Mathematical Statistics
Stockholm University

Intercensal Population Estimates by Age and Sex

Gabriella Lundquist

Examensarbete 2006:10

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Examensarbete **2006:10**,
http://www.math.su.se/matstat

Intercensal Population Estimates by Age and Sex

Gabriella Lundquist[*]

May 2006

Abstract

Population censuses are conducted regularly in most countries in order to determine the size and structure of the population. This data is used when planning the right locations for schools, roads, and hospitals; identifying trends over time that can help predict future needs; distribution of funds for government programs etc.

As the cost of conducting a census is so high, estimates are calculated for the intercensal years based upon the number of births, deaths and migrations in that population. Often a break can be observed in the population time series for the census years, due to the uncertainty in these rates. This is referred to as the error of closure.

This thesis discusses the causes of the error of closure and investigates several methods for performing population estimates in an effort to reduce the error. The recommendation is that population estimates should be analysed and, if possible, recalculated, before an appropriate method is used to eliminate the error of closure. The method proposed is an adaptation of Denton's Quadratic Minimization.

[*]Postal address: Dept of Mathematical Statistics, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: gabriella.lundquist@gmail.com. Supervisor: Anders Björkström.

Foreword

This report is written as part of an internship with the United Nations Economic Commission for Europe (UNECE), situated in Geneva, Switzerland. The UNECE has several databases with country specific data. As gaps may appear in the population time series, a deeper review of the causes and possible solutions was required. This report has focused on that problem in particular, and has resulted in a Master's Thesis at the University of Stockholm, Sweden.

I would like to thank the entire team of the Demographic Section of the Statistical Division of the UNECE for making my internship a great experience in every way. Extra thanks go to my supervisor Paolo Valente, for his great support and positive attitude. I would also like to thank my supervisor Anders Björkström at the University of Stockholm for his valuable thoughts and comments.

Gabriella Lundquist
April 2006

Glossary

Crude birth rate	Total annual number of childbirths per 1000 people.
Error of closure	The difference between the census result and the postcensal estimate. When using the term as a percentage, the result is divided by the census result.
Intercensal estimate	Estimate of the population between two censuses based upon the two census results.
Intercensal period	Period in between two censuses.
Postcensal estimate	Estimates for the period followed a census, based on the previous census results.
Vital statistics	Statistics on births and deaths.

Contents

Chapter 1

Introduction

1.1 Census

Most countries conduct a regular census; typically every ten years. The purpose is to collect information about a population and to find its size and structure. A census covers many different areas such as education, health, family and property. The implementation of and questions asked in a census differs from country to country, but there are international recommendations and guidelines published in order to make international comparisons possible[1].

In many developed countries a census is carried out with the help of questionnaires, and the goal is to include the entire population. Naturally there will always be people who do not send in the census form for many reasons, but in each census round an effort is made to obtain information about all of the people living in that census area. In the 2000 census round in the United States, more than 500.000 people were employed by the U.S Census Bureau just to visit homes from which census forms were missing, in order to increase the accuracy of the census. This gives us an idea of what enormous undertaking a census is.

There are other ways to count and gather information about a population though. Finland is an example of a country which carries out a census only with the help of registers. In Sweden a census hasn't been carried out using questionnaires since 1990.

A census is important for many reasons. It helps in planning the right locations for schools, roads, and hospitals; identifying trends over time that can help predict future needs; distribution of funds for government programs etc. It is also the most accurate source of the size of a population, which is an important variable in many applications.

As a census is such a large and expensive project, the population is estimated for the intercensal period. This is possible using vital statistics and rates on international migration, in order to move the population forward in time and predict its size and composition the following years after a census. These population estimates will be referred to as the *postcensal estimates* throughout the text.

1.2 Problem Description

For every non-census year the population is updated from vital statistics to obtain the estimated population size for the current year. The vital statistics and the assumptions made about migrations can naturally be incorrect, which will lead to population estimates that do not reflect the real size and structure. The error will then get worse with time, since each estimate of the population is an update of the size from the year before. When a new census is conducted the population size obtained can be significantly higher or lower compared to the postcensal estimate for the year before. Consequently there will be a break in the population time series, which is referred to as the *error of closure*. This is a problem since it reflects an error in the postcensal estimates. The break will also cause problems for economic applications in particular, since many of them requires time series that are smooth.

In a large majority of cases the error of closure is caused by unrecorded migration. Most European countries have reliable vital statistics, but in some countries international migration is hard to measure because of war, boundary changes or other events during the intercensal period that makes people move in and out of the country unrecorded. Most countries choose to recalculate the population estimates after a new census is taken, in order to produce more accurate population figures. There are many different ways to do so, and different methods produce different results.

Many international organizations who collect country specific data on population are also faced with this problem. Due to the large number of countries in their databases, the possibilities to recalculate the population estimates are limited.

1.3 Objective

The objective of this thesis is to analyse different methods to recalculate postcensal population estimates by age and sex. The aim is to analyse the problem from an international organization's point of view as well as a general one, and to analyse the different estimates available today.

1.4 Delimitations

The analysis is dependent on the availability of country specific data on population estimates as well as census results. Data from national statistical offices has been used where available, for certain countries other sources have been used.

Albania is used as an example throughout the thesis, but it is important to point out that this is not a thesis about the Albanian population. There are two reasons to why Albania is a good example. First, the country experienced great demographic changes due to high and unrecorded migration between the two censuses of 1989 and 2001. Second, it is one of the few countries with an error of closure of more than |10%|, that has both postcensal and intercensal estimates available from a reliable source.

1.5 Disposition

The thesis starts with a discussion in Chapter 2 about the reasons behind an error of closure, in order to understand why the problem occurs. This chapter also gives an understanding for the reality behind the numbers, and the fact that population estimates and the demographic changes in a country should follow the same trend. Chapter 3 demonstrates the impact of using data from different sources. The aim is to demonstrate the need for good estimates, and the problem of having different sets of estimates in frequently used databases. Chapter 4 presents three different ways to recalculate intercensal estimates. The methods are described and the necessary data is listed, in order to understand the available options. Chapter 5 analyses the methods more in detail. In Chapter 6 the conclusions and recommendations for the future are presented.

Chapter 2

The Reasons Behind an Error of Closure

In this chapter we are going to look at the error of closure that occured in Albania followed the 2001 census. The reason to why the Albanian population was so hard to estimate for the intercensal period of 1989 to 2001, has its explaination in the mass migration that occured at the time. We will therefore start this chapter with an overview of the political and demographic situation during the intercensal period.

2.1 The Migration in Albania 1989-2001

During the intercensal period of the 1989 and 2001 censuses Albania experienced one of the greater emigrations of recent times. 600 000 to 700 000 Albanians[2] left the country and changed the demographic landscape. Little research has been done in the area, and since a major part of the migration was unrecorded the data available is poor. In 2003 the Sussex Centre for Migration Research published a report[3] on the subject, in order to make an overview of the migration that took place during the 1990's. In 2004 the Albanian Institute of Statistics (INSTAT) released a similar research report, where the migration that took place between the same period was analysed and defined[2]. The numbers in the two reports differ slightly at some points, but overall they paint a similar picture of the migration in Albania 1989 to 2001.

The origins of this mass migrations begin with the leader of the Albanian Party of Labour, Envar Hoxha, who ruled the country from 1941 until his death in 1985. Hoxha turned Albania into one of the most isolated communist regimes in the world. Emigration was strictly forbidden, punishable by death or prison. An electric fence and sentry points ran the length of the Greek and Yugoslavian borders, to ensure that population of Albania remained within its borders.

In 1990 5000 Albanians invaded western embassies in Tirana to seek asylum. After this restrictions were relaxed and passports began to become available. This was the first step to the great migrations that were about to come. By the end of the same year approximately 20 000 Albanians emigrated to Greece. In March 1991, just before the first democratic elections, 25 000 Albanians headed towards the Italian coast by boats. The Italian government decided to accept them, but when 18 000 new Albanians arrived a few months later most of them were sent back, since a democratic election had taken place and they could not be accepted as political refugees[2].

In March 1992 a new election was held and the Democrats, led by Sali Berisha, came to the power. Within three years, 1991-1993, 2-300 000 Albanians emigrated to neighbouring countries. A result of the collapse of the communist or state-socialist regimes in Eastern Europe, as well as the political situation within the country. This turbulent period was followed by some years of economic progress in 1993-1996, with some emigrants even returning to their home country. Even with this promising development the unemployment rate was still at 20%, and by 1995 20% of the working population had emigrated[3].

The 1996 elections, and the collapse of a pyramid investment schemes that around half of all Albanians had invested in, resulted in another phase of mass emigration. In six days in 1997, 10 600 Albanians crossed the Adriatic Sea to seek refugee in Italy. During 1999 half a million Kosovo Albanian refugees arrived to the northern part of Albania creating instability, but the years since 2000 have been dominated by peace and political quiet, with the Socialists by the power[3].

The demographic situation in Albania made it difficult to estimate the population for the years after the 1989 census was taken. Consequently an error of closure of -11.1%[1] occurred when a new census was conducted in 2001 [2]. Looking at the postcensal estimates, see Figure 2.1, the population increased by 3,3% between 1989 and 1990. This is relatively high if we compare the number with the average growth rate in Europe at the time of approximately 0.53% [4], but is explained by the high fertility rate of 3.1 children per women. According to the same postcensal estimates the

[1] A negative error of closure means that the population was over estimated, meaning the census result was less than the postcensal estimate.

[2] No postcensal estimates for 2001 were available, which is why the growth rate of the postcensal estimates of 1999 and 2000 has been used to calculate an approximate postcensal estimate for the census year.

population decreased by about 120 000 people in 1990-1993, followed by a steady increase in 1994-2000 with an average growth rate of 1.0%.

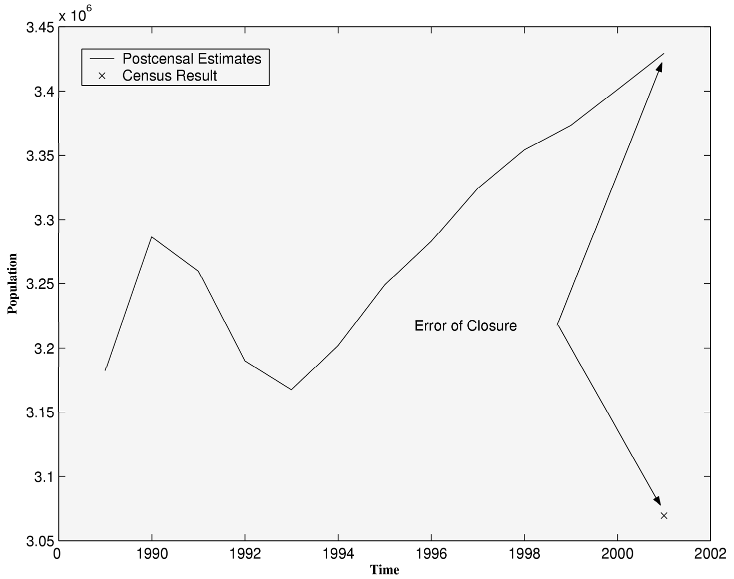

Figure 2.1: Postcensal Estimates of the Albanian Population

The situation in Albania is a good example of the reasons behind an error of closure, and shows in particular why some countries have more problems in predicting its population than others. This example illustrates how important it is to look behind the numbers, and to realize that the postcensal estimates and the real situation in a country can differ significantly.

There are many similar examples where a turbulent political situation and a high flow of migration makes it difficult to estimate the population in a country. Armenia had an error of closure of less than -18% following the 2001 census, and Georgia experienced one of less than -12%. Compare these to countries with lower migration rates and more stable economies. Both the United States and the United Kingdom experienced an error of closure of more than 2% in the census round of 2000.

For Albania the main reason for the error of closure in 2001 was the high and unrecorded migration that occurred throughout the intercensal period. All countries experience an error of closure of some kind after a census is taken, since estimating a population is hard even with accurate rates on birth, death and migration. In countries where the migration flow can be well recorded and occurs steadily over time, the error of closure is therefore caused by an uncertainty in the assumptions more balanced between all three factors. The postcensal estimates can then be considered as the best guess in the growth trend, but must be adjusted to fit the census results.

Chapter 3

Population Data Provided by Different Sources

Country specific data such as population is available from a wide variety of sources. Many international organizations provide an online database free of charge, and most national statistical offices present population numbers as well as rates on births and deaths on their websites. In this chapter we are going to look more closely at the differences in the data from different sources and what we can infer from the data.

3.1 Albania

Albania will once again serve as an example, since we are now familiar with the demographic background. This will help us when analysing the population estimates provided by different sources.

3.1.1 Intercensal Estimates produced by INSTAT

INSTAT produced a series of postcensal estimates after the census of 1989. These numbers were adjusted after the census of 2001, in order to correct the error of closure. In Figure 3.1 the estimates are shown together with a curve illustrating the Albanian population with the assumptions of no migration[1].

As discussed in the previous chapter, the demographic change that took place in Albania were one of the greatest in recent times. The recalculated

[1] The estimates are calculated with the 1989 census as a base, using data on natural increase obtained from INSTAT [6]. No information was available on the natural increase in 1989. However, since the natural increase was almost constant in the intercensal period, the average rate for 1990-2000 has been used when calculating the population for 1990.

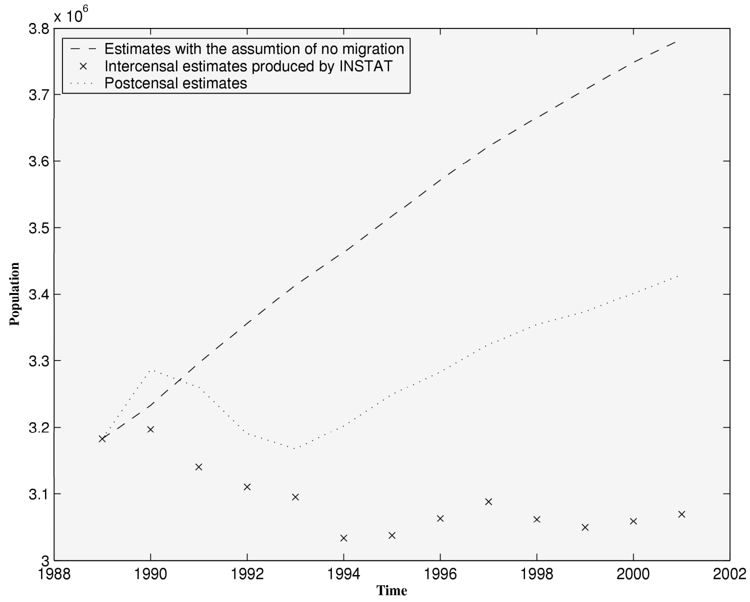

Figure 3.1: Population Estimates for Albania

intercensal estimates should therefore give an idea of the assumptions made about migration, when subtracting them from the estimates with the assumption of no migration. This is illustrated in Figure 3.2, where we can see the yearly migration flow according to this reasoning.

A total of 710 000 people emigrated during the intercensal period of 1989-2001. We have to keep in mind that this number is including the number of live births by the female migrants during the intercensal period and that the number of people that actually emigrated from Albania is less. The average crude birth rate during 1989-2001 was 21,7 [6], which gives, when applied to the yearly migration flow, a number of 15 500 Albanian children born outside of Albania between the two censuses. Considering that a majority of the female emigrants were in the fertile age of 18-32 [3], this number is probably higher. The total number of (710 000 − number of children born by emigrants outside of Albania) migrants could therefore match the number of 600 000 to 700 000 as discussed about in the previous chapter. Figure 3.2 also show a similar migration trend throughout the period, and the recalculated intercensal estimates from INSTAT seem to give an appropriate picture of the real situation in Albania during 1989-2001.

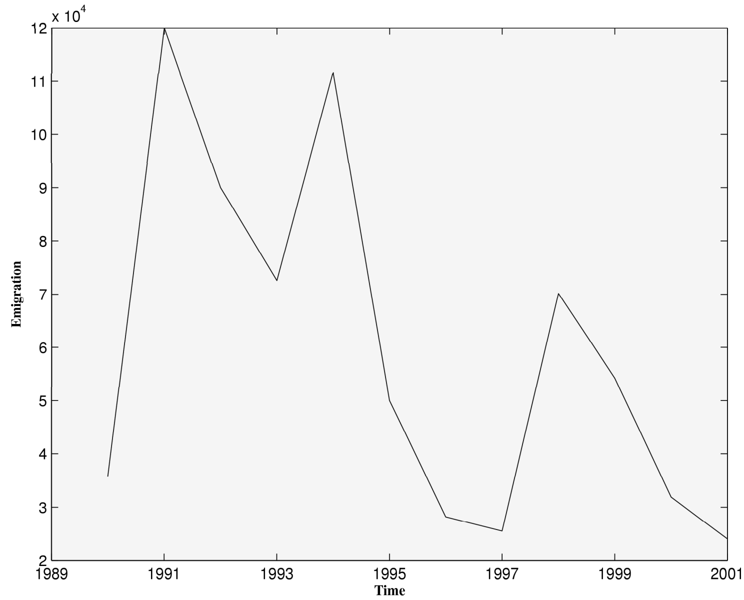

Figure 3.2: Emigration in Albania According to the Intercensal Estimates Produced by INSTAT

3.1.2 The United Nations

The United Nations Economic Commission for Europe (UNECE) and the United Nations Population Division (UNPD) are two examples of organizations within the United Nations (UN) that provide population data online. The figures differ though, and at the moment these two sources provide three different sets of series on population, as seen in Figure 3.3[2]. The UNPD produce their own estimates, whereas the UNECE upload postcensal estimates from the national statistical offices. The economic section of the UNECE sometimes recalculates the figures, or match them with available intercensal estimates, when a significant error of closure occurs, in order to produce as smooth a time series as possible for application reasons. In the case of Albania they have a similar trend as the intercensal estimates produced by INSTAT.

The estimates from the UNPD look different from both the postcensal and the intercensal estimates. To find out what assumptions are made about migration, we follow the same procedure as for Figure 3.2[3]. Results are shown

[2] The UNECE is aware of this problem, and a discussion has been held about what population estimates to use.

[3] This means making the assumption that the UNPD has used the same rates on natural increase as available from INSTAT, which might not be the case.

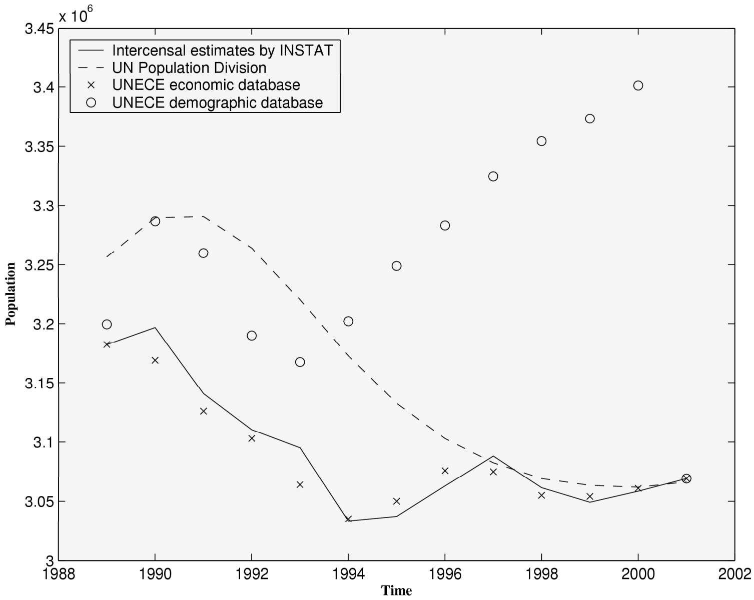

Figure 3.3: Population in Albania According to Different Sources

in Figure 3.4. The total migration adds up to approximately 790 000 for the intercensal period and about 250 000 people emigrated 1991-1993 according to the figure. The conclusion is that the migration flow is over estimated, and show a different trend to the one observed in Figure 3.2.

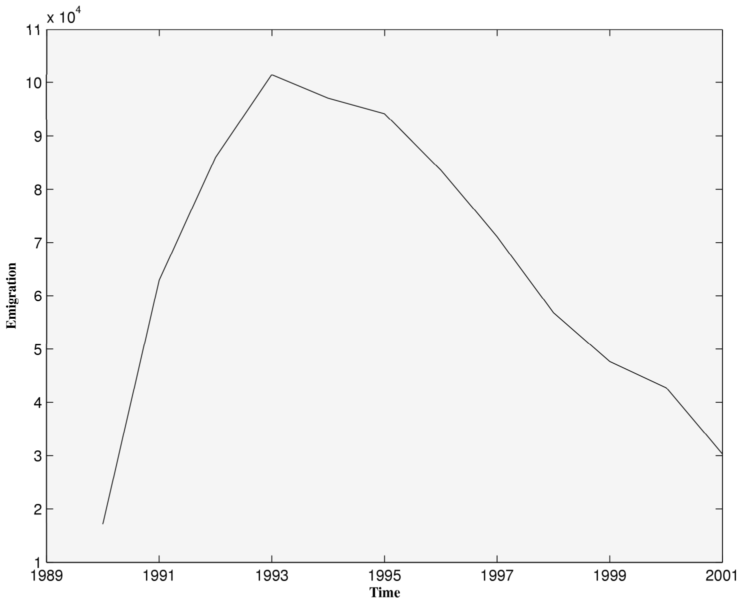

Figure 3.4: Emigration in Albania According to the United Nations Population Division

Chapter 4

Recalculation Methods

There are many options when recalculating intercensal estimates. In this chapter the different methods are presented, in order to understand the idea behind the intercensal estimates.

4.1 Linear and Exponential Interpolation

The easiest way to calculate population estimates in between two censuses is to use the census results, while making assumptions about how the population grow during the intercensal period.

Data required:

- Population at national level at the time of the first census.
- Population at national level at the time of the second census.

4.1.1 Linear Interpolation

This is a method based on the assumption that the growth rate of a population is linear. The estimates for the intercensal period are given by the formula[7]:

$$P_s = kP_i + (1-k)P_j$$

where

$$P_s,\ P_i \text{ and } P_j \;=\; \text{population at time } i,\ s \text{ and } j$$

$$\begin{aligned} s &= \text{time for the interpolation, where } i < s < j \\ k &= \frac{j-s}{j-i}, \text{ a constant for each year s} \end{aligned}$$

4.1.2 Exponential Interpolation

This is based on the same idea as above, but with the assumption that population grows exponentially not linearly. The formula used is[7]:

$$P_s = P_i e^{hr}$$

where

$$\begin{aligned} P_s,\ P_i \text{ and } P_j &= \text{population at time } i,\ s \text{ and } j \\ s &= \text{time for the interpolation, where } i < s < j \\ h &= s - i \\ r &= \frac{ln(P_j/P_i)}{j-i} \end{aligned}$$

4.2 Methods Based on Postcensal Estimates

Another approach when producing intercensal estimates is to keep the yearly fluctuations from the postcensal estimates, but to fit the new curve to the second census result. Several methods are based upon on this idea and all produce similar results. All of the methods base the intercensal estimates on the trend obtained from the postcensal estimates. These are calculated with consideration to the first census results, birth, death and migration rates, but in such a way as to ensure there is no error of closure. The intercensal estimates will therefore form a curve with a different slope, while keeping the fluctuations from year to year almost as they were.

Data required:

- Population at national level at the time of the first census.
- Population at national level at the time of the second census.
- Postcensal estimates for the intercensal period.
- Postcensal estimate at the time of the second census date.

4.2.1 Method Used by the U.S Census Bureau

The U.S Census Bureau[8] uses this method to recalculates the postcensal estimates after a new census. The following formula is used:[1]

$$P_t = Q_t(\frac{P_i}{Q_i})^{(t/i)}$$

where

$$\begin{aligned} P_t &= \text{intercensal population estimate at time } t \\ P_i &= \text{census result for the second census at time } i \\ Q_t &= \text{postcensal estimate at time } t \\ i &= \text{years in between the two censuses} \\ t &= \text{time in years elapsed since the first census} \end{aligned}$$

4.2.2 Method based on Denton's Quadratic Minimization

A problem often faced when preparing economic time series, is to adjust high frequency data to make them accord with low frequency ones. Annual values for a specific indicator might be available from one source, while monthly or quarterly values are obtained from another. It is therefore desirable that the monthly or quarterly values add up to the annual totals, while at the same time trying not to change the trend obtained from the high frequency data. A possible method to use which respects these restraints, is an adaptation of Denton's Quadratic Minimization[9][10].

A similar situation occurs when the total population needs to be re-estimated for the intercensal period, where the aim is to keep the yearly fluctuations from the postcensal estimates. We have a problem where high frequency data need to be adjusted to low frequency ones, while at the same time following the trend of a curve already obtained. An adaptation of Denton's Quadratic Minimization could therefore be used to calculate intercensal population estimates.

Let $\widetilde{Q}_t$ be the series of postcensal estimates and $\widetilde{P}_t$ be the series of the recalculated population estimates we are looking for. Also assume that the first census was conducted in year $t = 1$ and the second census in year $t = n$.

[1] The United States produces estimates for each month and not just per year. Since this thesis discuss the recalculation of postcensal estimates per *year*, the formula is adjusted to fit yearly calculations rather than monthly.

Since we want $\tilde{P}_t$ to have a growth rate as similar to the one of $\tilde{Q}_t$ as possible, we want to minimize the expression:

$$\sum_{t=1}^{n} \Big(\frac{\tilde{P}_t}{\tilde{Q}_t} - \frac{\tilde{P}_{t-1}}{\tilde{Q}_{t-1}}\Big)^2 \tag{4.1}$$

If we let

$$\mathbf{M} = \begin{pmatrix} \frac{1}{\tilde{Q}_1} & 0 & \dots & 0 & 0 \\ 0 & \frac{1}{\tilde{Q}_2} & \dots & 0 & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & \dots & \frac{1}{\tilde{Q}_{n-1}} & 0 \\ 0 & 0 & \dots & 0 & \frac{1}{\tilde{Q}_n} \end{pmatrix} , \quad \mathbf{D} = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & \dots & -1 & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & -1 & 1 \end{pmatrix}$$

and $\tilde{\mathbf{P}}' = \begin{pmatrix} \tilde{P}_1 & \tilde{P}_2 & \dots & \tilde{P}_n \end{pmatrix}$, we can write (4.1) as:

$$\sum_{t=1}^{n} \Big(\frac{\tilde{P}_t}{\tilde{Q}_t} - \frac{\tilde{P}_{t-1}}{\tilde{Q}_{t-1}}\Big)^2 = \big(DM\tilde{P}\big)' \big(DM\tilde{P}\big) = \tilde{P}'M'D'DM\tilde{P}$$

At the same time we want to make sure that $\tilde{P}_1 = C_1$ and $\tilde{P}_n = C_n$, where $C_i =$ census result at year i. This is to eliminate the error of closure. In matrix notation the constraint can be expressed as $B\tilde{P}_t = C_c$ where

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{C_c'} = \begin{pmatrix} C_1 & C_n \end{pmatrix}$$

The Lagrangian to be minimized is then $L = \tilde{P}'M'D'DM\tilde{P} + \lambda'(B\tilde{P} - C_c)$ which gives us the solution:

$$\begin{pmatrix} \tilde{P} \\ \lambda \end{pmatrix} = \begin{pmatrix} 2M'D'DM & B' \\ B & 0 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ C_c \end{pmatrix}$$

where the first n values gives us the series $\widetilde{P_t}$ we are looking for.

4.3 A Demographic Analysis

Poulain[11] presents a methodology which uses all available data in order to produce as good intercensal population estimates as possible. Before starting the recalculation process the reliability of the census data and the vital statistics should be examined and, if possible, corrected in an appropriate way.

- If the census date and the date for which the postcensal estimates are calculated is not the same, the census results have to be adjusted using vital statistics.
- One group of people might have been treated differently in the two censuses because of changed census rules.
- A coverage survey might be available, so that census errors can be corrected.
- It is important that both censuses and vital statistics relate to the same population. If this is not the case, the group missing in one data source should be eliminated from the other data sources as well.
- Throughout the recalculation process data is considered by year of birth rather than age. If data is only available by age, it should be transformed into year of birth using an appropriate method.

Data required:

- Birth, death, and migration rates for the period between the census date and the date used for the postcensal estimates
- Population by sex and year of birth at national level at the time of the first census
- Population by sex and year of birth at national level at the time of the second census
- Postcensal estimates by sex and year of birth at the time of the second census date
- Yearly international migration data by sex and year of birth where both flows are measured

  or

- A standard migration schedule

4.3.1 Step 1

The residuals[2] are calculated for each group by sex and year of age. The relative differences are also calculated and are defined as the absolute residuals divided by corresponding census numbers.

4.3.2 Step 2

In this step the residuals are examined in detail. One particular sex or birth cohort might have a large residual compared to the other groups and reasons for this have to be identified together with an appropriate solution. Examples of problems found in this step are unrecorded emigration, short-term migrations, refugee flows and unrecorded births and deaths etc. In these cases the residuals will be negative. Positive residuals naturally refers to unrecorded immigration. In the following text only negative residuals are considered but the theory is equally valid for positive residuals.

By looking at the relative differences we can discover errors caused by migration. If the relative differences are not randomly distributed around 0%, but seem to have a trend around a negative percentage, that might be an indicator for unrecorded emigration. The idea is then to add as many emigrations as needed to cover the residuals. In step 3a the method proposed is based on international emigration data, while in step 3b we can proceed without this information but with the help of a standard migration schedule. Residuals can also occur because of errors in age reporting in censuses and vital statistics. This is dealt with in step 4.

4.3.3 Step 3a

If the undercoverage is considered to be constant over time, all the observed annual numbers of emigration for a given age and birth cohort will be increased by the factor $\frac{EMI_i/RES_i}{EMI_i}$, where

EMI_i = total number of observed emigrations during the intercensal period for sex and birth cohort number i

RES_i = number of emigrants to be added corresponding to residual valu-

[2] The words *residual* and *relative difference* will be used in this section rather than error of closure, since we are talking about the error within each age and sex group as well as for the entire population.

es for age and sex group number i

If the undercoverage cannot be considered constant, an under coverage factor should be estimated in order to calculate a non-linear distribution of the residuals by years. Often countries have better records on immigration than emigration so it is often possible to compare emmigration from one country with immigration in another.

COV_j (undercoverage factor in year j) = total emigration to destination or total immigration in to destination in year j

COV_j should be counted for all years j in the intercensal period. If it is not uniform, all sex and birth cohorts should be treated separately.

The estimated number of unrecorded emigrations for sex and birth cohort i at time j is given by:

$$EST_{i,j} = \frac{EMI_i(1 - COV_j)}{COV_j}$$

$\sum_j EST_{i,j}$ will then be the total number of estimated unrecorded emigrations for each sex and birth cohort for the intercensal period. The distribution per year can then be applied to each residual by direct proportional transformation[3]. Each postcensus estimate will then be corrected by adding the calculated number of unrecorded emigrations.

4.3.4 Step 3b

Since migration is usually responsible for residuals with a negative trend, it is important not to oversee this even though data on international migration is missing. In this step a standard migration schedule is proposed, in order to distribute the residuals during the intercensal period for each sex and birth cohort. The idea is that there are certain trends in how different age groups tend to emigrate, and from a standard behaviour different weights can be calculated.

[3] If unrecorded emigration is A for year i, and B for the entire intercensal period, the distribution found would be $\frac{A}{B}$ for year i. Applying this to the residual means multiplying $\frac{A}{B}$ by the total residual, which will give us the correct number of unrecorded emigrations for year i.

$w_{i,j}$ = weight for the distribution of residuals for sex and birth cohort i in year j

RES_i = residual for sex and birth cohort i

This time $EST_{i,j}$, the estimated number of unrecorded emigrations for sex and birth cohort i at time j, is given by:

$$EST_{i,j} = RES_i \cdot w_{i,j}$$

This gives us a first distribution of residuals. The problem with this distribution is that it gives yearly fluctuations that are more or less the same, which might not be the case. We can look at the "total immigration in to destination" to see if the distribution is equal by years or not. To solve this problem a second distribution of residuals is calculated by using the "total immigration in to destination" as the total each year. The numbers are then transformed proportionally from the first distribution of residuals. We then have the right distribution for each sex and birth cohorts together with yearly fluctuations that correspond to the real observed immigration numbers. The last step is to adjust the numbers to fit the total residuals for each sex and birth cohort. This is also done by proportional transformation which gives us the final distribution of residuals.

This method is called the "bi-proportional iterative method" and adjusts the distribution by year as well as for age.

4.4 Age and Sex Distribution

Population numbers are often separated into groups by age and sex. A specific group might not follow the same trend over time as the entire population. It is therefore important to find the right age and sex distribution even for the intercensal period. The reason why it is more common with groups by age and sex, compare to groups by cohorts and sex, is that many applications depend on the trend of a specific age group. In the planning of future schools, the distribution of pensions etc, it is more common to talk about a specific age group and not about the number of people born at a certain time. It is important to understand the difference though. A cohort will decrease with time, while an age group will increase. This is because a cohort looses people as time goes by, whereas an age group will increase along with the growth of the entire population. Most statistical institutions publish population numbers by sex and five year age groups, or by sex and one group for each age.

The challenge when calculating population estimates for a population by age and sex is to find the correct age and sex distribution. Since all groups might not develop according to the same trend, they naturally have to be treated differently. The method chosen is therefore applied to each group separately, which will produce estimates for each group throughout the intercensal period. When adding the estimates by year, we will notice that they do not equal the numbers we get when applying the method to the entire population, as seen in Figure 4.1[4].

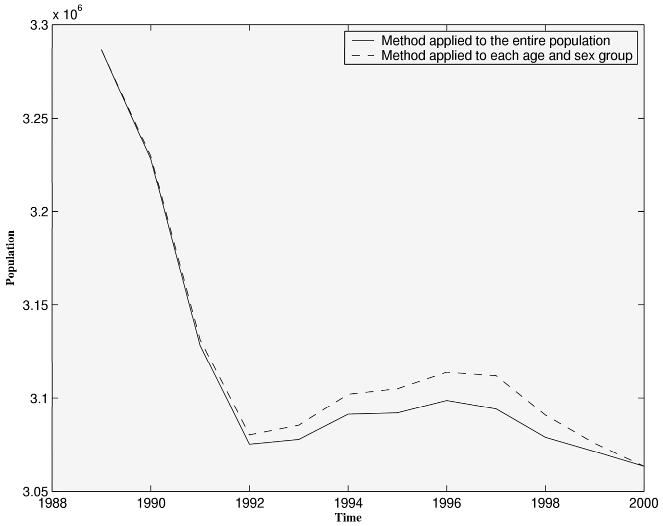

Figure 4.1: The Population of Albania - Different Results

Therefore we apply the distribution found on the total population, which will give us intercensal estimates that add up to the right total, while at the same time follow the desired trend even when looking at the different groups separately.[11]

[4] The method used to illustrate the different results, is the adaptation of Denton's Quadratic Minimization

Chapter 5

Analysis of Recalculation Methods

Many methods are based on the idea of comparing the estimated population with the one obtained from a census. The availability of a postcensal estimate for the census year is therefore important. Before moving on to the analysis of the different methods, we will therefore have a discussion about this problem in particular.

5.1 Missing Postcensal Estimate for the Census Year

In order to calculate a set of intercensal estimates there is often the need for a postcensal estimate of the census year. It is also essential when calculating the error of closure since the idea is to compare the size and structure of the actual population being compared to the estimated one. The numbers therefore need to refer to the same year. Most online databases do not provide this kind of information, simply because the organizations receive postcensal estimates for the intercensal period, not including the census years. The economic section of the UNECE therefore uses the growth rate from the two years before a census is taken, and uses this to create an approximate postcensal estimate for the census year. The problem of a missing postcensal estimate is handled in this way throughout the thesis. In this section we are going to look closer on how much this approximation differs from a real postcensal estimate, in order to get a measure of how much information we loose in the forthcoming analyse.

In 2000 the population of the United States was 281 421 906 according to the census. The postcensal estimate for the same year was 274 608 346, which gives an error of closure of 2,42%[8]. When creating an estimate based on the postcensal growth rate from the previous two years, we get an error of closure of 2,43%. The difference in between the two estimates is $< 0.01\%$, which could justify using an approximate postcensal estimate in the analysis when no other data is available. Of course using other countries as an example could give different results. No postcensal estimate for the census year of 2001 is available for Albania, but instead we can make the same analysis for the years before. We can then see that the differences range from 0,2% to a maximum of 4,1%, when calculating new estimates for the entire intercensal period[1]. The percentage of error is higher, but since it is below 1,7% for all years but the first, we will carry out the analysis using approximate postcensal estimates when necessary.

5.2 The Intercensal Estimates

Figure 5.1 shows the results from the different recalculation methods. The error of closure of -11,1%, or 359 900 people, is distributed between the intercensal years, and each method has its own way of calculating this. The interesting aspect is which migration patterns are shown according to which method. When doing the same migration analysis as in the previous chapters, we can see in Figure 5.3 that the trend differs significantly between the methods. Surprisingly all methods have the same total of 714 000 emigrants during the intercensal period. This gives a realistic picture of the situation in Albania 1989-2001, according to the discussion in Chapter 2. Again we have to have in mind that the total number of migrants include the number of live births by the female migrants during the intercensal period. According to the analysis in Chapter 3, approximately 15 500 children were born by Albanian emigrants, when applying the available average crude birth rate for the intercensal period to the total number of migrants. The number could be much higher though, since the age and sex composition of the migrants does not reflect the composition of the Albanian population as a whole. Overall the total number of migrants according to the recalculation methods match the previous analysis of 600 000 to 700 000 emigrated people during 1989-2001.

The integral between the curves of the population according to the different methods, and the curve illustrating the population with the assumption of no migration, equals the total number of migrants for each method. The

[1] By this we mean starting with the first and second postcensal estimate, in order to create a third one. This estimate is compared to the original third postcensal estimate and so on.

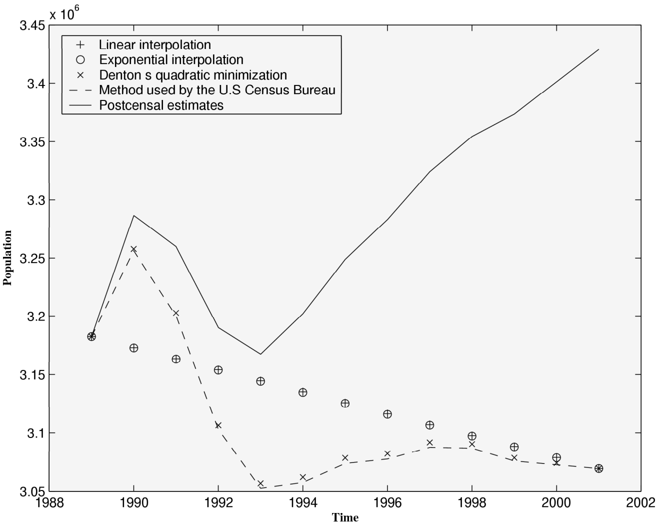

Figure 5.1: Albania - Results of Different Recalculation Methods

phenomenon that all methods has the same total number of migrants during the intercensal period, can be explained by the fact that they are benchmarked to the same numbers, in this case the results of the 1989 and 2001 Albanian censuses, see Figure 5.2. The integral will therefore be the same, no matter how much the population trends differ throughout the intercensal period.

As seen in Figure 5.3 the migration trend differs depending on the method used. Linear and exponential interpolation show an almost constant trend for the period, while both the adaptation of Denton's Quadratic Minimization and the method used by the U.S Census Bureau show a trend more similar to Figure 3.2.

5.2.1 Adaptation of Denton's Quadratic Minimization vs the Method Used by the U.S Census Bureau

The adaptation of Denton's Quadratic Minimization and the method used by the U.S Census Bureau give almost identical results. In Figure 5.4 we can see the difference between the two set of estimates for Albania.

The difference is marginal compared to the size of the population, 0,15% at the most in 1996[2], but as seen in the figure the relation between the two methods is a squared function. That means each estimate based on Denton's Quadratic Minimization could be written as:

[2] $\frac{4725}{3248836} = 0.0015$

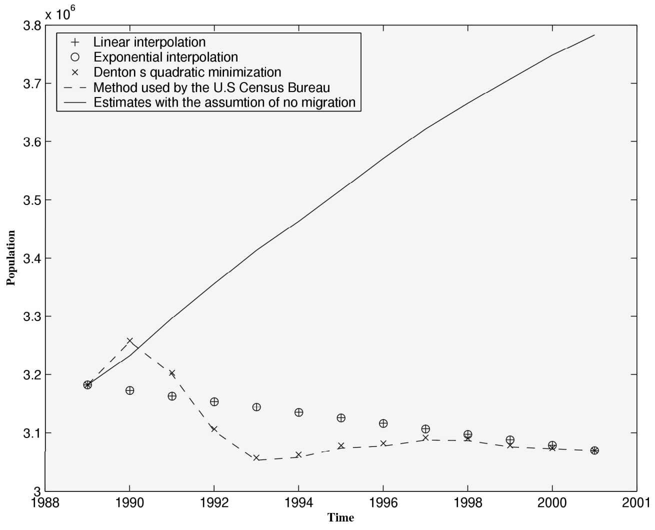

Figure 5.2: Albania - Comparing Different Methods with the Assumption of No Migration

$$P_t = Q_t(\frac{P_i}{Q_i})^{(t/i)} + f(i-t) \tag{5.1}$$

for some function $f(x) = ax^2 + bx + c$.

When looking more closely at the Albanian case, the function can be written as:[3]

$$f(x) = -131,3x^2 + 1837,4x - 1706,1 \tag{5.2}$$

When doing the same analysis for Armenia, we can see that the difference between the method based on Denton's Quadratic Minimization and the method used by the U.S Census Bureau shows the same relation, as seen in Figure 5.5. The function can now be written as:

$$f(x) = -346.3x^2 + 4849.4x - 4503.1 \tag{5.3}$$

From equation 5.2 and 5.3, we can see that a general form can be written as:

$$f(x) = -\frac{A}{36}(x^2 - 14x + 13) \tag{5.4}$$

[3] The equation is an approximation of $f(x)$, since the number of significant digits is one. In the equation the first census year is referred to as year number 1 and so on, and not by the actual year.

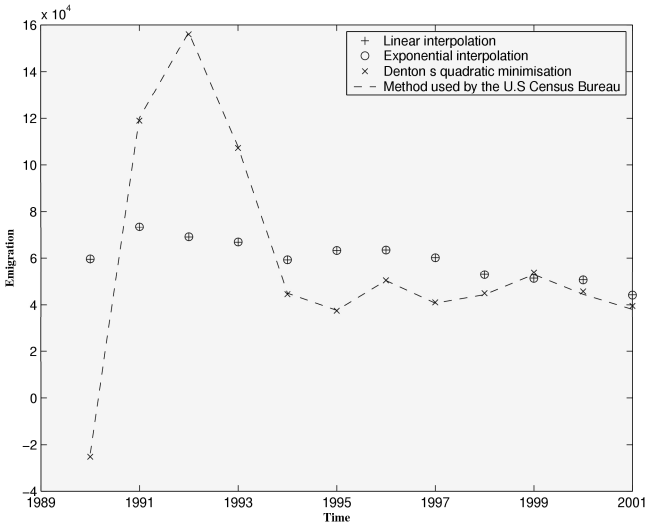

Figure 5.3: Albania - Emigration According to the Different Methods

where A is a country specific constant equal the midterm value for the difference between the method based on Denton's Quadratic Minimization and the method used by the U.S Census Bureau. In this case $A_{Albania} = 4722.8$ and $A_{Armenia} = 12470.0$. Both Albania and Armenia conducted a census in 1989 and 2001. Since equation 5.4 has the restriction of an intercensal period equal to the ones for Albania and Armenia, we will look at a country with other census dates. The United States conducted a census in 1990 and 2000, which gives a different intercensal period. When doing the same analyse on the postcensal estimates from the United States[4], we get the general function:

$$f(x) = -\frac{B}{25}(x^2 - 12x + 11) \tag{5.5}$$

where $B_{usa} = 19912.0$, see Figure 5.6.

It is interesting to see that the relation between the method based on Denton's Quadratic Minimization and the method used by the U.S Census Bureau is a squared function in the *first* quadrant for both positive and negative error of closures. This, and the results from the three examples of Albania, Armenia and the United States, leads us to think that there might be a general expression for the differences between the two methods that is valid for all possible intercensal periods. Based on the same reasoning as

[4] The U.S Census Bureau only provides intercensal estimates for 1990-2000, why equation 4.2.1 has been used to obtain postcensal estimates.

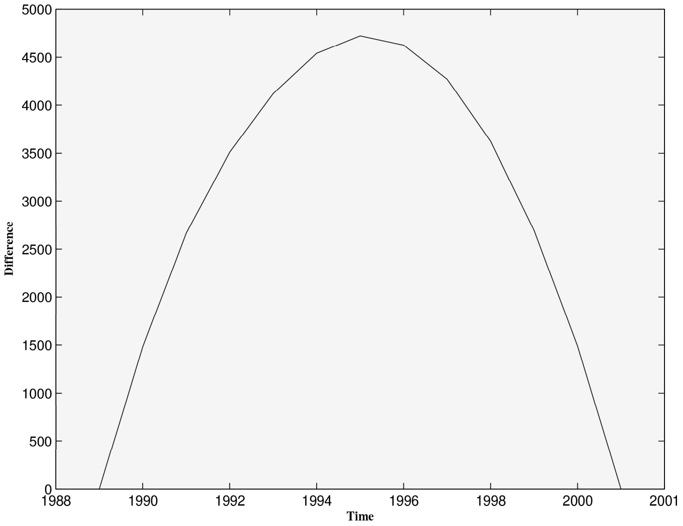

Figure 5.4: Albania - Difference Between the Adaptation of Denton's Quadratic Minimization and the Method Used by the U.S Census Bureau

above, the general formula would be:

$$f(x) = -\frac{4C}{(n-1)^2}(x^2 + (-1-n)x + n) \tag{5.6}$$

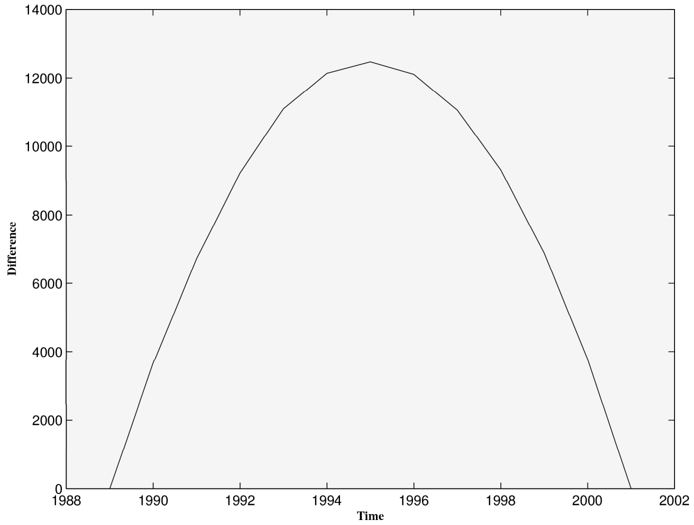

Figure 5.5: Armenia - Difference Between the Adaptation of Denton's Quadratic Minimization and the Method Used by the U.S Census Bureau

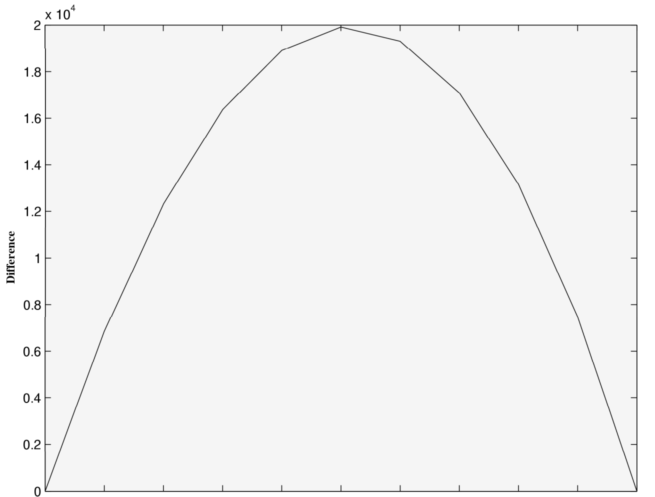

Figure 5.6: United States - Difference Between the Adaptation of Denton's Quadratic Minimization and the Method Used by the U.S Census Bureau

Chapter 6

Conclusion

6.1 Recommendations

All countries will experience an error of closure at some level after a census is taken, since postcensal estimates are projections about a population and not real numbers. Since the reasons behind an error of closure differs between countries, there is naturally not one identical solution to recommend in all cases. As seen in the example of Albania, unrecorded migration was the number one factor that caused an error of closure of -11.1%. When calculating intercensal estimates for Albania using the different methods, we obtained estimates that gave an incorrect picture of the migration flow in the past.

However, the problem is not whether or not a country has experienced a mass migration during the intercensal period or not, the problem we should focus on is how accurate the postcensal estimates are. My suggestion for all countries and cases, is to start with an analysis of the postcensal estimates. This analysis should include a comparison between the estimates and the real situation, and an investigation on what new data on vital statistics and migration is available for the intercensal period. This is in order to determine if better postcensal estimates can be calculated. If it is decided that the postcensal estimates does not reflect a trustworthy situation in the country, a second set of postcensal estimates should be calculated based on new rates on vital statistics and migration and/or a demographic analyse like the one presented in Chapter 4.3.

Since an error of closure will occur even for the second set of postcensal estimates, they should be used in a method that keeps the yearly fluctuations as close as possible, but fits the curve to the two census results. As we cannot justify creating estimates with a new trend, the curve obtained from the first or second set of postcensal estimates is considered the best attempt. According to the analysis in Chapter 5, the best method to use would be the adaptation of Denton's Quadratic Minimization. In theory these estimates work best with two benchmark restrictions, while at the same time wanting to minimize the difference between the yearly fluctuations for the postcensal and the intercensal estimates (see expression 4.1).

The method used by the U.S Census Bureau is based on the same idea and produces similar results. However, when comparing the two it appears that better estimates could be produced by the method based on Denton's Quadratic Minimization, with a difference according to $f(x)$ in equation 5.6. A more careful analysis and a mathematical proof are required though, in order to state which of the two methods that produce the most accurate estimates according to the underlying theory.

Linear and exponential interpolation cannot be recommended in any case, as the estimates are only based on two census results and they ignore that demographic changes occurs during the intercensal years.

My recommendations for international organizations in particular, are to encourage the national statistical offices to calculate intercensal estimates after a census is taken. As discussed in the thesis, population estimates can differ depending on the source, and it should be in everyone's interest to provide country specific data as accurately as possible. I would not recommend for individual institutions or organizations to apply a recalculation method to the postcensal estimates directly without further analysis. That could result in population estimates as seen in Figure 5.3, which is not desirable in any database.

In general I would like to encourage better cooperation between the national statistical offices and the organizations that provide country specific data. My wish is to find the same set of estimates for a country, independent on what source is used. Today the UNPD produce their own estimates, while the UNECE has estimates from the national statistical offices in their databases. It is of course in the UN's best interests that the data provided by its different organization is consistent. It also shows a lack of communication when UNPD publish population estimates that differs from the ones provided by the member states.

6.2 Suggestions for Further Research

It is important to bear in mind the reasons why population estimates are produced. Facts about the size and structure of a population are used in a variety of applications and for decision making. During the thesis work it became clear how much population estimates can differ depending on the source. They will also differ depending on what method is used when calculating intercensal estimates.

My suggestion for future research in the area of population estimates is a more in depth review of the relation between the method used by the U.S Census Bureau and the method based on Denton's Quadratic Minimization.

I would also like to suggest an investigation on what the long-term effects will be when using different estimates for a population. BNP per capita is one example of a quantity based on population estimates that is frequently used in a variety of applications and decisions worldwide. What difference does using different population estimates have in these calculations? Will decisions further down in the chain have a different outcome, depending on what source or method is used? What sources are used by decision makers today?

An investigation like this is outside the scope and time available for this thesis. It is therefore my hope that more research will be done in the area, since population estimates can be part of decisions that in the end affects the daily lives of people.

Bibliography

[1] *Recommendations for the 2000 Censuses of Population and Housing in the ECE Region.* United Nations Economic Commission for Europe, 1998

[2] E. Galanxhi, E. Misja, D. Lameborshi, M. Lerch, P. Wanner, J. Dahinden. *Migration in Albania.* INSTAT, 2004

[3] R. King, J. Vullnetari. *Migration and Development in Albania.* Sussex Centre for Migration Research, 2003

[4] *Analysis of Urban and Rural Population Growth at Regional Level.* United Nations Population Division, Table 20, 2001

[5] The Government of Albania. *National Report*, United Nations Economic Commissions for Europe, Government of Hungary, United Nations Population Fund, 2, 1998

[6] www.instat.gov.al

[7] E. Arriaga. *Population Analysis with Microcomputers.* Bureau of the Census, USAID, UNFPA, Volume 1, 1994

[8] www.census.gov/popest/archives/methodology/intercensal_nat_meth.html

[9] D. Rhoades. *Interpolating Annual Estimates of Purchasing Power Parity Between Tri-annual Benchmarks.* United Nations Economic Commission for Europe, 4-5, 2003

[10] F. Denton. *Adjustment of Monthly or Quartely Series to Annual Totals: An Approach Based on Quadratic Minimization.* Journal of the American Statistical Association, Volume 66, 1971

[11] A. Herm, M. Poulain. *Basic Methodology for the Recalculation of Intercensal Population Estimates.* European Commission, 2003