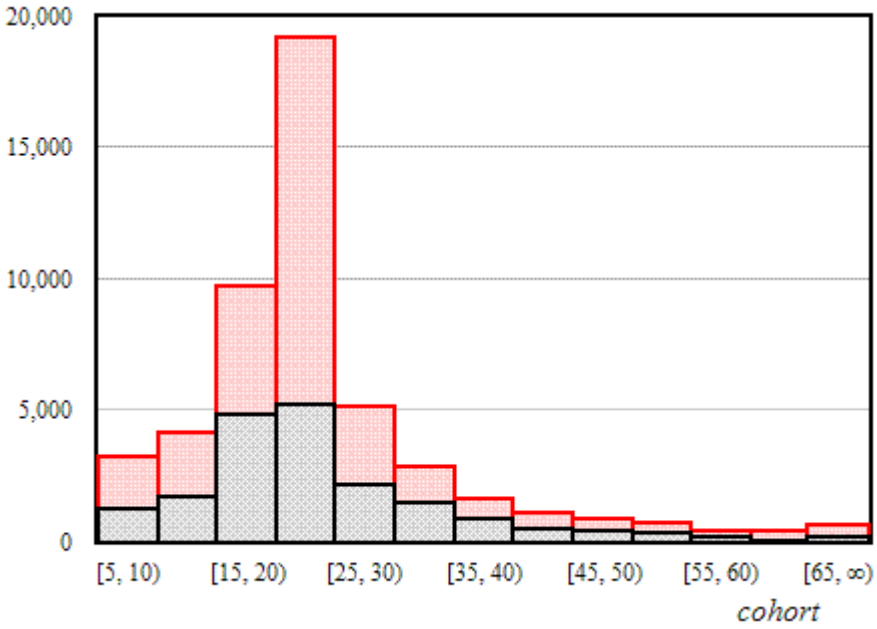
The results for all cohorts are given in table 10, below. The data in this table have been computed to a greater degree of accuracy than those in the worked example above. The cohort mass data in this table are in units, rather than in hundreds as in the previous tables based on the 1970 Thai PHC. Further, the column labelled Total contains the sums of duration of residence classes [0, 1), [1, 2), … , [4, 5). So it excludes duration of residence class [5, ∞). Note that this latter class is included in the totals of the worked example given above.

Table 10 Distribution of the Completed Duration of Residence (Years), Bangkok Male Cohorts 1970, Before and After Adjustment for Underenumeration

	Cohort	[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, 5)	Total
Enumerated	[5, 10)	1,213	2,718	2,495	2,505	1,732	10,664
Estimated	[5, 10)	3,198	2,747	2,522	2,531	1,750	12,748
Adj Multipl	[5, 10)	2.6369	1.0105	1.0105	1.0105	1.0105	1.1954
Enumerated	[10, 15)	1,682	3,491	2,938	2,499	1,781	12,391
Estimated	[10, 15)	4,088	3,442	2,897	2,464	1,756	14,648
Adj Multipl	[10, 15)	2.4310	0.9860	0.9860	0.9860	0.9860	1.1821
Enumerated	[15, 20)	4,822	7,729	5,822	4,706	3,491	26,570
Estimated	[15, 20)	9,706	8,228	6,197	5,009	3,716	32,856
Adj Multipl	[15, 20)	2.0126	1.0645	1.0645	1.0645	1.0645	1.2366
Enumerated	[20, 25)	5,209	11,947	8,402	5,388	3,150	34,096
Estimated	[20, 25)	19,154	14,323	10,072	6,460	3,776	53,785
Adj Multipl	[20, 25)	3.6769	1.1988	1.1988	1.1988	1.1988	1.5774
Enumerated	[25, 30)	2,106	3,831	3,337	3,219	2,262	14,755
Estimated	[25, 30)	5,109	4,351	3,790	3,656	2,569	19,474
Adj Multipl	[25, 30)	2.4261	1.1357	1.1357	1.1357	1.1357	1.3199
Enumerated	[30, 35)	1,463	2,585	2,233	2,227	1,443	9,950
Estimated	[30, 35)	2,816	2,384	2,060	2,054	1,331	10,645
Adj Multipl	[30, 35)	1.9251	0.9224	0.9224	0.9224	0.9224	1.0698
Enumerated	[35, 40)	834	1,534	1,384	1,334	1,008	6,094
Estimated	[35, 40)	1,615	1,414	1,276	1,230	929	6,465
Adj Multipl	[35, 40)	1.9362	0.9221	0.9221	0.9221	0.9221	1.0610
Enumerated	[40, 45)	484	929	835	808	602	3,658
Estimated	[40, 45)	1,041	912	820	793	591	4,156
Adj Multipl	[40, 45)	2.1507	0.9815	0.9815	0.9815	0.9815	1.1363
Enumerated	[45, 50)	348	648	620	618	389	2,623
Estimated	[45, 50)	844	692	662	660	415	3,273
Adj Multipl	[45, 50)	2.4243	1.0681	1.0681	1.0681	1.0681	1.2481
Enumerated	[50, 55)	326	566	570	434	369	2,266
Estimated	[50, 55)	716	583	587	447	380	2,713
Adj Multipl	[50, 55)	2.1960	1.0296	1.0296	1.0296	1.0296	1.1976
Enumerated	[55, 60)	205	358	298	373	220	1,454
Estimated	[55, 60)	405	366	305	382	226	1,684
Adj Multipl	[55, 60)	1.9741	1.0247	1.0247	1.0247	1.0247	1.1586
Enumerated	[60, 65)	90	283	308	221	170	1,072
Estimated	[60, 65)	373	281	306	219	169	1,347
Adj Multipl	[60, 65)	4.1468	0.9925	0.9925	0.9925	0.9925	1.2568
Enumerated	[65, ∞)	191	470	547	547	368	2,123
Estimated	[65, ∞)	594	467	543	543	365	2,512
Adj Multipl	[65, ∞)	3.1130	0.9927	0.9927	0.9927	0.9927	1.1834
Enumerated	Total	18,973	37,088	29,789	24,878	16,985	127,715
Estimated	Total	49,660	40,189	32,037	26,448	17,974	166,308
Adj Multipl	Total	2.6173	1.0836	1.0754	1.0631	1.0582	1.3022

Figure 5, finally, shows the underenumeration for recent migrants. It was directly derived from table 10, duration of residence class [0, 1).

Figure 5 Underenumeration of Recent Migrants:
Bangkok Male Cohorts 1970, Duration of Residence Class [0, 1),
Before and After Adjustment for Underenumeration
(*Black: enumerated data; Red: data after correction for underenumeration*)



Perhaps more than table 10 does this figure 5 highlight the exceedingly poor degree to which recent migrants are enumerated: The census only recorded the black bars. *The areas coloured red in this graph represent the migrants who arrived in the past 12 months and who have not been observed in the census*.

This completes our description of the procedure to adjust migration data for incomplete enumeration, or, in the case of a population registration system, for incomplete registration.

Summarizing briefly, we have achieved two important results. Based on demographic theory, we have established *theoretically justified direct methods of measuring instantaneous rates of migration as a function of continuous time*.

Second, given a standard population census from a developing country, we have demonstrated -- in a non-data set specific manner -- how demographic theory

enables us to recover detailed quantitative information on the incompleteness of the enumeration of migrants, *allowing us to correct the observed migration data for such incompleteness*.

Further, the theoretical and methodological results are general, in the sense that they apply equally to *developing countries* and to *developed countries*; and that they apply equally to *internal migration* and to *international migration*.

Also, the methods of measuring instantaneous rates described are not specific to *migration*. They apply equally to the measurement of *mortality*, *fertility*, and *other formally similar processes*.

Finally, all methods developed apply equally to *population census data*, to *population registration data*, and to *sample survey data*.

CONCLUSIONS

In this paper, we set out by highlighting the societal relevance of the systematic measurement and analysis of internal and international migration. Next we review to what extent demography currently contributes to this field. Here, we conclude that, while instruments for analysis and forecasting have reached considerable maturity, this cannot be said of methods of measurement.

In order to explore theoretically justified approaches to the measurement of *internal and international migration*, we start with a reformulation of analytical concepts and tools. Here we adopt an approach *from first principles*. After establishing a number of elementary definitions and after formulating three axiomatic postulates, we are able to arrive at a number of results through logical deduction. Many of these results are familiar to demographers, some may be less familiar or even new, at least within the discipline.

One of the principal benefits of an approach from first principles is that it provides a clear and complete insight into the assumptions underlying the theoretical results. Traditionally in demography, such assumptions often remain partly undiscussed and therefore to some extent hidden from view. This can sometimes make the application of theoretical results difficult to justify explicitly.

A full insight in the underlying assumptions also provides clear guidelines to redefine one or more of the definitions and/or postulates if it is found that any of the assumptions violates empirical conditions, or if one merely wishes to explore the effect of changing one or more of the assumptions. The latter may for example be useful in sensitivity analysis when comparing alternative theoretical formulations.

A second benefit of an approach from first principles is that it contributes to the development of demography as a science, a science with wide-ranging and valuable practical applications -- or, in other words, an applicable science --, away from demography merely as an applied science. A science has explicitly specified abstract, general and unifying theory; an applied science at best has well-defined concepts and models.

It thereby distances the discipline from the perspective of mathematical demography and analytical demography. For, mathematical demography, for instance, has developed into a collection of -- sometimes rather disparate, though frequently ingeniously designed and useful -- tools, techniques and results from mathematics and statistics applied to demographic questions and issues. Although there are several strong and coherent strands, it has not developed into a truly

fully-integrated body of theory. Too often, too, the specification of its concepts and models lacks complete explicitness. As a consequence, its systematic falsifiability leaves to be desired.

It is probably no exaggeration to claim that, in no small measure due to its focused paradigm, demography is the social science *par excellence* that lends itself to abstract, general, coherent, internally consistent and empirically falsifiable theory construction.

The third benefit of an approach from first principles is that it, again logically, leads to the specification of measurement instruments. Demography has a tradition of developing concepts and tools around available empirical data. However, theory construction based on, and building on, empirical data is an approach which is epistemologically and methodologically unsound. We referred to this distinction earlier as theory-based measurement versus measurement-based theory. It is theory that should allow one to deduce which data are required and how they should be specified so as to allow the testing of new theory and so as to apply well-validated theory.

We defined demography as the social science which describes and explains the generation and the behaviour over time and age of human cohorts. So, implicitly, we make a clear distinction between demography and population studies. Fundamental in the generation of cohorts and in the behaviour of cohorts are *individual demographic events*, events such as getting married, giving birth, making a migratory move, and dying. In other contexts, such as in labour force analysis, manpower planning, the study of illness and medical intervention, there are other events, but they may be defined in a formally similar manner.

Cohort members are defined as being at risk of experiencing such demographic events. Our point of departure then are the intensities at which such individual events occur in continuous time to cohort members over the life history of cohorts, that is, the instantaneous (birth, death, migration, and so on) rates as a function of time and age. Often we use the short term hazard functions. Hazard functions are *the sole governors of the behaviour of cohorts and of the creation of new cohorts*. Knowledge of the relevant hazard functions equates to full knowledge of cohort behaviour and cohort creation.

The question of the measurement of migration, therefore, resolves to the measurement of migration hazard functions in a theoretically and methodologically sound manner.

Clearly, in the case of migration, we do not consider the well-established indirect methods of measurement. They do not measure migration. They merely deduce the net result of unobserved migration from the study of other events.

For the direct measurement, one requires information on the *timing* of each of the individual migratory events, as well as on the *direction* of the move in question. Such information can be obtained from population registers in countries where these are kept, provided that they record migratory events. However, most countries in the world do not maintain such registers. Then, the best alternative source of information is a full population census.

Not all censuses accommodate for the measurement of timing and direction of migratory events, however. While almost all population censuses attempt to collect information on migration, some of the measurement instruments selected are of relatively inferior design. The only standard census (and survey) questions that are suitable, are questions on the *duration of residence* combined with questions on the *associated previous place(s) of residence*. Duration of residence allows the direct deduction of the timing of the events.

Methodologically, a question on durations yields information which is rich in contents. It allows for the direct measurement of migration hazard functions, that is, of *instantaneous migration rates as a function of time and age*.

In addition, the approach to the measurement of migration developed is *general*, in that it applies equally to the methodologically valid measurement of instantaneous birth rates, death rates, and to the measurement of the instantaneous rates at which any other formally similar demographic events occur.

However, the data resulting from a question on the duration of residence yield more information. They also enable one accurately to *quantify the degree of underenumeration of migrants*, and they allow for the *adjustment of the measured events for incompleteness*.

As described in this paper, this is based on the fact that, from the point of view of statistical observability, the longer a migrant has been resident in the place of enumeration, the less there will be to distinguish this person from a person who has lived there all his or her life. Specific incompleteness in the registration or enumeration of migrants is directly related to the recentness of the move.

Migrants suffer disproportionately from being missed out in data collection procedures. Table 10 and figure 5 clearly illustrate how erroneous reported migration data may be.

Underenumeration rates amongst recent migrants are uniformly high. As we have seen in the case of our empirical data for Bangkok, the lowest rate of underenumeration occurring for any cohort is just under 50%, and the highest rate is in excess of 75%. In other words, *only between 25% and 50% of all recent migrants have actually been enumerated*.

As discussed, there are no grounds to assume that the quality of the data for Thailand used in this paper are in any way very exceptional. So, if the Thai census which we have used is indeed of a reasonable quality, then one therefore has to conclude that the use of uncorrected observed migration data to describe the phenomenon of migration will generally be difficult to justify.

It is well-known that the quality of direct data on mortality and fertility in developing countries is often doubtful. It is likely that the same holds true for migration data. However, in the case of migration, such data in many developed countries might well be suspect, too. Further studies in other countries will have to be conducted to verify this.

Incidentally, high rates of underenumeration of migrants at the same time indicate that the importance of migration as a component in the process of population change is greater than hitherto assumed on the basis of direct measurements.

The potential occurrence of high rates of underenumeration in the direct measurement of migration is a point often overlooked by proponents of other direct migration measurement instruments, in particular of questions on the *place of residence a fixed number of years prior to the enumeration*.

Such questions merely measure the net effect of migration over the fixed time interval chosen. The events themselves are not recorded. As a consequence, it is impossible to use such data for the measurement of instantaneous migration rates in a theoretically sound manner. Also, as a consequence, the powerful analytical instruments derived in the process of theory development, such as the probabilities to experience multiple moves, cannot be applied.

But most important of all, such questions result in data which suffer from serious incompleteness, and which, at the same time, do *not* allow for the adjustment for such incompleteness.

To give one example, consider the highest instantaneous rate of migration measured in the case of our data for Bangkok: 0.11106 at the time point of the enumeration in 1970 for the cohort aged [20, 25). Using theorem 3A, we know that at this migration rate the probability to experience multiple moves within one year is just over 10.5%.

This means that the measurements obtained using a question on the place of residence one year prior to the census would almost match our observed data for this cohort aged [20, 25) which is defined based on the combination of a question on duration of residence and on place of last previous residence. For all other cohorts studied, the match would be even closer.

Therefore, figure 5 also gives a very good indication of how many migrants would have been missed in the enumeration if a question on the place of residence one year prior to the census would have been asked. And, as mentioned, with the latter question, the extent of such underenumeration cannot be assessed.

Finally, our findings have important implications for those organizations which advise countries on good population registration, census and survey practice, in particular the *United Nations* and its regional commissions for Africa, Asia, Europe and Latin America.

As we have explained in detail in section 4, the influential recommendations of the United Nations on the methods of measuring internal migration and international migration (United Nations, 1997; United Nations, 1998) provide insufficient guidelines so as to allow national statistical agencies to make theoretically and methodologically sound decisions on how to measure migration from population registers and in population censuses and surveys.

In addition, where the question on the duration of residence is discussed (United Nations, 1997), the current recommendations suggest measurements which are unnecessarily crude, and which severely limit the informational value of the data obtained.

REFERENCES

Al Mamun, A (2003) *Life History of Cardiovascular Disease and its Risk Factors: Multistate Life Table Approach and Application to the Framingham Heart Study*. Rozenberg Publishers

Bell, M (2005) Towards Rigorous Cross-National Comparison of Internal Migration: Who Collects What? Paper prepared for the XXV International Population Conference of the International Union for the Scientific Study of Population (IUSSP), 18-23 July 2005, Tours, France (unpublished, available from IUSSP)

Bishop, Y M M, S E Fienberg and P W Holland (1975) *Discrete Multivariate Analysis: Theory and Practice*. MIT Press

Blumen, I, M Kogan and P J McCarthy (1955) *The Industrial Mobility of Labor as a Probability Process*. Cornell studies in industrial and labor relations, vol 6. Cornell University Press

Bocquier, P (2005) World Urbanization Prospects: An Alternative to the UN Model of Projection Compatible with the Mobility Transition Theory. *Demographic Research* 12(9) <http://www.demographic-research.org/>

Burch, T K (2003) Demography in a New Key: a Theory of Population Theory. *Demographic Research* 9(11) <http://www.demographic-research.org/>

Chiang, C L (1968) *Introduction to Stochastic Processes in Biostatistics*. John Wiley & Sons

Chiang, C L (1972) On Constructing Current Life Tables. *Journal of the American Statistical Association* 67:538-541

Chiang, C L (1984) *The Life Table and its Applications*. Robert E Krieger Publishing Co

Coale, A J (1972) *The Growth and Structure of Human Populations: A Mathematical Investigation*. Princeton University Press

Coale, A J and P G Demeny (1966) *Regional Model Life Tables and Stable Populations*. Princeton University Press

Coale, A J, P G Demeny and B Vaughan (1983) *Regional Model Life Tables and Stable Populations*. 2nd Edition. Academic Press

Coale, A J and T J Trussell (1974) Model Fertility Schedules: Variation in the Age Structure of Childbearing in Human Populations. *Population Index* 40(2):185-258

Courgeau, D (1980) *Analyse Quantitative des Migrations Humaines*. Masson

Courgeau, D (1988) *Méthodes de Mesure de la Mobilité Spatiale: Migrations Internes, Mobilité Temporaire, Navettes*. INED

Courgeau, D and É Lelièvre (1992) *Event History Analysis in Demography*. Clarendon Press

Doeve W L J (1987) How Do We Measure Migration? The Preferred Migration Questions for the Global 1990 Round of Population Censuses. Paper presented at the International Conference on Urbanization and Urban Population Problems (ICUUPP), Tianjin, People's Republic of China, October 1987

Gardiner, P and Mayling Oey-Gardiner (1990) *Indonesia*. In: Nam, C B, W J Serow and D F Sly (eds) (1990) *International Handbook on Internal Migration*. Greenwood Press, Chapter 11, p207-224

Goodman, L A (1961) Statistical Methods for the Mover-Stayer Model. *Journal of the American Statistical Association* 56:841-868

Greville, T N E (1943) Short Methods of Constructing Abridged Life Tables. *Record of the American Institute of Actuaries* 32(Part I):29-42

Henry, L (1972) *Démographie. Analyse et Modèles*. Larousse

Keyfitz, N (1966) A Life Table that Agrees with the Data. *Journal of the American Statistical Association* 61:305-312

Keyfitz, N (1968a) A Life Table that Agrees with the Data, Part II. *Journal of the American Statistical Association* 63:1253-1268

Keyfitz, N (1968b) *Introduction to the Mathematics of Population*. Addison-Wesley Publishing Company

Keyfitz, N (1970) Finding Probabilities from Observed Rates or How to Make a Life Table. *The American Statistician* 24:28-33

Keyfitz, N (1977) *Introduction to the Mathematics of Population. With Revisions*. Addison-Wesley Publishing Company. (Revised edition of Keyfitz, 1968b)

Keyfitz, N and W Flieger (1971) *Population: Facts and Methods of Demography*. W H Freeman

Lederman, S (1969) *Nouvelles Tables-Types de Mortalité*. INED - Presses Universitaires de France

Nam, C B, W J Serow and D F Sly (eds) (1990) *International Handbook on Internal Migration*. Greenwood Press

NSO [National Statistical Office] (1972-1977) *1970 Population and Housing Census. 76 Volumes and 3 Subject Reports*. NSO, Bangkok

NSO [National Statistical Office] (1982) Personal Communication

Pressat, R (1983) *L'Analyse Démographique: Concepts, Méthodes, Résultats*. 4th Revised Edition. Presses Universitaires de France. (First edition published in 1961)

RAWG [Research and Analysis Working Group] (2003) Poverty and Human Development Report 2003. Dar es Salaam (RAWG operates within the framework of Tanzania's Poverty Reduction Strategy; the secretariat is provided by REPOA <http://www.repoa.or.tz/rawg/>)

Reed, L and M Merrell (1939) A Short Method for Constructing an Abridged Life Table. *American Journal of Hygiene* 30:33-62 [also reprinted in 1995 in *American Journal of Epidemiology* 141(11):993-1022]

Rees, P (1984) Does It Really Matter Which Migration Data You Use in a Population Model? Working Paper 383, School of Geography, University of Leeds

Rees, P and M Kupiszewski (1996) Internal Migration and Regional Population Dynamics: What Data are Available in the Council of Europe Member States? Working Paper 96/01, School of Geography, University of Leeds

Rees, P and A G Wilson (1977) *Spatial Population Analysis*. Edward Arnold

Rogers A (1973) Estimating Internal Migration from Incomplete Data Using Model Multiregional Life Tables. *Demography* 10(2):277-287

Rogers, A (1975) *Introduction to Multiregional Mathematical Demography*. John Wiley & Sons

Rogers, A (1982) Personal Communication, IIASA

Rogers, A (ed) (1999) The Indirect Estimation of Migration. *Mathematical Population Studies* 7(3):181-309 (Special Issue)

Rogers, A and L J Castro (1976) Model Multiregional Life Tables and Stable Populations. Research Report RR-76-9. IIASA

Rogers, A and L J Castro (1981) Model Migration Schedules. Research Report RR-81-30. IIASA

Schoen, R (1975) Constructing Increment-Decrement Life Tables. *Demography* 12(2):313-324

Schoen, R and S H Jonsson (2003) Estimating Multistate Transition Rates from Population Distributions. *Demographic Research* 9(1) <http://www.demographic-research.org/>

Smith, D P and N Keyfitz (1977) *Mathematical Demography: Selected Readings*. Springer-Verlag

UN-HABITAT [United Nations Human Settlements Programme] (2002) Cities without Slums. Paper prepared for the World Urban Forum, Nairobi, 29 April-3 May 2002. United Nations

UN-HABITAT [United Nations Human Settlements Programme] (2003) *The Challenge of Slums. Global Report on Human Settlements 2003*. Earthscan Publications

UNECE [United Nations Economic Commission for Europe] (2005) Report of the March 2005 Joint UNECE/EUROSTAT Seminar on Migration Statistics. United Nations

UNESCAP [Economic and Social Commission for Asia and the Pacific] (1982) *Migration, Urbanization, and Development in Thailand. Comparative Study on Migration, Urbanization, and Development in the ESCAP Region. Country Reports*. United Nations.

United Nations (1967) *Manual IV. Methods of Estimating Basic Demographic Measures from Incomplete Data*. United Nations

United Nations (1970) *Manual VI. Methods of Measuring Internal Migration*. United Nations

United Nations (1983) *Manual X. Indirect Techniques for Demographic Estimation*. United Nations

United Nations (1997) *Principles and Recommendations for Population and Housing Censuses, Revision 1*. United Nations

United Nations (1998) *Recommendations on Statistics of International Migration, Revision 1*. United Nations

United Nations (2002a) *International Migration Report 2002*. United Nations

United Nations (2002b) *Methods for Estimating Adult Mortality*. United Nations

United Nations (2004a) *World Urbanization Prospects: the 2003 Revision*. United Nations

United Nations (2004b) *Handbook on the Collection of Fertility and Mortality Data*. United Nations

Wanglee, Anuri [then, former Director, National Statistical Office, Thailand] (1982) Personal Communication, Bangkok

Willekens, F J (1980) Multistate Analysis: Tables of Working Life. *Environment and Planning A* 12:563-588

Willekens, F J (1999) Modeling Approaches to the Indirect Estimation of Migration Flows: From Entropy to EM. *Mathematical Population Studies* 7(3):239-278

Willekens, F J, I Shah, J M Shah and P Ramachandran (1982) Multi-state Analysis of Marital Status Life Tables: Theory and Application. *Population Studies* 36(1):129-144

Zlotnik, H (2002) Assessing Past Trends and Future Urbanization Prospects: The Limitations of Available Data. Paper prepared for the conference 'New Forms of Urbanization: Conceptualizing and Measuring Human Settlement in the Twenty-first Century', organized by the IUSSP Working Group on Urbanization and held at the Rockefeller Foundation's Study and Conference Center, Bellagio, Italy, 11-15 March 2002 (unpublished, available from IUSSP)