2011
2012

BEDRIJFSECONOMISCHE WETENSCHAPPEN

master in de verkeerskunde: mobiliteitsmanagement
(Interfacultaire opleiding)

Masterproef

Refining synthetic population generation: development of a migration data model

Promotor :
Prof.dr.ir Tom BELLEMANS

Copromotor :
Prof. dr. Mario COOLS

Thomas Rayen

Masterproef voorgedragen tot het bekomen van de graad van master in de verkeerskunde , afstudeerrichting mobiliteitsmanagement

universiteit hasselt
KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt

universiteit hasselt
KNOWLEDGE IN ACTION

2011
2012

BEDRIJFSECONOMISCHE WETENSCHAPPEN

master in de verkeerskunde: mobiliteitsmanagement
(Interfacultaire opleiding)

Masterproef

Refining synthetic population generation: development of a migration data model

Promotor :
Prof.dr.ir Tom BELLEMANS

Copromotor :
Prof. dr. Mario COOLS

Thomas Rayen

Masterproef voorgedragen tot het bekomen van de graad van master in de verkeerskunde , afstudeerrichting mobiliteitsmanagement

universiteit hasselt
KNOWLEDGE IN ACTION

Refining Synthetic Population Generation: Development of a Migration Data Model

Thomas Rayen*
Prof. dr. Ir. Tom Bellemans
Prof. dr. Mario Cools
Katrien Declercq

Transportation Research Institute (IMOB)
Hasselt University
Wetenschapspark 5, bus 6
B-3590 Diepenbeek
Belgium

Fax: +32 (0)11 26 91 11
Tel.: +32 (0)11 26 91 99

E-mail:
Thomas.Rayen@student.uhasselt.be
Tom.Bellemans@uhasselt.be
Mario.Cools@uhasselt.be
Katrien.Declercq@uhasselt.be

* Corresponding author

Number of words = 9.365
Number of Tables = 0
Number of Figures = 12
Words counted: + 12 * 250 = 12.365
Paper submitted:
June 1, 2012

ABSTRACT
This study deals with the concept of synthetic data populations as input for traffic modeling purposes. The term ‘synthetic’ applies to a combination of data sets, often collected for other research purposes. Over the years, they have been growing in popularity in the field of transportation because they form an alternative to the extensive data collection process. A synthetic population for Belgium was developed containing socio-demographic variables of all inhabitants for the year 2001. These variables can be used to model transportation demand because socio-demographic characteristics influence travel behavior. There are 5 main components included in this model: birth and death processes, the forming of new couples, employment, drivers’ license and income. From a demographic perspective, it’s essential that migration is added to this equation. Three migration probabilities were calculated based on data of the Flemish Government: internal migration, immigration and emigration. A random number generator in SAS was used to assign these probabilities. As a result, a correction mechanism was necessary because migration probabilities were calculated on a person-level, resulting in problems on household-level. The analysis and results indicated that a population growth was found. However, due to the random nature of the approach, areas such as the coastal region and city centers suffered large population losses. Also, a shift in the age distributions was noted resulting in a rejuvenated population. These findings force us to conclude that a person-level approach is perhaps not the best way to adopt migration in a population prediction model.

Key words: Synthetic Data Population, Synthetic Data Generation, Migration

1. INTRODUCTION

Demography is a science that studies population structures using statistics and mathematical concepts. Some examples of demography include the births and deaths that have occurred in a certain time span and the household size development of a region. Changes in the population structure are of great importance to policy makers and governments. In Belgium, for example, there has been a continuous trend of an increasing population and it is likely to continue at least until 2060 [1]. This trend does not only occur in Belgium. Many sociologists have confirmed a worldwide exponential growth of populations [2] [3]. The cause(s) for this explosion can be found in the roots of the industrial revolution at the end of the 18th century. These positively influenced social standards and living conditions. Demographic trends have impacts in numerous sectors: health care, economy, but for this study, the field of transportation is of most importance. Congestions result in economic losses and are harmful for the environment. Transportation demand is partly influenced by socioeconomic and demographic characteristics of individuals so it is essential to estimate any future developments in this domain [4] [5] [6].

Synthetic data populations can play an important role in modeling studies. They are often the result of a combination of datasets that were previously collected for other research purposes. Nakamya et al. [7] of 'the Institute of Mobility' (Imob) has already constructed a synthetic dataset for Belgium based on the SocioEconomic Survey (SES) of 2001. Also, data from the 'Onderzoek Verplaatsingsgedrag' (OVG) or national household travel survey was merged to result in a cleaned dataset including variables such as the drivers' license rate, income, etc. Synthetic datasets are more often used in modeling processes because they form a good alternative to intensive data collection for large studies. Another reason for their growing popularity is the fact that we see an international shifting trend from classic four-step models to activity-based models [8] [9]. These models attempt to predict the transport demand based on how individuals' activities are spread in time and space.

This study aims to improve the synthetic data population of Belgium by adding migration to the current model in use. We take the coastal region of this province as an example to explain the need of this study. One of the 10 provinces in Belgium is West-Flanders. Its coastal region consists of 10 municipalities. Data analysis shows us that the average age of the citizens in the segment '< 18' is 2% less than the Flanders region and the province of West-Flanders as a whole (Figure 1).

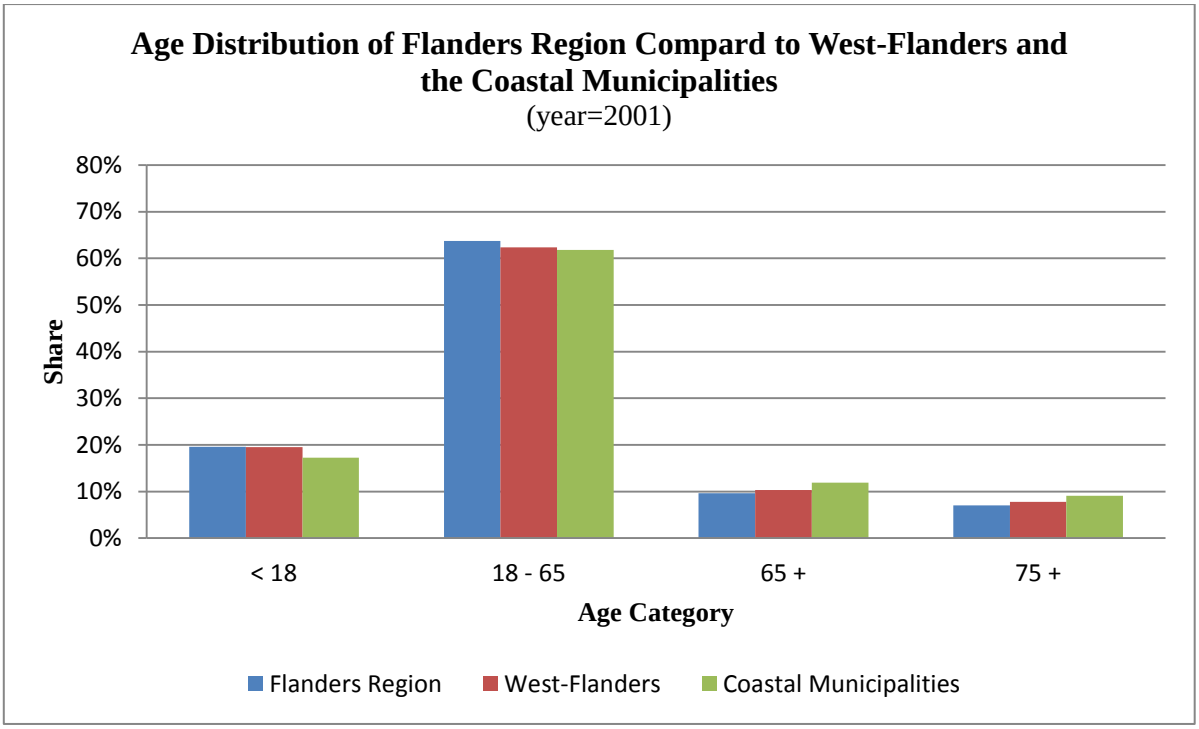

FIGURE 1 Age distribution of Flanders region, West-Flanders and the coastal municipalities according to the base-year population (2001).

For the 'not working' population (65 + and 75 +), we note a small discrepancy. The coastal region municipalities show a different distribution than the province and the region of Flanders. On a regional level, nine out of ten municipalities having the eldest population are located near the coast [10]. According to De Nauw [11] and Lodewijckx [12], the ageing of the population will continue to take place in the future. Figure 1 indicates that there is a surplus of elderly in the coastal region, combined with a smaller share of non-adults (<18). Suppose there were no migration streams active in the coastal region. This would influence the population distribution when modeling long term periods in the future. Due to the surplus of elderly and a deficit of newborns, a most likely outcome will be the loss of the elder segment of the population (75 +). However, this phase is followed by a decrease of the death rate, resulting in a small population with a lot of young people. We can therefore assume that a new population balance will occur in the future.

The absence of migration streams is not a realistic scenario because internal migration in Belgium is taking place in large cities and also in the coastal region [13]. The importance of migration is also stressed in global literature. It plays a large role for the growth of cities, therefore creating a link with traffic conditions in urban areas [14]. In general, internal migration is considered as a key determinant for changes in the population distribution [15]. Also, estimating migration has been a theme of growing interest to policy makers since the 1990's. A problem remains that migration is still poorly registered in many countries. Different criteria are often used, making it therefore not easy to process the scarce migration data [15] [16].

2. PROBLEM DEFINITION

Nakamya et al. [7] has constructed a synthetic dataset for 2001. We can make a distinction between the main components of this model:

1. Birth and death processes
2. Forming of new couples
3. Employment
4. Income
5. Drivers' license rate

From a demographic perspective, it immediately becomes clear that an essential component is missing: migration [16] [17]. As we stated in the introduction, there are several problems with the current method in practice. It was found that, when modeling the population long term periods in the future, the coastal population suffered large losses. Because this region is populated with many elderly (Figure 1), a population deficit will be generated since there are too few births and many deaths. By implementing migration in our synthetic dataset, we aim to contribute to a better reflection of the real population. A natural consequence is the improved prediction strength of the model.

Other drawbacks of the model include the forming of new couples. Aging a population for several years will improve the likelihood of a changed marital state so this is an aspect that must be taken into account. Once the modeling process is conducted, persons in the age class of 24 to 32 years that are single and living in the same municipality, are simply coupled to another single individual of a different sex, based on a random number. These assumptions are rudimentary and result in the absence of gay couples for example. It is therefore not realistic. There are also modules for the employment rate, the driver license and income. The (3) employment rate is based on similar assumptions meaning that the likelihood of someone to change age-class will also influence the chances of employment. The (4) income module depends on socioeconomic status. It calculates and predicts the income of the different individuals of the households via a multinomial logit model. Finally, (5) driver's license rate is assumed to be constant over time: once an individual has passed his or her driver's test, it is assumed that they will not lose their driver's license in the future. By adding a sixth component to this model, we hope to resolve the problems at the coastal region in particular.

3. OBJECTIVES & RESEARCH QUESTIONS

This study aims to contribute to the relationship of demographic attributes and mobility characteristics. A synthetic population of high quality opens doors for other fields of study. It can be used as input for Feathers, which is an activity-based model used to estimate transportation demand in Flanders [18]. Because of this, we can formulate the main goal as upgrading the synthetic population with a migration module so it becomes of higher quality. To achieve the main research goal, four research questions were developed:

1. Which data sources are available and provide significant information for the development of population prediction models?
2. What does the current model look like and what are its drawbacks?
3. How can we calculate migration and which model provides the best fit?
4. Given the use of an updated synthetic data population, to what extent can we detect an influence of the migration component?

The first research question calls for a detailed review of existing data sources. The Imob has already developed a synthetic data set based on the SES of 2001. They combined it with data from the OVG to create a cleaned database for base-year 2001. However, in order to correctly integrate the migration module, more data are needed. These data are provided by the Flemish government and will be briefly discussed. This is done in Section 5.1.

The second question reflects the current model in use and aims to describe it in summary. This way, problems and drawbacks are exposed, opening the field for recommendations and future fine-tuning of the model. The biggest problem of all is the absence of migration, but there are also other issues such as the rudimentary state of the 'new couples' module for example. More of this can be found in the problem statement. Question number three explores the options for the calculation of future migration rates. Future migration numbers are calculated using trend extrapolations. This method was applied by Siegel and Swanson [16] to estimate population sizes in the future. Nevertheless, the mathematical concept remains the same. Three types of extrapolations are used: linear, geometrical and exponential. Linear regression is used to select the method that provides the best fit. Section 5.2. will be dealing with this matter. The final research question, and perhaps the most important one, was set up to critically reflect on the used methodology. The output can be used for further analysis to see which trends occur in the future. This will be extensively discussed in Section 6 by means of several analyses that focus on a selection of demographic variables, e.g. age and population size.

4. LITERATURE STUDY

This section of the study provides general information of demographic modeling concepts, an insight in migration types and a brief summary of applications of synthetic datasets in the context of transportation. Demography deals with concepts that are not known to traffic modelers and thus it is interesting to set out some important findings that can be used for the purpose of this study. Also, a review of papers is given in which synthetic data populations have successfully been developed in the past.

4.1. Fundamental Equation of Demography

The fundamental equation is often considered as one of the most important equations within the field of demography [16] [17]. The equation has the following structure:

$$P_t - P_0 = B - D + I - O$$

P_t is the population at the end of a given time period, while P_0 is a population at the beginning of a time period. Terms *B* and *D* represent the number of births and deaths respectively. *I* and *O* stand for the immigration and emigration components. Put otherwise: This formula is a description of the net difference between the number of births and deaths and the net difference between immigration and emigration. Therefore, terms *B* and *D* form the 'natural' component while the final two terms are stated as the 'net migration' component. Siegel and Swanson [16] interpret this equation as an 'inflow-outflow relationship' in which several conditions must be met to guarantee correct results. These conditions imply a geographically defined area that does not deal with measure errors. When this is not the case, a final term ε must be added. It is a representation of the residual error term. This equation allows us to transfer terms to the other side:

$$P_t = P_0 + B - D + I - O + \varepsilon$$

We are now able to predict the population at time *t*. There are, however, difficulties that go alongside the use of this equation. Siegel and Swanson [16] indicate that it is difficult to calculate the migration component. The 'Federaal PlanBureau' (FPB), an important governmental agency in Belgium also states that international migration is the most difficult component to estimate. Migration is based on several aspects, labor being one, and therefore making it not easy to estimate it. Also, countries sometimes apply different migration criteria, making it therefore difficult for comparison [16]. Nevertheless, migration is an important aspect that cannot be overlooked. In 2005, nearly 3 % of the global population left their county. This corresponds with a number of 191 million people [19].
Another important element to point out is the use of terms such as 'base-year' and 'launch-year'. For this study, the base-year of the migration data is 1990. The launch-year is that of 2009.

From this year on, projections are made until 2020. More detailed information about the migration data can be found in Section 5 and Appendices I, II and III (p. 26 – 28).

4.2. Migration

The Flemish Government adopts a set of definitions to correctly distinct between migration types. Table 3 is a summary table with migration definitions of the 'Federal Governmental Agency of Economy' (F.O.D. Economie) and is found in Appendix I (p. 26).

The most important distinction is that of internal and external migration. Internal migration can be defined by individuals moving across municipal boundaries in their country while external migration refers to people crossing national boundaries. The latter type can be divided in two more categories: immigration and emigration. Immigrants are people who enter a country while emigrants decide to leave and move abroad. Robila [20] found two major factors influencing immigration streams in the United States: (1) education attainment and (2) language use. Immigrants are more likely to find a job when they have an educational degree. Young children are better able to adopt and learn new languages. Since Belgium is a bi/tri-lingual country with high profile education, this can encourage immigrants to move to Belgium. Pelfrene [1] also states that international migration is the most difficult component to estimate in the future. Because of the complexity, authors often apply the approach of a 'constant migration balance' or a 'constant evolution of the migration balance' using several assumptions. The latter approach is applied in this study.

An important migration condition can be found with respect to age. Willaert [21] researched the migration profiles of Belgian individuals with relation to spatial areas and age groups. Willaert's findings indicate a peak in the social mobility for young adults (Figure 2).

FIGURE 2 Social mobility per 1000 inhabitants with relation to age according to Willaert [21].

The vertical axis represents the social mobility which is expressed per 1000 inhabitants. 'Age' is found on the horizontal axis ranging from 0 to 100 +. Immediately we can see that individuals in the age category from 18 to 28 are very mobile. The peak value can be found at the age of 28 and implies that 1.3 % of the 28-year olds are very mobile. Overall, it is found that children do not migrate, except for those in the age class of 0 to 4. This phenomenon is called 'associated migration' [21]. The infants are forced to migrate with the other members of the household, their parents in particular.

On the other hand, parents are less likely to migrate when their children are going to school and have established a social network. From the age of 70, we see an increase in the social mobility. This has to do with elderly people moving in retirement homes or with their families due to health problems.

4.3. Synthetic Data Populations

From a transport-related point of view, creating a synthetic dataset is an alternative data collection method that is often applied to build activity-based models. Activity-based models are used to estimate the demand for transport based on how activities of individuals are distributed in time and space. According to Davidson et al. [8], an activity-based model therefore requires three building blocks:

1. An activity-based platform
2. A tour-based structure
3. Microsimulation modeling techniques

The first aspect refers to the use of activities as a starting point. When creating an activity-based model, researchers must step away from traditional modeling techniques and look at the behavioral component of transportation. The best way to do this is to collect and describe information about ones activities. This information could be gathered by using a daily log with a GPS tracker for example. By calculating the number of tours, and not the number of trips, modelers can obtain more information because this approach results in a consistent trip scheme that is part of a tour. A trip-based structure is often considered as simplistic and does not describe the behavior of an individual well [22]. The third element, microsimulation, is important when modeling on a disaggregated or an agent-based level. Microsimulation has a lot of advantages. Modelers are able to take the variability of the travel demand into account instead of applying average values [8]. In order to simulate the behavior of individual agents, we therefore need a dataset that represents these agents. Thanks to synthetic data populations, modelers can achieve datasets that can act as input for traffic models, given the condition that they are of high quality.

4.3.1. Review of Case Studies

Building a synthetic data population is a process where a combination of various data sources is needed. A technique that is often applied is the ‘Iterative Proportional Fitting process’ (IPF) and was described in detail by Beckmann et al. [23]. The IPF-technique exists of a proportional fitting algorithm that estimates the households in a zone, with respect to a certain list of demographic attributes: age, gender, income, employment, family class and race. In order to do so, they used census data of the ‘Public Use Microdata Samples’ (PUMS), which represents a sample of 5 % of the total census data file that corresponds with merely 100.000 individuals.

The first step of the algorithm is to create summary tables with demographic attributes. Once this is done, a multiway table is estimated to generate the number of households per cell. This is done by multiplying the total number of households and their probabilities or by generating a random number to assign probabilities [24]. The method, however, is not waterproof. There are several problems that arise when using the IPF technique: (1) the zero-cell value problem and (2) a problem with the distribution of attributes [25]. This implies the correction of round-off errors. In his study, Beckmann et al. [23] encountered a matrix with a total of 11.760 cells of which 11.151 were empty meaning that no value was assigned to these cells. These cells would be assigned a value of ‘0’ during the IPF-algorithm resulting in a failure to converge. In order to resolve this problem, Beckmann et al. [23] proposed a method called ‘tweaking’. This means that the empty cells were assigned a value of 0.1 or 0.01 before the IPF-algorithm was conducted. Guo and Bhat [25] recently proposed another method to solve this zero-cell value problem. One solution is to terminate the IPF-algorithm when a maximum number of iterations are reached. Another method is designed to reduce the number of incorrect zero-cell values by defining the variable class intervals [25].

The authors suggest a strict definition of the variable class intervals implying that a more aggregate classification (e.g. a 6-way classification of a household type) is preferred above a less aggregate classification (e.g. a 12-way classification) because it will reduce the number of empty cells.

It is necessary that a trade-off is taken into account between the accuracy of the IPF-technique and the level of detail of the population. Coming back to back to the second problem, Guo and Bhat [25] argue that it is not possible to control the algorithm for household and individual level at the same time. Some attributes are defined on an individual level (e.g. the distribution of gender) while others are formulated on household level (e.g. household size). This problem is resolved by formulating a new algorithm that deals with these kinds of situations. Even though Beckmann's method has been the subject of many discussions, it is often applied in practice [24]. Other methods within the field of generating synthetic populations include that of Mohammadian et al. [26].

Mohammadian et al. [26] define control variables for the synthetic population. These variables are: household size, household income, number of vehicles in the household, number of workers in the household, the presence of children and the age of the householder. After defining these variables, the IPF-method is applied to generate a joint distribution between the household proportions and the demographic attributes. Once this joint distribution is known, weights are assigned to create a fully synthesized population. The next step implies the generation of travel household attributes. The authors therefore apply an Artificial Neural Network (ANN) model that identifies homogenous groups (clusters) in the synthetic population. Then, for all the travel attributes that are taken into account, the best-fitted distribution is applied for new clusters [26]. The authors suggest two updating methods for generating a good-as-possible population: (1) an expert's opinion and (2) the Bayesian updating technique.

In order to use the first one, the authors have to make sure that the average values of the travel attributes are calculated so that they could update these mean values during the modeling process [26]. Bayesian updating implies taking the effects of local attributes into account that were overlooked in earlier stages of the modeling process. The technique is applied to update the best-fitted distributions in earlier stages of the process. This way, researchers can make a distinction between prior-and posterior distributions.

5. METHODOLOGY

In order to effectively test the migration model in SAS, a sample (n=2359) was drawn from the population using simple random sampling. Once the model was completed and tested, SAS ran the script using the entire population (N=10.296.350) for the base-year (2001). This was done in combination with the module of birth and death processes.

5.1. Data

Several data sources were used to calculate the correct migration probabilities. Table 2 from Appendix II (p. 27) provides a summary description of the data files that were used for constructing the synthetic dataset. Most important for this study are the migration numbers from the Flemish Government (Appendix III, p. 28). The data range from 1990 until 2009 and describe the migration totals per year. The SES contains information of all Belgian individuals. It was conducted in 2001 and was set up as a survey for all persons and households. The main goal was to obtain relevant statistical information for other governments (local, regional and federal) and scientific research purposes [27]. The OVG was the third of its kind. It was conducted in 1996, 2001 and 2008. In 2013, a new report will appear with relevant travel information [28]. Variables such as the distribution of the driver's license rates and the number of vehicles per family are, for example, derived from data of the OVG. For upgrading the synthetic population, an extra data set containing migration numbers of the Belgian population is used. Based on these numbers, predictions are made for future migration numbers (to 2020). This way, we can calculate the probability of an individual to migrate in the future. Once this is done, we can assign the migration probabilities in SAS using a RNG.

5.2. Determining Migration Probabilities

A first step was to assign the migration probabilities. In order to correctly do this, data from the Flemish Government was used. Migration totals from 1990 to 2009 were extrapolated to 2020 in three different ways using a linear, geometric and an exponential model. Linear regression was used to determine the determination coefficient (R^2) for each method and for all types of migration. The linear projection method resulted in the best fit (Figure 3):

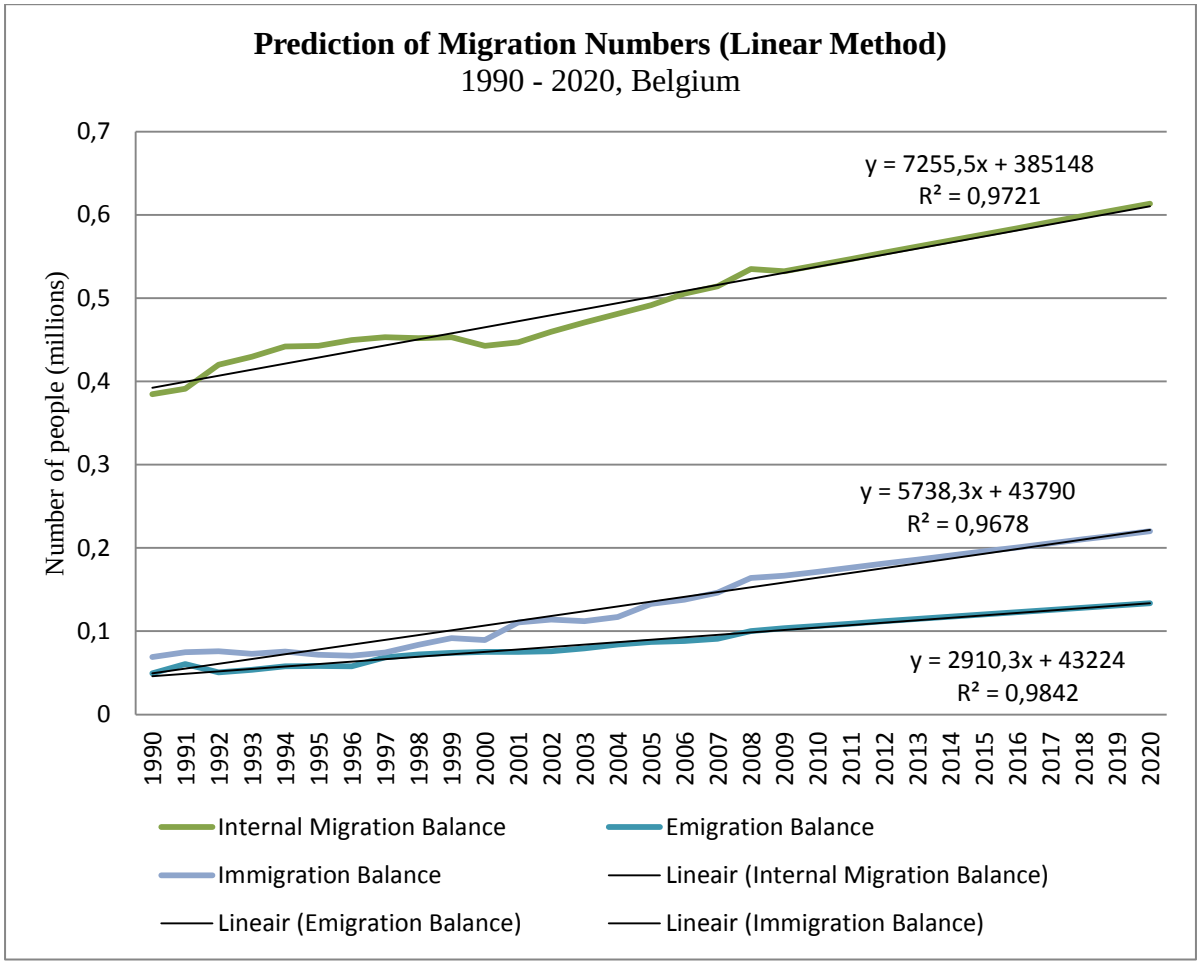

FIGURE 3 Extrapolation of migration data using a linear method.

All three migration types were extrapolated: internal migration, international out-migration (emigration) and international in-migration (immigration). For each of the three types, R^2-values of more than 96 % were noted, marking the linear approach as the best method. This was to be expected due to the linear nature of the data. In general, we can see an increase of the social mobility for all three types. Internal migration has the largest share and implies individuals moving across municipal borders while the other two reflect streams of people that cross national boundaries. In 2020, an approximate total of 613.451 individuals will move across municipal boundaries. The number of immigrants will rise to 220.132 persons that year and 133.678 people will eventually leave the country in 2020. For detailed numbers per year, we refer to Appendix IV (p.29).

After having calculated the total number of migrants by using trend extrapolations, we can now calculate the migration probability per person based on population data of the FPB [29]. We divide the migration numbers by the totals of the population predictions that were made by the FPB. This way, we obtain the probability per individual per year to migrate.

This approach implies that we are surrendered to the assumptions made by the FPB by using their population numbers. With this information, migration probabilities for every individual were calculated, resulting in Figure 4:

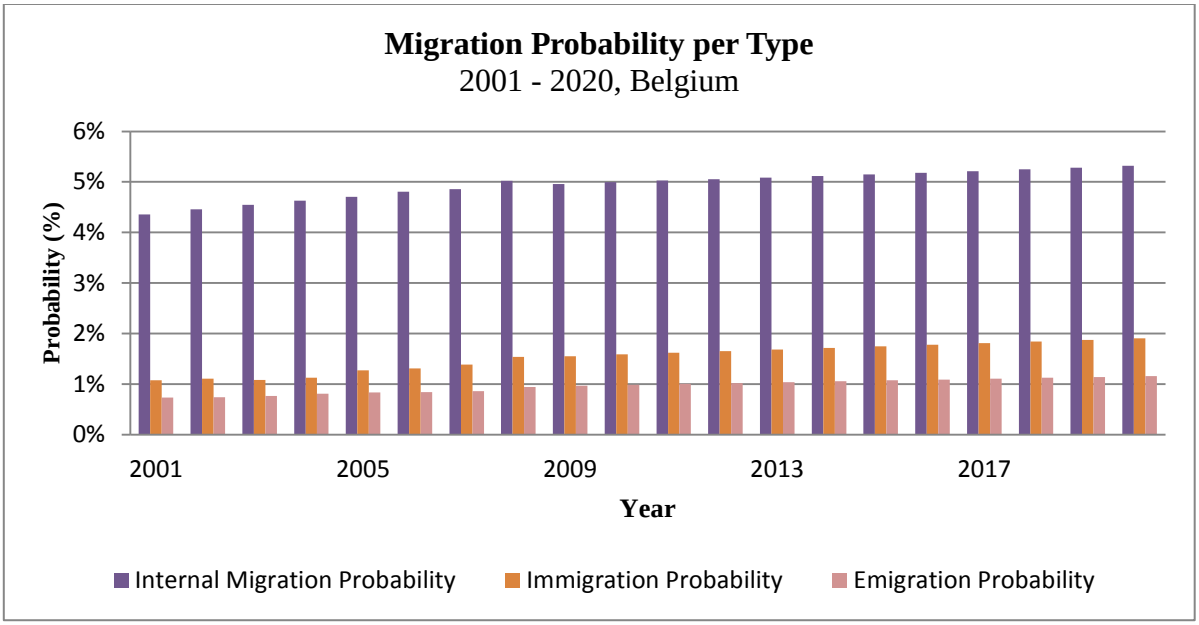

FIGURE 4 Migration probabilities per person per year.

Figure 4 represents the migration probability for each individual of the Belgian population from 2001 to 2020. The internal migration probability increases from 4.36 % in 2001 to 5.32 % in 2020. The other two types, immigration and emigration, increase from 1.08 % to 1.91 % and from 0.73 % to 1.16 % respectively. This implies that the probability of an individual entering Belgium is larger than that of leaving the country. A stabilizing trend for all migration types is noted. This can be explained by a stagnation of the population growth. A detailed table for the probabilities can be found in Appendix V (p. 30).

5.3. Assigning Migration Probabilities Using the RNG

The probabilities were assigned in the following order because of practical reasons:

1. International out-migration (emigration)
2. Internal migration
3. International in-migration (immigration)

A random number (RN) in SAS was added to the dataset. If the RN was found to be smaller than the probability, the person in question will migrate. If this was not the case, no migration will take place and nothing will happen for that person.

5.3.1. Emigration

There is, however, need for a correction when applying this approach. After a person is selected for emigration, he/she will leave the country and the ID in question will be deleted from the dataset. These emigrants, and all their characteristics (age, sex, income, number of cars, drivers' license, etc.), are stored in a separate dataset. This is done in order to assign the third migration probability: immigration. We do this based on the emigrants and their attributes (Section 5.3.3).

5.3.2. Internal Migration

Because some members of the population left due to emigration, a correction was needed for the probability of assigning internal migration in the next step. This was done the following way:

$$P_{internal\ migration\ corrected} = \frac{P_{internal\ migration}}{(1 - P_{international\ out})}$$

Where *P* is a representation of the migration probability. This way, the correct probability is assigned to the remaining population. After the out-migration is assigned, the RNG is used to select those who will internally migrate and to randomly assign a new municipality for the 'movers'. There are 589 municipalities in Belgium but simply generating a random number between 0 and 590 would not result in the correct output. The municipalities are coded using the NIS-reference or 'mc_gem' variable. Since they are coded this way, a procedure called 'survey select' was used to randomly select a new municipality from the dataset for every mover.

5.3.3. Immigration

Based on available data of the Flemish Government, we can see that more people are entering Belgium than leaving it, meaning that the net balance is positive (Figure 13, Appendix III). This trend is also likely to take place in the future based on a scenario of constant evolution of the migration balance [1].

One important assumption that we make in this study is that the immigrants will have the same characteristics of those who have left the country: the emigrants. A technique called 'oversampling' was therefore applied. For example: if 100 people left the city of Antwerp, a sample of 300 immigrants will be generated with the exact same attributes of the emigrants, given the assumption that there are three times as many immigrants as emigrants. Oversampling, however, generates an additional problem: It is unlikely that immigrants will have the exact same characteristics as the emigrants. Also, the fact that they will move to same municipality as that of the emigrants, is questionable. In real life, large cities, and their suburbs, such as Brussels, Antwerp en Liège will have a larger probability of being selected by immigrants [30]. Here, a new municipality was assigned randomly to the immigrants.

5.4. Corrections

As mentioned before, assigning migration probabilities calculated on a person-level, will eventually lead to problems with the remaining population. Imagine a household of four members: a father, a mother and two children. Suppose that the father was selected for emigration. This would result in a three-member household that is left behind (a mother and two children). This is, of course, not a realistic scenario. In real life, it is often the case that the entire family would migrate or all household members decide to stay home. In order to resolve this problem, corrections for the remaining population were carried out and are described below.

5.4.1. Corrections for Municipality and Statistical Sector

By randomly assigning the internal migrants and the immigrants to a new municipality, a correction is needed for the relationship between the municipality variable and the statistical sector (SS). The SS is a more detailed level of modeling than the municipality, which was not taken into account at first sight. Once a new municipality was generated for every internal migrant and immigrant, the SS was still linked to the 'old' municipality. This was not correct and as a result, problems occurred. An SQL procedure in SAS was used to calculate the cumulative probability for every internal (im)migrant. This resulted in in an inflated dataset with every possible combination of municipality and SS. Next, a RN was used to assign a new SS to the migrants.

5.4.2. Corrections for Households

After assigning the migration probabilities, the resulting population is disturbed because of the discrepancy between the household and person database. Therefore, we need to rearrange the remaining households within the remaining population. In order to keep the influence on the socio-demographic attributes to a minimum, we used, again, a random approach to compose new households. First, a query was built to select the children that were younger than 16 and were living alone. After they have been stored in a new dataset, they needed to be reassigned to a new household. These households were also selected and stored in a separate dataset, taking the household size in account. It is not realistic that a household with 5 children already, will be reassigned a new one.

6. ANALYSIS & RESULTS

The analysis aims to describe what trends take place during the simulation and to investigate whether or not this is due to the migration module. We will be doing this by taking two fields of exploration into account: the population size and the age distribution.

First, a short introduction of Belgium is given to explain the context. Belgium exists of three regions: the Flemish, Walloon and Brussels region. In Figure 14 (Appendix VI, p. 31), the Flanders region is indicated in green and the Walloon region in blue. The city of Brussels, which is the capital, is considered as a separate region together with its suburbs (yellow). The Flemish and Walloon regions are divided in 10 provinces where they both have 5. For Flanders they are: Limburg, Antwerp, East-Flanders, West-Flanders (coastal region) and Flemish-Brabant. In the Walloon region the 5 provinces are: Walloon-Brabant, Liège, Hainaut, Namur and Luxembourg. The combination of these 10 provinces result in a total of 589 municipalities of which 308 are located in Flanders, 262 in Wallonia and 19 in Brussels. There is also an additional level between the municipalities and the provinces namely the district. In total there are 43 districts, but this level is not taken into account for further analysis.

6.1. Selection of Variables

The synthetic dataset comprises many variables for each individual in the dataset. Due to the private nature of this data, two variables are selected for exploring the migration effects: the population size and age. The results are compared with data of the Flemish Government in Table 6 (Appendix VIII, p. 33). The population size provides us with an overall view on the population. By offsetting these results on map, we can easily make comparisons with previous years and see what trends occur. Age is taken into account to see what effect the simulation process will have on the age distribution.

6.2. Selection of Municipalities

Belgium has 589 municipalities. Analyzing all of them is time consuming and not the purpose of this study. Instead, we aim to present a clear picture by taking three types of municipalities into account based on information of the ‘Structuurplan Vlaanderen’ [31]. This is an official document that prescribes guidelines for the spatial development in Flanders. In this study, we make a distinction between three types of regions:

1. Highly populated municipalities: Brussels, Antwerp, Ghent and Louvain.
2. Municipalities in rural areas: Borgloon, Hamont-Achel, Bree.
3. Coastal region municipalities Knokke-Heist, Bruges, Blankenberge, de Haan, Bredene, Ostend, Middelkerke, Nieuwpoort, Koksijde and De Panne.

The first group has municipalities with a number larger than 75.000. All of the rural area municipalities have totals less than 15.000. Cities regarded as ‘highly populated’ are: Brussels, Ghent, Louvain and Antwerp. These cities form what is called the ‘Flemish Diamond’ because their location on map is diamond-shaped (red) (Figure 5). These cities form one of the densest regions in the whole of Europe and are considered of great economic importance [32]. A selection of three municipalities was done for the province of Limburg.

This province is often considered as the 'green' province with lots of open spaces (green) due to the availability of building parcels [31]. Finally, the coastal region is taken into account for investigation of the modeling problems that occur in this region (blue). The selected municipalities are highlighted in yellow (Figure 5):

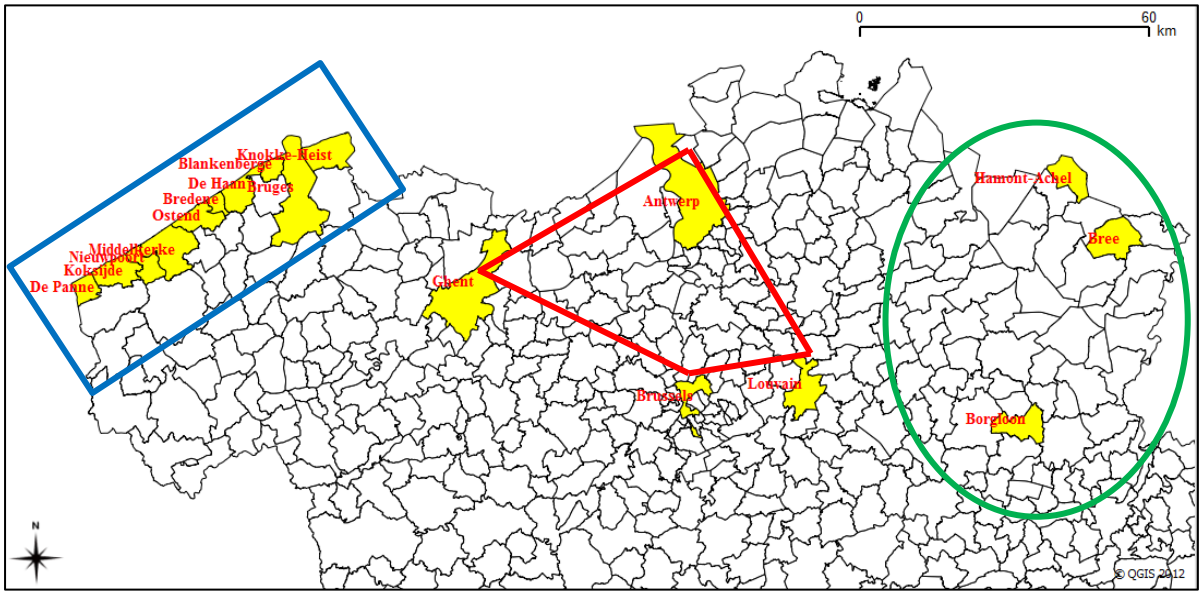

FIGURE 5 Map of selected municipalities.

Please note that the city of Brussels is selected for further analysis and not the entire region. Next, an analysis of the base-year data (2001) is explained and followed by a more detailed look of the simulated results from the SAS iteration process. These results range from 2001 to 2005. The initial goal was to simulate until 2020. However, due to practical reasons such as operationalizing the calculating capacity of the server, a four-year simulation interval was selected.

6.3. Base-Year Analysis

In chronological order, we will discuss the population size and the age distributions of the selected municipalities.

6.3.1. Population Size

The total population of the synthetic dataset for 2001 was 10.296.350 individuals. An SQL procedure in SAS was used to provide us with Table 5 in Appendix VII (p. 32). It contains the NIS-and population numbers of the selected municipalities. The NIS-numbers from table 5 form a unique code and are given by the government to each municipality in Belgium [33]. As we expected, the Flemish Diamond accounted for most inhabitants of the sample with almost 900.000 persons (Figure 6):

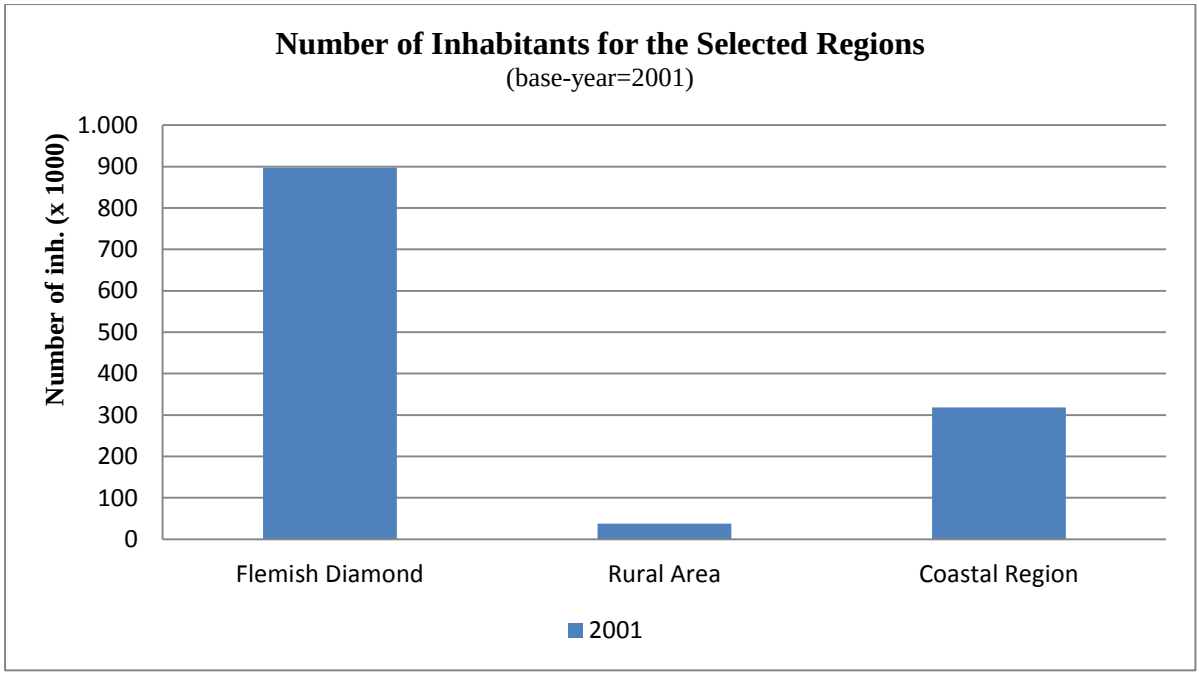

FIGURE 6 Number of inhabitants for the selected regions (year=2001).

The city of Antwerp alone had already more than 400.000 inhabitants. In total, the Flemish Diamond accounted for almost 900.000 persons. The 10 municipalities of the coastal region were good for 318.531 inhabitants while the rural area had a total of less than 40.000 inhabitants. The rural areas and coastal region owned a smaller share of inhabitants ranging from 7.545 to 116.868, with Bruges and Ostend being the exceptions. These three regions combined for a sample with the size of n=1.253.992 corresponding with 12.18 % of the total population in 2001.

The spatial distribution for the number of inhabitants of the 17 municipalities is displayed in Figure 7. Because of the small number of iterations (4), it is difficult to display large differences in the population distribution. To effectively display any possible changes, classifications with increments of 20.000 inhabitants were used.

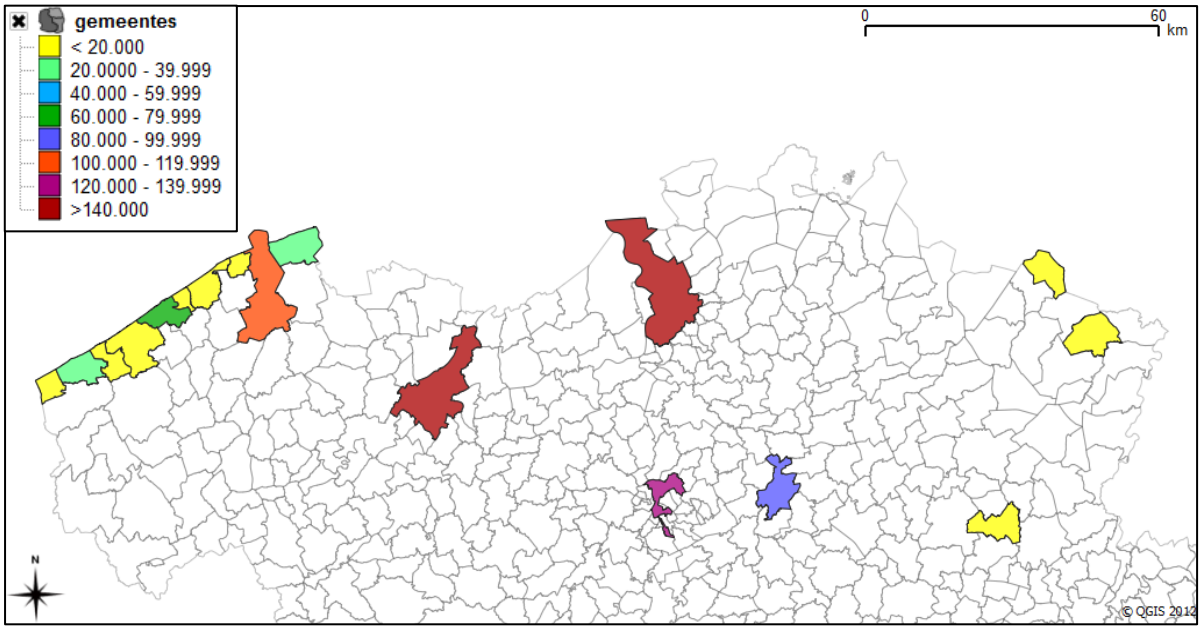

FIGURE 7 Spatial distribution of the number of inhabitants for the selected municipalities (base-year=2001).

The highest population numbers were found in the large cities: Ghent and Antwerp had inhabitants of over 140.000 followed by Brussels. Bruges was next in line with a population that ranges between 100.000 and 119.999 individuals. Hence, we can conclude that the highly populated municipalities of Flanders are located in the center. The three municipalities in Limburg had a number that was less than 20.000 persons. This implies the more rural nature of the province. The population of the coastal region showed more variability having municipalities with less than 20.000 persons but also exceptions with more than 100.000 persons.

6.3.2. Age Distribution

The three regions were selected for an analysis of the age distribution (Figure 8). The 17 municipalities are presented in shades of blue that have no specific purpose. The pie charts represent the relative number of people per age class for the selected regions. The age classes were:

- < 18 For persons younger than 18 years of age, non-adults.
- 18 – 65 The working population ranging from 18 to 65 years of age.
- 65 + The part of the population that has reached a pension age.
- 75 + The oldest segment of the population.

This selection is based on a study of Keyfitz and Flieger [34] because it describes the stages of life very well. People younger than 18 do not own a drivers' license and are not considered as adults in Belgium. The adult and working population is found in the blue segment. The pension age varies from sector to sector but here it is assumed to be at 65. Therefore, the number of people older than 65 is not part of the working population. Finally, there is a small segment of people older than 75.

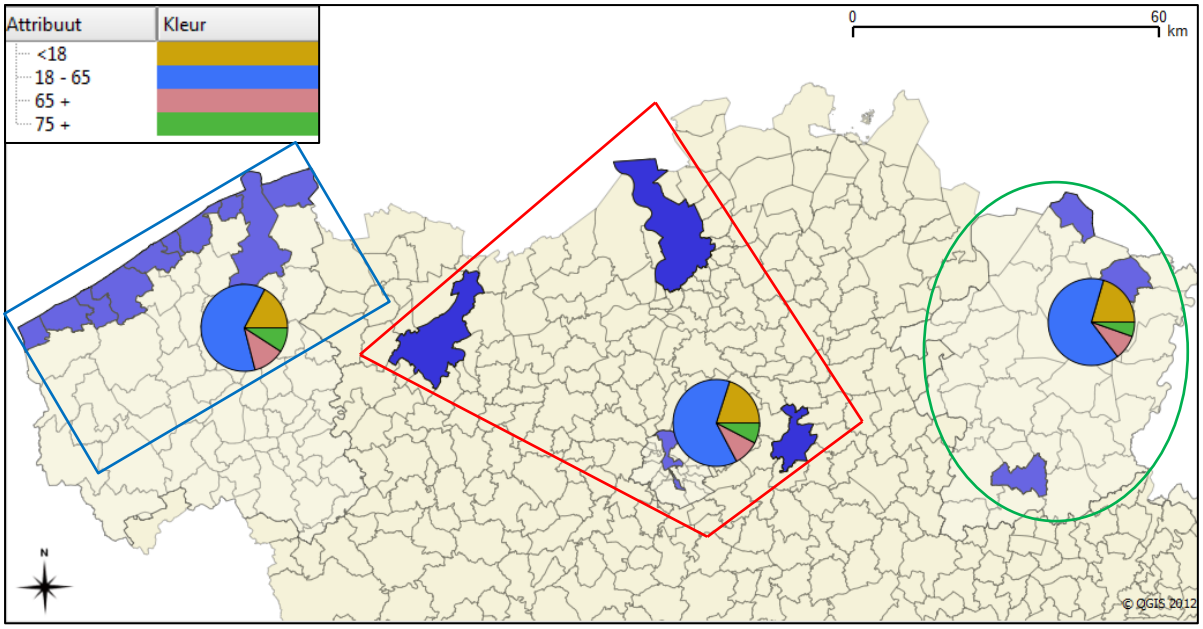

FIGURE 8 Age distribution for selected regions (base-year=2001).

As was expected, the brown segment (<18) was the smallest for the coastal region. It had a share of 17 % and was followed by the Flemish Diamond and the rural region with values of 20.05 and 20.47 % respectively. The working population was found in the blue segment. Of the population in the rural region, 64.89 % was part of the working class. This was higher than the Flemish Diamond: 62.74 %. This is probably due to the high number of commuters for the Flemish Diamond that are living in other cities or provinces. The distribution also showed that the coastal region had the smallest share in this segment. A detailed look of the green segment for the coastal region, revealed a total of approximate 9 %. This was also confirmed by Figure 1. The green segment seems to be the smallest for the rural region (5%) indicating that these municipalities are relatively young while we noted a share of almost 8 % in the Diamond area. For detailed numbers, we refer to Tables 7-11 in Appendix IX (p. 34).

6.4. Simulation Results

For every year of the simulation process, a database with population numbers was achieved. Parallel with the previous section, we describe the population size and age distribution.

6.4.1. Population Size

The total population increased from almost 10.3 million to approximately 10.7 million in 2005. This was an average increase of 0.53 % per year. Compared to the actual population numbers, the model overestimated the total population size by several thousand individuals (Figure 9).

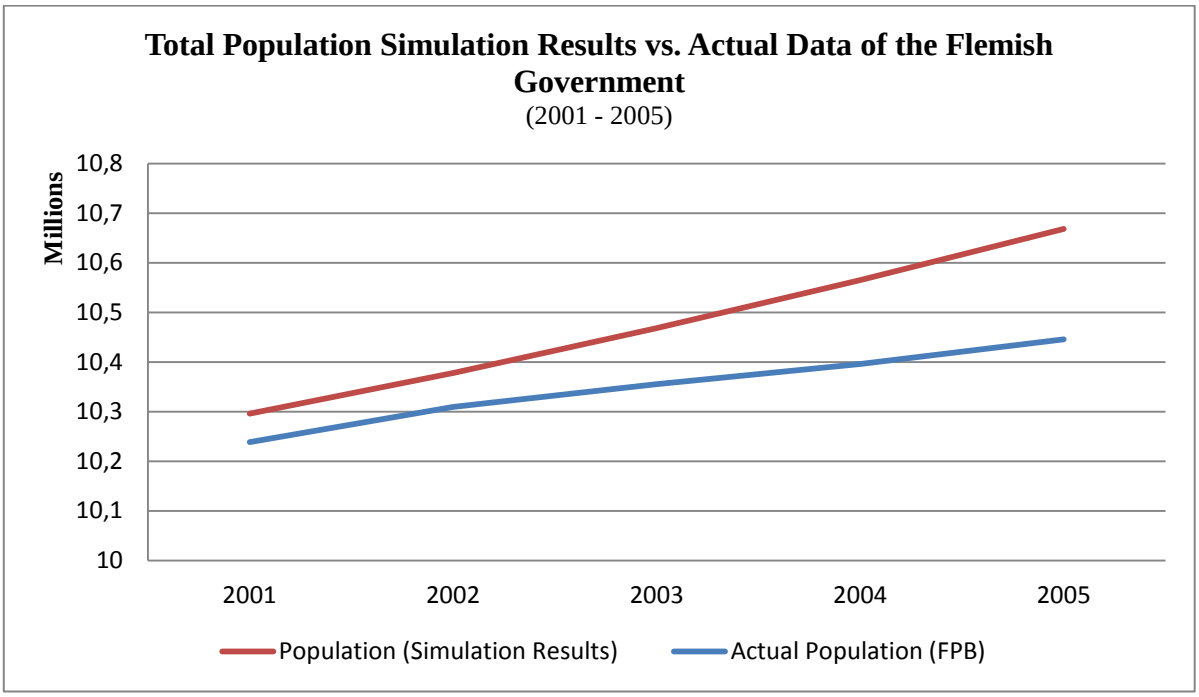

FIGURE 9 Evolution of yearly population totals compared with actual data of the Flemish Government.

A linear growing trend for the simulation results was noted while the actual population showed growth that was less intense. In 2005, our synthetic population total reached 10.668.337 persons while the actual number was 10.445.852 according to the Flemish Government. This is a deviation of 2.13 % and corresponds with a mere 222.485 individuals. The overestimation tended to increase for the entire modeling range. If this trend continues when modeling long terms in the future, larger differences are likely to be noted. According to the Pelfrene [1], the Belgian population growth slows down from 2030 on. If this is really the case, our synthetic population results in even larger differences with the actual population.

The population totals per municipality changed during the modeling processes (Figure 10).

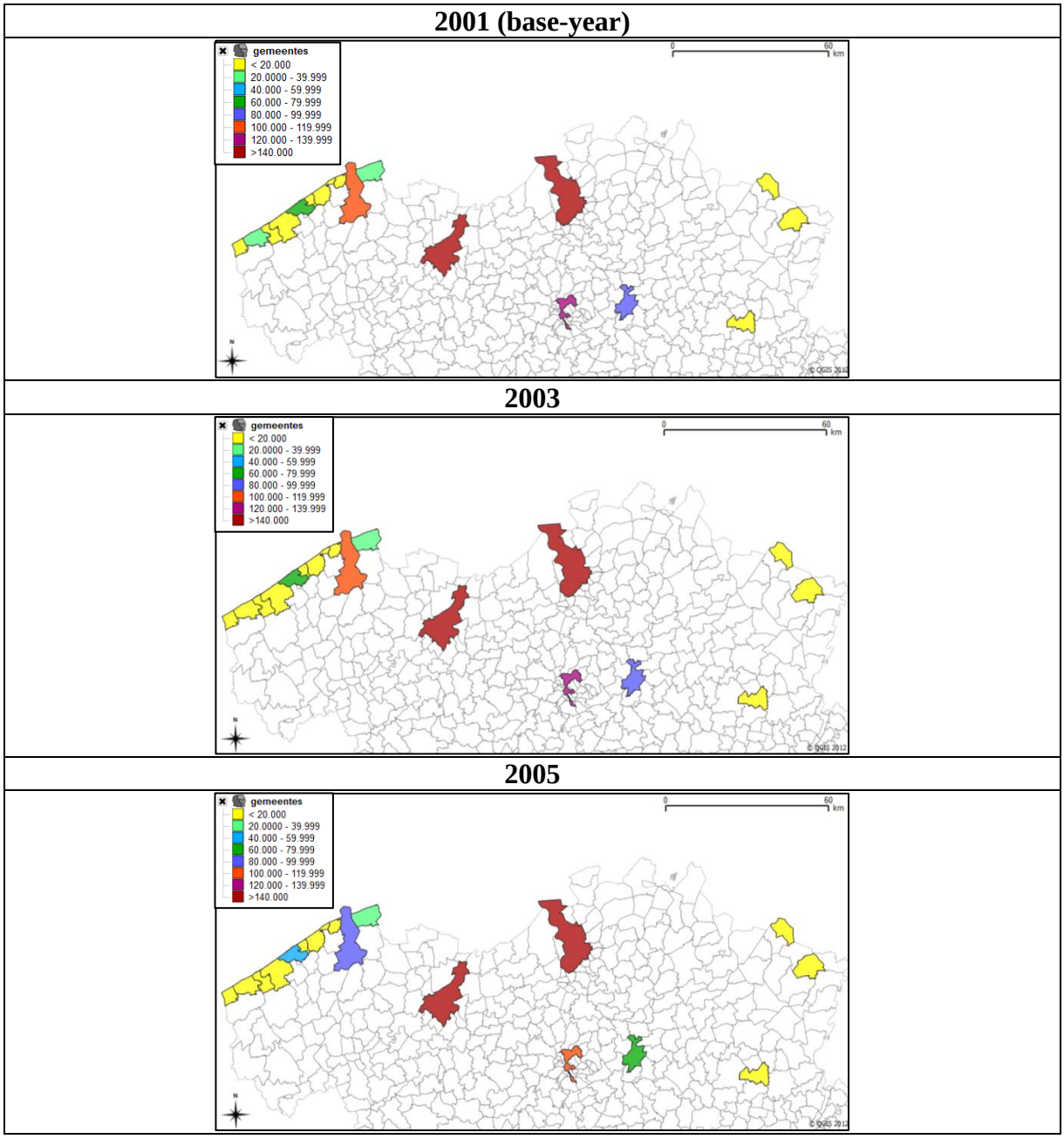

Figure 10 Evolution of population size during the modeling process.

In contrast with total population development, a closer look at the selection showed a decreased population size for almost all municipalities during the modeling process, except the rural ones. The coastal region had 3 municipalities that changed population class (Koksijde, Bruges and Ostend). In the Flemish Diamond, cities like Brussels and Louvain also lost a part of their population. The rural area remained unchanged with the exception of a small increase that is not visible on the map (few thousand individuals). Also not visible on the map, is the major population loss of Antwerp. In 2001, the population of Antwerp was 447.664 persons compared 369.285 persons in 2005. This is an absolute difference of 78.379 individuals in 5 years' time. A decrease was thus noted for all regions except for the rural area. These findings are again displayed in Figure 11.

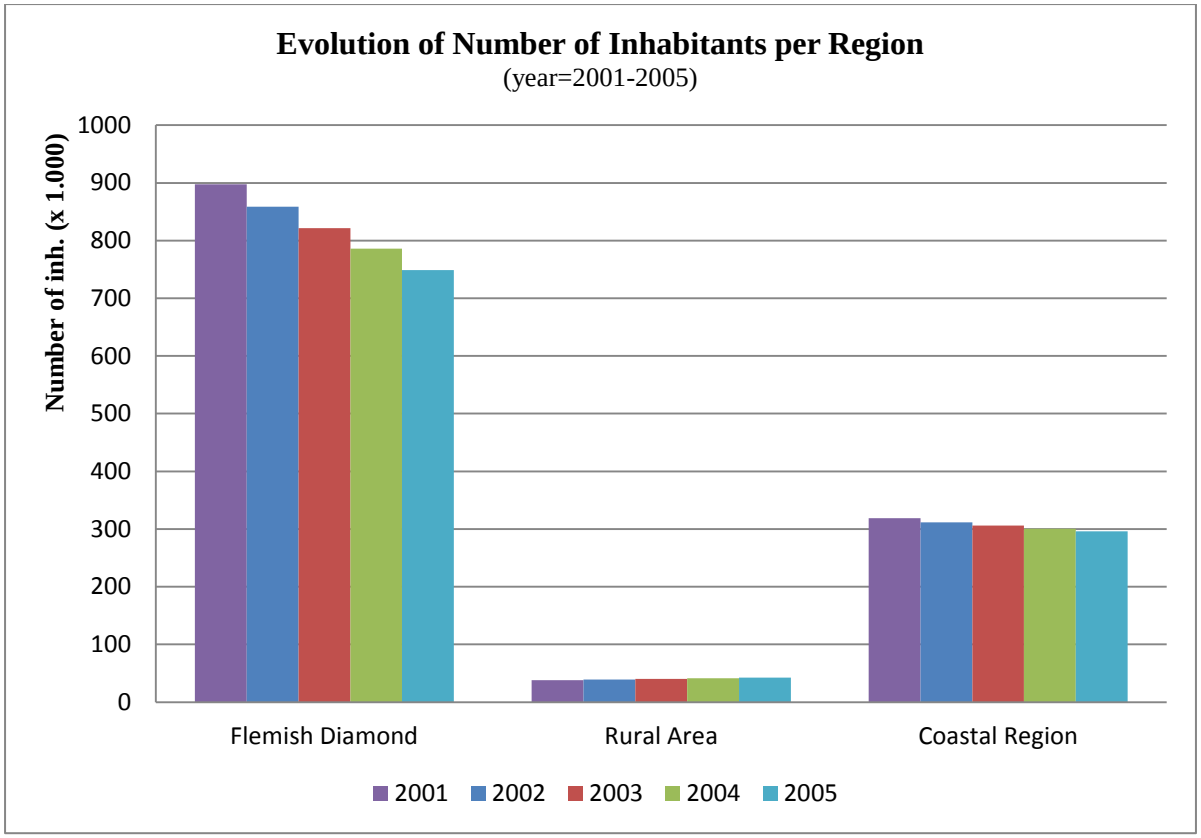

FIGURE 11 Evolution of the number of inhabitants per Region.

An almost constant decrease for the number of persons located in the Flemish Diamond was noted for the entire range of the simulation. This decrease also took place in the coastal region, while the rural area tended to increase. Some of these losses were due to a part of the elder population dying, because birth and death processes were incorporated in the simulation. However, the migration process is assumed to play a large role here. Since a uniform probability was calculated for every person, the model does not take municipal boundaries into account. Because there are more individuals in Antwerp then in Borgloon, more persons from Antwerp could potentially be selected for migration. In real life, the migrants will likely migrate to the large cities whereas now they were assigned over the entire country with a probability of one 589^{th} for every municipality. The randomized approach can therefore explain the loss of population for the large cities and the small increase for the rural municipalities. The migration component did not seem to positively influence the coastal region since it was still suffering from a population loss.

6.4.2. Age Distribution

The iteration process resulted in a shift of the age distributions (Figure 12). In Appendix IX (p. 34), the detailed numbers can be found for the entire simulation range.

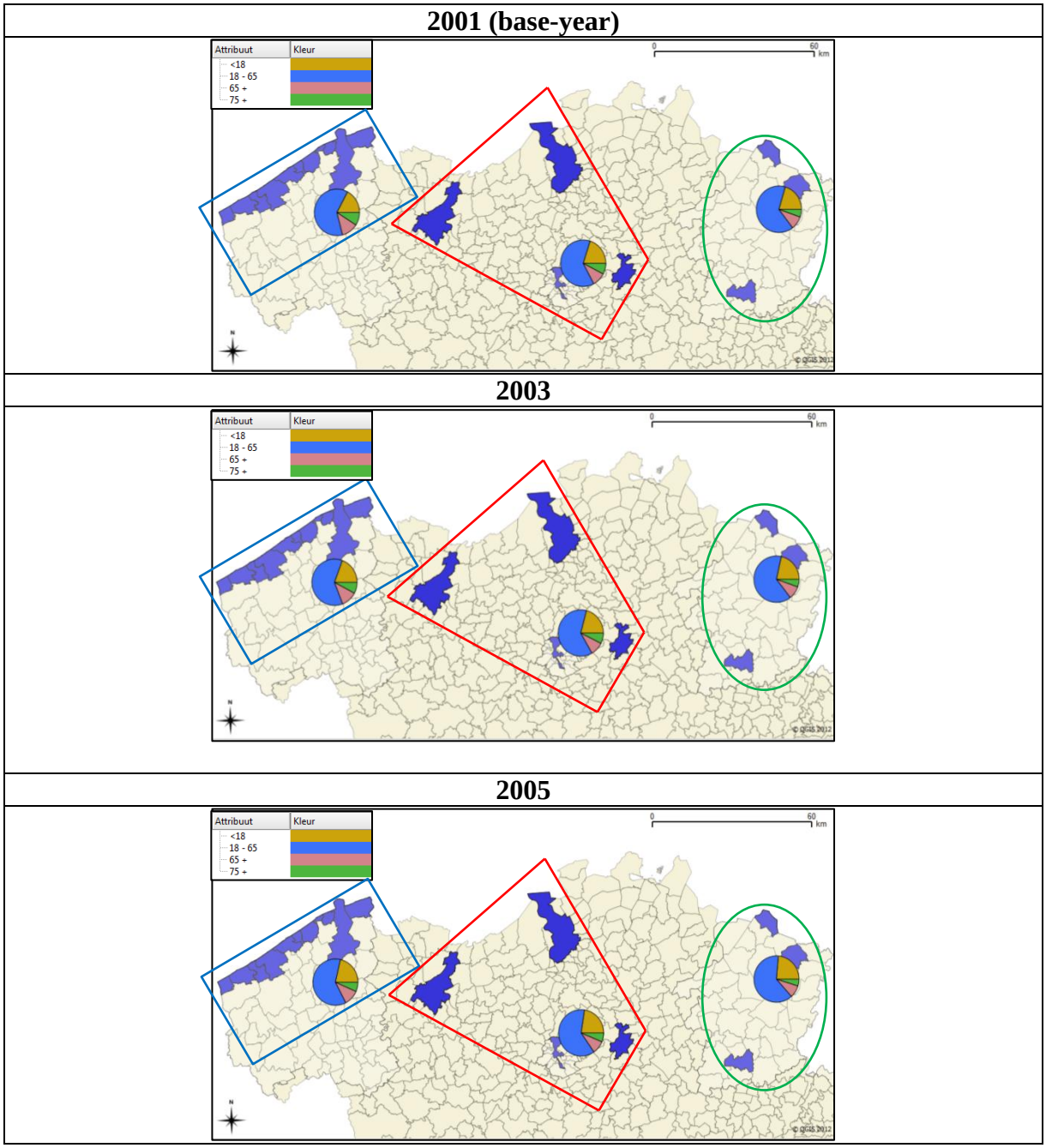

Figure 12 Comparison of age distributions.

For all three regions, an increase in the segment '<18' was found. This increase was approximately 3 % for the entire range of the modeling process. We also noted a decrease in the share of the '75 +' part of the population. This was especially the case in the coastal region. This share went from 9.09 % in 2001 to 6.55 % in 2005. The working population (blue) remained mostly constant with small differences varying from year to year. The segment of '65 +' was also decreasing for all regions during the simulation range, though it was less intense as the drop in the green segment. The migration component did not seem to feed the regions with elderly. However, a gain in non-adults took place which was also the case for the coastal region. We assumed this was due to a combination of the migration module and the births of new children.

This forces us to conclude that the elder part of the population made room for the younger. Nevertheless, the large cities suffered from population losses. Most of these changes were coming from losing the elder part of the population.

7. DISCUSSION

A major problem with the previous version of the model was the loss of populations in certain regions, the coastal region in particular. Using randomized migration, based on person-level probabilities, did not seem to provide a solution for this problem.

Even though an increase in the total population was noted, thus confirming literature [1], large cities such as Brussels and Antwerp showed major losses in the four-year modeling interval. A small increase was noted in the rural region while the coastal region showed signs of small population losses. The total population increase overestimated the actual data with several hundred thousand individuals. Due to the linear growing trend, we assume that this difference even becomes larger in the future if the modeling process is continued. After all, the population growth shows signs of stagnation from 2030 on [1]. Also, a change in the age distribution took place resulting in a rejuvenated population in all of the selected municipalities.

These findings are somewhat contradictory with the literature. According to the Belgische Federale Overheid [35], an increasing trend of juveniles is found. However, it is combined with an increase in the oldest segment of the population (65 + and 75 +). This remains a point of discussion since a stratified analysis at the level of the three Belgian regions indicates that not all regions are suffering from an ageing population. In the region of Brussels, for example, a rejuvenating trend is found while in Flanders, the opposite is true [1]. This was not entirely the case in our study. Our study produced a decrease in the share of elderly for all selected regions and a younger population as replacement. This was partly due to a combination of a large number of newborns and the migration process. However, since we did not take all municipalities into account, we cannot generalize these findings to the national level.

Another explanation for these results is the differences in used data. The migration probabilities in this study were based on migration totals from the Flemish Government and yearly population numbers from the FPB [29]. However, there appears to be a discrepancy with the population size of the synthetic dataset and the totals provided by the FPB. According to this latter, the population of 2001 was 10.263.414 persons while the total of the synthetic population was 10.296.350 persons. Our migration probabilities were based on the numbers of the FPB and thus not taking this deviation into account. This resulted in an underestimation of the migration probabilities which corresponded with a few thousand individuals. Also, assumptions according the migration components were made due to the longitudinal approach that was used for this study e.g. oversampling and random assignment. For a detailed overview of the migration module, we refer to Appendix X (p. 35).

8. CONCLUSIONS

This study aimed to develop a migration module in SAS in order to overcome certain problems of the current model in use e.g. a loss of the elder population in the coastal region. We believe that the randomized nature of the methodology, selecting individuals and reallocating them in a random municipality, is not the best approach to incorporate migration in the model.

A major factor influencing the model is the used data. Data from the Flemish Government were used to establish migration probabilities. These data are of high quality, yet they showed differences with the synthetic data population of 2001. This influenced the estimation strength of the model resulting in a small deviation of a few thousand individuals.

The analysis and results have indicated that the methodology provided questionable results. The migration component, in combination with birth and death processes, resulted in an overestimation of the population and a change in the age distribution of all selected municipalities.

The share of elderly in the coastal region dropped from 9.09 % to 6.55 %. We expect that, if the modeling process is continued, a more equal population distribution will be found in all municipalities. There were no migration weights introduced resulting in an absence of preferences and in a uniform reallocation probability of one 589^{th} for every municipality. More migrants from the large cities were selected and assigned to new municipalities, on a national level, resulting in an exodus of the large cities. In real life, large cities such as Brussels and Antwerp are attractive to migrants [30].

On the other hand, literature is partially confirmed. The model increases the population size significantly. However, we are forced to conclude that using a randomized migration approach on person-level is not the best methodology to model future populations since there are many flaws in the output. The model is thus in need of more fine-tuning.

9. LIMITATIONS AND FUTURE RESEARCH

Recalculating migration on a municipal level, results in a 'weight' per municipality, which is not the case when applying a person-level approach. If the persons-approach of this study is again to be used in the future, additional literature can be consulted to establish weights for the cities/regions, hence increasing their attractiveness for migrants. Another recommendation is necessary concerning the computer capacity. In order to successfully complete the iteration process for the entire population, the script was run on the server of the Imob. However, simulating a single year would take multiple days resulting in a very time consuming process. If this method is applied for other countries or regions with large populations, it is highly recommended that a strong server with high calculating capacity is available. A suggestion for future research is to see if these trends continue if more data becomes available. If the model is optimized, a next step could include the estimation of the transport demand by using the synthetic dataset as input for the Feathers model.

10. ACKNOWLEDGMENTS

I would like to thank Prof. dr. Ir. Tom Bellemans and Prof. dr. Mario Cools for their support and guidance throughout the process. Thanks to our conversations and meetings, I was able to absorb the necessary information that has put me in the right direction. I would also like to thank Katrien Declercq, researcher of the Imob, for helping me with the development of the model. Thanks to her SAS expertise, the migration model was successfully developed and introduced on the Imob server. I would also like to thank my friends, family and girlfriend for supporting me throughout the development. In conclusion, I would like to dedicate this study to my grandfather, who passed away 2 years ago and did not have the chance to see me graduate.

REFERENCES

[1] Pelfrene, E. (2008). *Bevolkingsvooruitzichten 2007-2060.* Brussel: Algemene Directie Statistiek en Economische Informatie.

[2] Van Bavel, J. (2008). Bevolkingssociologie: Omvang en Evolutie van de Bevolking. Katholieke Universiteit Leuven.

[3] Economic and Social Affairs. (2004). *World Population to 2300.* New York: United Nations.

[4] Kitamura, R., & Susilo, Y. (2005). Is Travel Demand Instable? A Study of Changes in Structural Relationships Underlying Travel. *Transportmetrica*, pp. 23 - 45.

[5] Wen, C.-H., & Koppelman, F. (2000). A Conceptual and Methodological Framework for the Generation of Activity-Travel Patterns. *Transportation*, pp. 5 - 23.

[6] Buckland, L., & Jones, M. G. (2008). *Estimating Bicycle and Pedestrian Demand in San Diego.* Washington DC: Transportation Research Board.

[7] Nakamya, J., Moons, E., & Wets, G. (2007). Combining Survey Data from Different Studies to Simulate a Local Travel Survey Sample Data Set: An Application to the Flemish Region. *11th World Conference on Transportation Research* (p. 23). Hasselt: Hasselt University.

[8] Davidson, W., Donnelly , R., Vovsha, P., Freedman, J., Ruegg, S., Hicks, J., et al. (2007, juni 1). Synthesis of First Practices and Operational Research Approaches in Activity-Based Travel Demand Modeling. *Transportation Research Part A: Policy and Practice*, pp. 464-488.

[9] Hall, R. W. (2003). *Handbook of Transportation Science.* Dordrecht: Kluwer Academic Publishers.

[10] De Klerck, P. (2011). *Vergrijzing en ouderenzorg aan de Kust: Moet er nog (Nieuw) Zand zijn?* Brussels: Studiedienst Vlaamse Regering.

[11] De Nauw, J. (2011). *Socio-Demografisch Profiel van de Gemeenten.* Brussel: Dexia.

[12] Lodewijckx, E. (2008). *Veranderende Leefvormen in het Vlaamse Gewest, 1990-2007 (en 2021): een Analyse van Gegevens uit het Rijksregister.* Brussel: Acco Drukkerij.

[13] Lievevrouw, P., Vandekerckhove, B., Moortgat, W., Somers, D., & Van Dorpe, H. (2006). *Ruimtelijke Analyse van de Migratie in en Naar Vlaanderen.* Brussel: SumReasearch.

[14] Iklé, F. C. (1954). *Sociological Relationship of Traffic to Population and Distance.* Connecticut, USA: Eno Foundation.

[15] Department of Economic and Social Affairs. (2011). *Population Distribution, Urbanization, Internal Migration and Development: An International Perspective.* United Nations Publications.

[16] Siegel, J., & Swanson, D. (2004). *The Methods and Materials of Demography.* San Diego, California: Elsevier.

[17] Shryock, H. S., Siegel, J. S., Larmon, E. A., & Bayo, F. (1973). *The Methods and Materials of Demography.* Washington: U.S. Government Printing Office.

[18] Janssens, D., Wets, G., Timmermans, H., & Arentze, T. (2010). *Modelling Short-Term Dynamics in Activity-Travel Patterns: the Feathers Model.* Hasselt: Hasselt Univeristy.

[19] Martin, P., & Zürcher, G. (2008). Managing Migration: The Global Challenge. *Population Reference Bureau*, pp. 3-20.

[20] Robila, M. (2007). Eastern European Immigrants in the United States: A Socio-Demographic Profile. *The Social Science Journal*, 113-125.

[21] Willaert, D. (1999). *Migratieprofielen Naar Leeftijd Voor de Migratiebekkens en Zones in de Nieuwe Ruimtelijke Indeling.* Brussel: Vakgroep Sociaal Onderzoek (SOCO).

[22] Bhat, C. R., & Koppelman, F. S. (2003). *Activity-Based Modeling of Travel Demand.* Dordrecht: Kluwer Academic Publishers.

[23] Beckmann, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating Synthetic Baseline Populations. *Transportation Research Part A: Policy and Practice*, 415-429.

[24] Moeckel, R., Spiekermann, K., & Wegener, M. (2003). Creating a Synthetic Population. *Proceedings of the 8th International Conference on Computers in Urban Planning and Urban Management*.

[25] Guo, J., & Bhat, C. (2007). Population Synthesis for Microsimulating Travel Behavior. *Transportation Research Record: Journal of the Transportation Research Board*, pp. 92-101.

[26] Mohammadian, A., Javanmardi, M., & Zhang, Y. (2010). Synthetic Household Travel Survey Data Simulation. *Transportation Research Part C: Emerging Technologies*, pp. 869-878.

[27] Vlaamse Overheid. (2005). *Socio-Economische Enquête 2001 (SEE2001).* Opgeroepen op januari 16, 2012, van aps.vlaanderen.be: http://aps.vlaanderen.be/sgml/largereeksen/1106.htm

[28] Vlaamse Overheid. (2012, januari 1). *Onderzoek Verplaatsingsgedrag Vlaanderen.* Opgeroepen op december 18, 2011, van MobielVlaanderen: http://www.mobielvlaanderen.be/ovg/

[29] Federaal Planbureau. (2012). *Data.* Opgeroepen op mei 20, 2012, van Federaal Planbureau - Economische Analyses en Vooruitzichten: http://www.plan.be/databases/database_det.php?lang=nl&TM=30&IS=60&DB=DEMOG11&ID=35

[30] De Corte, S., Raymaekers, P., Thaens, K., Vandekerckhove, B., & François, G. (2003). *Onderzoek naar de Migratiebewegingen van de Grote Steden in de drie Gewesten van België.* Brussel: Cel Grootstedenbeleid.

[31] Vlaamse Overheid. (2011). *Ruimteliijk Structuurplan Vlaanderen.* Brussel: Vlaamse Overheid.

[32] Vanhaverbeke, W. (1997). *Het Belang van de Vlaamse Ruit Vanuit Economisch Perspectief.* Maastricht, Nederland: Universiteit Maastricht.

[33] Belgische Federale Overheid. (2010, januari 1). *Administratieve Geografie*. Opgeroepen op april 20, 2012, van Statbel.fgov.be: http://statbel.fgov.be/nl/statistieken/gegevensinzameling/nomenclaturen/admin-geo/#namen

[34] Keyfitz, N., & Flieger, W. (1968). *World Population Growth and Aging: Demographic Trends in the Late Twentieth Century.* United States of America: University of Chicago Press.

[35] Belgische Federale Overheid. (2010). *Structuur van de bevolking*. Opgeroepen op mei 20, 2012, van Statistics Belgium: http://statbel.fgov.be/nl/statistieken/cijfers/bevolking/structuur/leeftijdgeslacht/belgie/

[36] Belgische Federale Overheid. (2010). *Migratie.* Opgeroepen op mei 12, 2012, van Economie - Statistics Belgium: http://statbel.fgov.be/nl/statistieken/cijfers/bevolking/migraties/

[37] Vlaamse Overheid. (2012). *Studiedienst van de Vlaamse Regering.* Opgeroepen op oktober 22, 2011, van aps.vlaanderen.be: http://www4dar.vlaanderen.be/sites/svr/Cijfers/Pages/Excel.aspx

APPENDIX I

TABLE 1 Migration Definitions Used by the Belgische Federale Overheid (2010) (Governmental Agency of Economy)

Type	Definition
(1) Internal migration	The movement of individuals across municipalities in Belgium.
(2) International migration	External migration movement implying movement to and from foreign countries.
(3) Administrable reformation	Refugees of countries that file for asylum will only be taking into account on a waiting list and no longer in the National Register.
(4) Immigration	In-movement of (2) + changes in the register + re-entries in the register after deleting the records.
(5) Emigration	Out-movement of (2) + persons who are cut from the national register.
(6) External migration balance	In-movement of (2) + changes in the register + re-entries in the register after deleting the record – out-movement of (2)

APPENDIX II

TABLE 2 Summary of Data Files

Dataset	Provider	Information
(1) Socio-economic survey (SES)	Federal Government	Conducted in Belgium in 2001. Starting point for synthetic population.
(2) Onderzoek VerplaatsingGedrag (OVG)	Flemish Government	Conducted in Flanders in 2008 for the third time. Input for synthetic population.
(3) Migration numbers	Flemish Government	Trend extrapolations based on data from 1990 – 2009 for upgrading the synthetic population.

APPENDIX III

Belgian migration balance from 1990 – 2009.

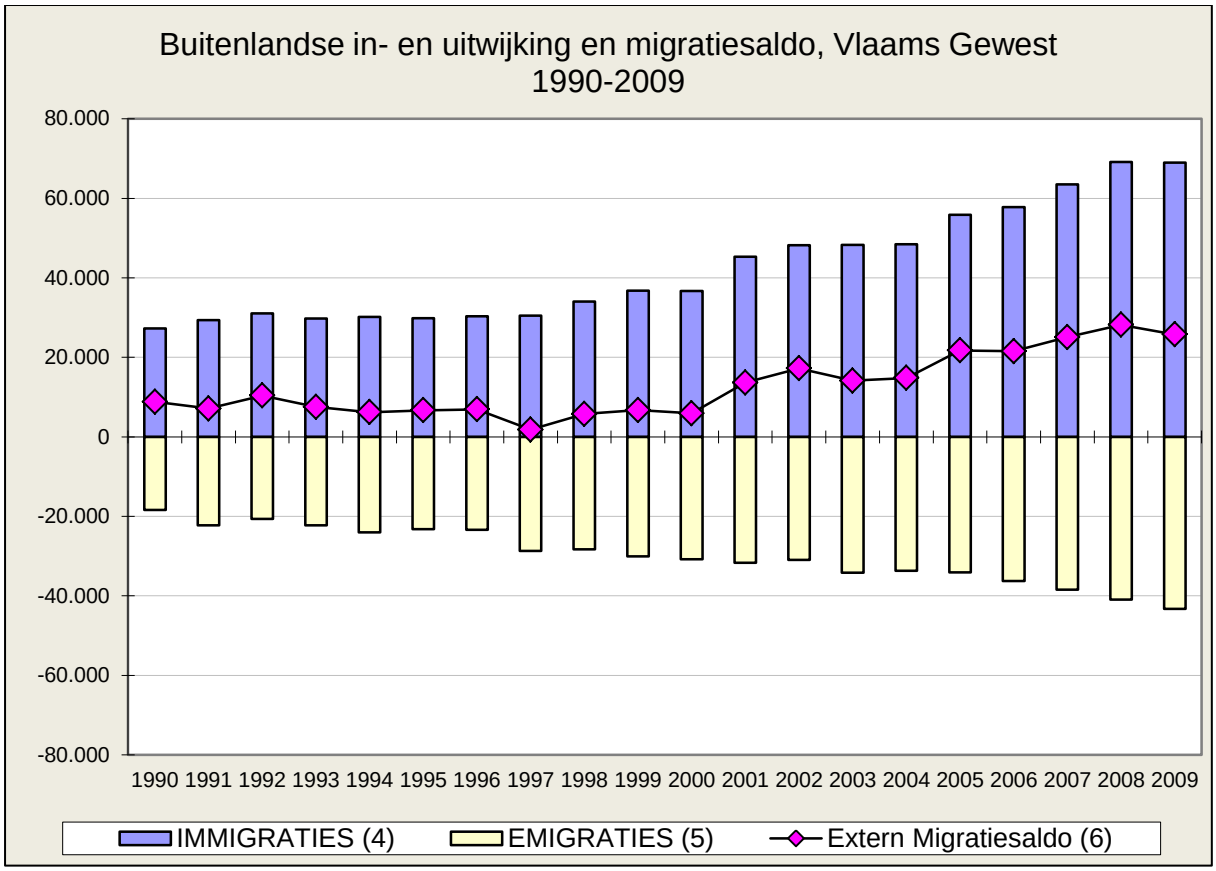

FIGURE 13 Immigration, emigration and international migration balance from 1990 – 2009 [36].

Blue = *immigration totals (international in-migration)*
Yellow = *emigration totals (international out-migration)*
Pink = *net difference (immigration – emigration)*

APPENDIX IV

Trend extrapolation results of migration numbers using the linear method.

TABLE 3 Yearly Migration Totals According to the Flemish Government and own Research

Year	Internal Migration	Immigration	Emigration
1990 (M_b)	384.561	68.929	49.246
1991	391.232	74.617	60.471
1992	420.052	75.940	50.551
1993	429.768	72.762	53.824
1994	441.860	75.621	57.987
1995	442.530	71.563	58.184
1996	449.573	70.581	57.867
1997	453.085	74.578	68.537
1998	451.759	83.812	72.087
1999	453.218	91.624	74.097
2000	442.564	89.388	75.320
2001(M_l)	447.042	110.410	75.261
2002	459.536	113.857	75.960
2003	470.695	112.060	79.399
2004	481.279	117.236	83.895
2005	491.338	132.810	86.899
2006	505.216	137.699	88.163
2007	514.084	146.409	91.052
2008	534.816	164.152	100.275
2009	532.232	166.479	103.718
2010	539.616	171.357	106.442
2011	546.999	176.234	109.165
2012	554.383	181.112	111.889
2013	561.766	185.989	114.612
2014	569.150	190.867	117.336
2015	576.533	195.744	120.060
2016	583.917	200.622	122.783
2017	591.300	205.499	125.507
2018	598.684	210.377	128.230
2019	606.068	215.254	130.954
2020	613.451	220.132	133.678
Total	15.538.306	4.203.711	2.783.449

M_b = *Base Year*
M_l = *Launch Year*

APPENDIX V

Migration probabilities for Belgian individuals.

TABLE 4 Migration Probabilities per Year

Year	Internal Migration	Immigration	Emigration
2001	4.36%	1.08%	0.73%
2002	4.46%	1.10%	0.74%
2003	4.55%	1.08%	0.77%
2004	4.63%	1.13%	0.81%
2005	4.70%	1.27%	0.83%
2006	4.81%	1.31%	0.84%
2007	4.86%	1.38%	0.86%
2008	5.02%	1.54%	0.94%
2009	4.96%	1.55%	0.97%
2010	4.99%	1.59%	0.98%
2011	5.02%	1.62%	1.00%
2012	5.06%	1.65%	1.02%
2013	5.09%	1.68%	1.04%
2014	5.12%	1.72%	1.05%
2015	5.15%	1.75%	1.07%
2016	5.18%	1.78%	1.09%
2017	5.21%	1.81%	1.11%
2018	5.25%	1.84%	1.12%
2019	5.28%	1.88%	1.14%
2020	5.32%	1.91%	1.16%

APPENDIX VI

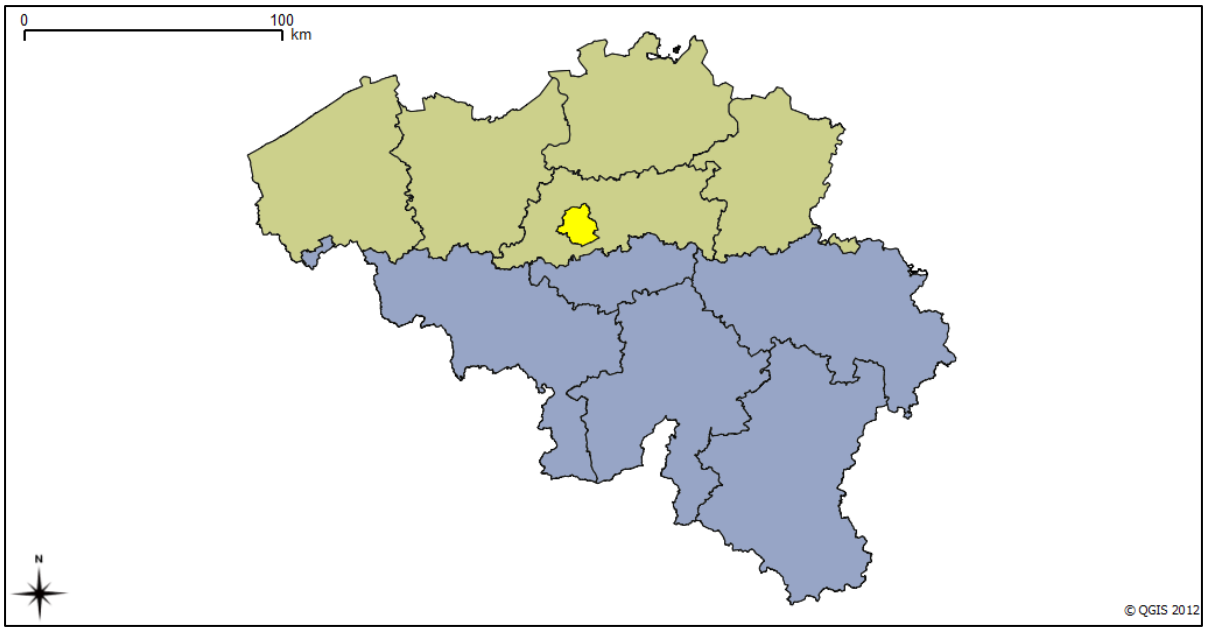

FIGURE 14 Regions of Belgium.

APPENDIX VII

TABLE 5 NIS-and Population Numbers of Selections for the Base-Year (2001) [37]

Region & Municipalities		NIS-number	Population	Total
Flemish Diamond	Brussels	21004	135.875	**897.577**
	Antwerp	11002	447.664	
	Ghent	44021	225.422	
	Louvain	24062	88.616	
Rural area	Borgloon	73009	10.076	**37.884**
	Hamont-Achel	72037	13.661	
	Bree	72004	14.147	
Coastal region	Knokke-Heist	31043	33.314	**318.531**
	Bruges	31005	116.868	
	Blankenberge	31004	17.690	
	De Haan	35029	11.570	
	Bredene	35002	14.436	
	Ostend	35013	67.500	
	Middelkerke	35011	16.821	
	Nieuwpoort	38016	10.410	
	Koksijde	38014	20.052	
	De Panne	38008	9.870	
TOTAL OF SELECTION				**1.253.992**
TOTAL POPULATION (2001)				**10.296.350**

APPENDIX VIII

TABLE 6 Population Totals of Simulation vs. Actual Data [29]

Year	Population (Simulation)	Population FPB
2001	10.296.350	10.263.414
2002	10.378.078	10.309.725
2003	10.468.437	10.355.844
2004	10.565.608	10.396.421
2005	10.668.317	10.445.852

APPENDIX IX

Age distributions for the selected regions.

TABLE 7 Age Distribution for Selected Municipalities and Regions (year=2001)

2001				
Region/Municipality	**Age segments**			
	<18 (%)	**18 – 65 (%)**	**65 + (%)**	**75 + (%)**
Coastal Region	17.24	61.80	11.88	9.09
Flemish Diamond	20.05	62.74	9.62	7.59
Rural Region	20.47	64.89	9.25	5.38

TABLE 8 Age Distribution for Selected Municipalities and Regions (year=2002)

2002				
Region/Municipality	**Age segments**			
	<18 (%)	**18 – 65 (%)**	**65 + (%)**	**75 + (%)**
Coastal Region	17.39	61.50	11.85	9.25
Flemish Diamond	21.08	62.48	9.42	7.03
Rural Region	21.66	64.36	8.95	5.03

TABLE 9 Age Distribution for Selected Municipalities and Regions (year=2003)

2003				
Region/Municipality	**Age segments**			
	<18 (%)	**18 – 65 (%)**	**65 + (%)**	**75 + (%)**
Coastal Region	19.40	61.50	11.32	7.78
Flemish Diamond	20.09	59.58	8.98	7.03
Rural Region	21.58	63.96	9.09	5.37

TABLE 10 Age Distribution for Selected Municipalities and Regions (year=2004)

2004				
Region/Municipality	**Age segments**			
	<18 (%)	**18 – 65 (%)**	**65 + (%)**	**75 + (%)**
Coastal Region	20.46	61.40	11.05	7.09
Flemish Diamond	21.20	63.97	8.60	6.24
Rural Region	22.83	63.27	8.86	5.04

TABLE 11 Age Distribution for Selected Municipalities and Regions (year=2005)

2005				
Region/Municipality	**Age segments**			
	<18 (%)	**18 – 65 (%)**	**65 + (%)**	**75 + (%)**
Coastal Region	21.49	61.17	10.79	6.55
Flemish Diamond	22.66	61.98	9.09	6.28
Rural Region	23.71	62.89	8.65	4.76

APPENDIX X

SAS script for migration module.

```
/* INFORMATIE: Thomas Rayen, 04/01/2012, 17u46 */
/* open de bibliotheken */
/*LIBNAME Testen 'C:\Users\Thomas\Desktop\2e Master\Thesis\Katrien\Testen'; /*
LIBNAME Thesis 'C:\Users\Thomas\Desktop\2e Master\Thesis\Data\Thomas_sas';*/
%LET PATH = E:\Katrien\Synthetic Data\Implementation;
libname DATA "&PATH\KAD\input datasets";
libname MARIO "&PATH\Mario";
libname THOMAS "E:\Katrien\Synthetic Data\Implementation\MARIO\Prognoses\Thomas";
libname PROGNOSE "&PATH\Mario\Prognoses";
*/
/* De variabele jaar wordt in een macro gestopt */
%let Year=2007;
/* Koppelen van de migratiekans (internationale uitmigratie) aan elke individu van de sample */
/* Omdat we werken met 3 onafhankelijke migratiekansen, hebben we een correctie moeten uitvoeren voor de interne migratiekans */
data Samplepers;
        if _n_=1 then set Thomas.Migration(where=(Year=&Year));
        set persons13;
        ranmigr1=ranuni(&Year);
        chance_internal_migra_corr=chance_internal_migra/(1-chance_international_out);
run;

/* We tellen het aantal personen van de sample en slaan deze op in de variabele 'popsize' */
proc sql;
  select count(*)
  into :popsize
  from persons13;
quit;

/* In deze stap kijken we naar de personen die bevolking gaan verlaten (internationale uitmigratie) en worden gestuurd naar de expats */
/* De randvoorwaarde van leeftijden 0 - 42 dient om een restrictie op te leggen voor de sociale mobiliteit */

/* STAP 1: UITSTROOM (INTERNATIONALE UITMIGRATIE) */
data Samplepers1 expats;
        set Samplepers;
        length migr $ 15;
        if /* 0 LE leeftijd LE 42 AND */ ranmigr1 <= chance_international_out then output expats;
   else do;
      if /* 0 LE leeftijd LE 42 AND */ ranmigr1 LE chance_internal_migra_corr then do;
                                        andere_gem=1;

                                                HHID = 10000000 + HHID;

                                                migr = "internal";
```

```
                                                                    end;
                    output Samplepers1;
                    end;
run;

/* STAP 2: INTERNE MIGRATIE */
/* We gaan de personen die intern verhuisd zijn een '1' toekennen voor variabele 'anderegem' */
proc sql;
          select count(*)
          into :toreplace
          from Samplepers1
          where andere_gem=1;
quit;
/* Maak tabel gemeente. Dit zijn alle verschillende gemeentes die in de sample voorkomen (met hun NIS code)
*/
/* Hier eventueel de gewichten toekennen van bepaalde steden */
proc sql;
  create table gemeentes as
  select distinct mc_gem
  from MARIO.synthpers;
quit;
/* Dit zijn de personen die van gemeente zijn veranderd (met al hun karakteristieken) */
data anderegem;
  set Samplepers1;
  where andere_gem = 1;
run;
/* Voor een lijst van macrovariabelen te maken voor alle gemeentes */
data _null_;
  set anderegem;
  call symput('ID'||left(_n_),ID);
  call symput('gem'||left(_n_),mc_gem);
run;

/* Macro die random gemeentes selecteert om toe te wijzen aan de personen die intern migreren */
/* De ID wordt eraan gekoppeld */

OPTIONS NONOTES;
proc datasets nolist;
delete newmunicipality;
run;

%macro newmun;
%do i = 1 %to &toreplace;
proc surveyselect data=gemeentes(where = (mc_gem NE &&gem&i)) out=out sampsize=1 noprint;
run;

data out;
```

```
  set out;
  ID = &&ID&i;
run;

proc datasets nolist;
 append base = newmunicipality data = out;
run;
%end;
%mend newmun;

%newmun
OPTIONS NOTES;

/* Interne migratie: mc_sec selecteren in de nieuwe gemeente */
/* We tellen het totaal per mc_sec */
/* STAP 3: BEREKEN DE PROBABILITEITEN PER GEM EN SEC VOOR PERSONEN VAN DE INTERNE MIGRATIE AD RANDOM TOE TE WIJZEN */
/* Hier worden ze 'opgeblazen' om achteraf eentje te selecteren */

proc sql;
  create table gemsec as
  select mc_gem, mc_sec, count(*) as seccount
  from persons13
  group by mc_gem, mc_sec;
quit;

/* het aantal huishoudens per gemeente bepalen voor de gehele bevolking */
proc sql;
  create table gemtotals as
  select mc_gem, count(*) as gemcount
  from persons13
  group by mc_gem;
quit;

/* per gemeente het percentage huishoudens per sector bepalen */
data gemsecprob;
  merge gemsec gemtotals;
  by mc_gem;
  secprob = seccount/gemcount;
run;

/* per gemeente per sector de cumulatieve kansen berekenen*/
data gemsecprob;
  set gemsecprob;
  by mc_gem;
  retain cumsecprob 0;
  if first.mc_gem then cumsecprob = 0;
  cumsecprob + secprob;
```

```
run;

/* aan elk interne migrant met nieuwe mc_gem een nieuwe mc_sec toekennen dmv een random number */
data newmunicipality;
 set newmunicipality;
 rannr = ranuni(&year + 20);
run;

/* voor de interne migranten met alle mogelijke mc_secs die bij de nieuwe mc_gem horen aan de data linken */
proc sql;
 create table newmunicipality2 as
 select a.*, b.mc_sec as imputed_mc_sec, b.cumsecprob
 from newmunicipality as a
     left join
          gemsecprob as b
          on a.mc_gem = b.mc_gem
 order by ID, cumsecprob
;
quit;

/* op basis van het random nummer een mc_sec selecteren per HH */
/* We gaan uiteindelijk één mc_sec en één mc_sec overhouden per persoon */
data imputed_mc_sec;
 set newmunicipality2;
 by ID;
 retain done;
 if first.ID then done = 0;
 if not done and rannr LE cumsecprob then do;
                                                output;
                          done = 1;
                                                                end;
 drop rannr done cumsecprob;
run;

/* De ge-update versie van de sample (Met de correcte koppeling tussen de mc_gem en mc_sec) */
proc sql;
 create table Samplepers2 as
 select a.*, b.mc_gem as new_mc_gem, b.imputed_mc_sec
 from Samplepers1 as a
     left join
     imputed_mc_sec as b
 on a.ID = b.ID;
quit;

data Samplepers2;
 set Samplepers2;
 if new_mc_gem NE . then mc_gem = new_mc_gem;
 if imputed_mc_sec NE "" then mc_sec=imputed_mc_sec;
```

```
  drop new_mc_gem imputed_mc_sec;
run;

/* chance_international_in opslaan in macro-variabelen om later te gebruiken */
/* mag voor de macro */
data _null_;
  set THOMAS.migration_chance(where = (year NE .));
  call symput('IMMIGR'||left(year),chance_international_in);
run;

%let sampsize = %sysevalf(&&IMMIGR&year * &popsize,integer);

/* STAP 4: BEREKENING VAN IMMIGRANTEN */
/* Belangrijke assumptie: de inwijkelingen beschikken over dezelfde karakteristieken als diegene die zijn uitgeweken en maken gebruik van oversampling */
/* We passen dezelfde methodiek toe als hierboven (berekenen van de probabiliteiten per gemeente en sector */

proc surveyselect data = expats out = immigrants method = urs sampsize = &sampsize outhits;
run;

data immigrants;
  set immigrants;
  HHID = 20000000 + _n_;
  ID = HHID;
  migr = "immigration";
run;

/* Tellen van het aantal immigranten */
proc sql;
        select count(*)
        into :toreplace
        from immigrants
;
quit;

data _null_;
  set immigrants;
  call symput('ID'||left(_n_),ID);
run;

/* De immigranten krijgen een nieuwe gemeente toegewezen en een nieuwe ID */
OPTIONS NONOTES;
proc datasets nolist;
delete newgem;
run;

%macro newgem;
%do i = 1 %to &toreplace;
```

```
proc surveyselect data=gemeentes out=out sampsize=1 noprint;
run;

data out;
  set out;
  ID = &&ID&i;
run;

proc datasets nolist;
 append base = newgem data = out;
run;
%end;
%mend newgem;

%newgem
OPTIONS NOTES;

/* aan alle immigranten een nieuwe mc_gem een nieuwe mc_sec toekennen dmv random number */
data newgem;
  set newgem;
  rannr = ranuni(&year + 40);
run;

/* voor de immigranten alle mogelijke mc_secs die bij de nieuwe mc_gem horen aan de data linken */
proc sql;
  create table newgem2 as
  select a.*, b.mc_sec as imputed_mc_sec, b.cumsecprob
  from newgem as a
      left join
            gemsecprob as b
            on a.mc_gem = b.mc_gem
  order by ID, cumsecprob
;
quit;

/* op basis van het random nummer een mc_sec selecteren per HH */
data newgem3;
  set newgem2;
  by ID;
  retain done;
  if first.ID then done = 0;
  if not done and rannr LE cumsecprob then do;
                                                output;
                              done = 1;
                                                                end;
  drop rannr done cumsecprob;
run;
```

```
/* We koppelen de nieuwe mc_gem en mc_sec eraan vast door te joinen met hun overige karakteristieken */
proc sql;
  create table immigrants2 as
  select a.*, b.mc_gem as new_mc_gem, b.imputed_mc_sec
  from immigrants as a
       left join
       newgem3 as b
  on a.ID = b.ID;
quit;

data immigrants2;
  set immigrants2;
  if new_mc_gem NE . then mc_gem = new_mc_gem;
  if imputed_mc_sec NE "" then mc_sec=imputed_mc_sec;
  drop new_mc_gem imputed_mc_sec;
run;

/* STAP 5: CORRECTIES OP HUISHOUDNIVEAU VOOR DE NIEUWE HUISHOUDENS */
data Samplepers3;
  set Samplepers2 immigrants2(in=inimm);
  if inimm then immigrant = 1;
  else immigrant = 0;
run;

OPTIONS NOTES;

/* aangepaste versie van persons13 */
data persons13;
  set Samplepers3;
run;
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Refining synthetic population generation: development of a migration data model

Richting: **master in de verkeerskunde-mobiliteitsmanagement**
Jaar: **2012**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Rayen, Thomas

Datum: **1/06/2012**