TECHNISCHE UNIVERSITÄT WIEN

DIPLOMARBEIT

Different Strategies for Modelling and Simulation of Regional Population Development

Ausgeführt am Institut für
Analysis und Scientific Computing
der Technischen Universität Wien

unter der Anleitung von
Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Felix Breitenecker
und
Projektass. Dipl.-Ing. Dr.techn. Martin Bicher

durch
Matthias Rudolf Obermair

....
Ort, Datum   Unterschrift

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen und Abbildungen – die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

I hereby declare to have written this Master Thesis independently and completely specified the utilized sources and resources. All parts of the thesis – tables and figures included – that are taken from other works including the internet, literally or extracted, have been marked by referencing the source as borrowed.

....	....
Ort, Datum	Unterschrift

Abstract

Population dynamics have always been a very important global and local issue, therefore reliable predictions can be useful for social, economical and political decisions, such as strategies in the health-care system. Demographic developments are not only a point of interest on a national but also on a sub-national level. Without real borders separating regions, migration rates between them are typically much higher compared to international movements.

The goal of this diploma thesis is the further upgrade of an existing agent-based population model: Whereas the basic population model only allows the prediction of demographic changes for one area, the upgrades presented in this thesis enhance the original model with the ability of simulating the local population development for an arbitrary number of regions, which are connected trough migration. Therefore it is necessary to develop equations describing population movements between these regions.

Doing research on reasons for internal migration leads to many, partly contradictory, motives. None the less the influence of an individual's age and sex on its mobility is a common feature discussed in literature. Therefore the models developed in this thesis manly focus on these influences. Several different approaches are presented and implemented. The first models use decoupled migration processes. To take into account an individual's wish of moving to a particular place and not being distributed to a random region, one additional approach is designed to simulate this situation. All models are further enhanced with a region-specific external migration tool, hence additional parametrisation data are needed. The object-oriented programming language Python 3 is used for the implementation of the models.

Public migration data for the Austrian federal states has been used for the validation process. The results of all implemented versions fit the data well. Especially the additional enhancement with region-specific external migration leads to improved results of the simulations. In addition, forecasting future regional population developments has been attempted.

Kurzfassung

Bevölkerungsentwicklungen bilden schon lange einen sehr wichtigen globalen und lokalen Faktor. Zuverlässige Prognosen können für soziale, wirtschaftliche und politische Entscheidungen, wie zum Beispiel in Bezug auf Standortentwicklungen im Gesundheitssystem, von großem Nutzen sein. Sowohl auf nationaler als auch auf subnationaler Ebene ist die Einschätzung demographischer Veränderungen von Interesse. Durch das Fehlen von harten Grenzen, welche Bewegungsfreiheiten zwischen Regionen einschränken, sind regionale Migrationsraten üblicherweise weitaus höher als internationale Wanderungsraten.

Das Ziel dieser Diplomarbeit ist die Weiterentwicklung eines bereits existierenden Agenten-basierten Populationsmodells. Während das als Ausgangspunkt dienende Modell lediglich die Vorhersage von demographischen Entwicklungen eines Gebietes erlaubt, können durch die im Zuge dieser Arbeit entwickelten Erweiterungen derartige Veränderungen für eine beliebige Anzahl verschiedener Regionen, die durch Migration verbunden sind, simulieren. Dieser Prozess erfordert die Entwicklung von Gleichungen, die Migrationsströme zwischen den jeweiligen Gegenden beschreiben.

Auf der Suche nach Gründen für interne Migration stößt man auf viele verschiedene, teils widersprüchliche, Motive. Nichtsdestotrotz ist quer durch die Literatur der Einfluss von Alter und Geschlecht eines Individuums auf dessen Mobilität ein wichtiger und unbestrittener Faktor. Aus diesem Grund konzentrieren sich die erstellten Modelle hauptsächlich auf diese Parameter. Mehrere verschiedene Ansätze werden vorgestellt und implementiert. Die ersten Modelle basieren auf einem entkoppelten Migrationsprozess. Um die Umzugshistorie einer Person realistischer nachzubilden und sie nicht zufällig, unabhängig von ihrer Herkunft, einer Region zuzuordnen, wird ein zusätzliches Modell präsentiert, welches diesen Anspruch erfüllt. Außerdem werden alle Modelle mit regionsspezifischer Migration in Bezug auf das Ausland mittels zusätzlicher Parametrisierungsdaten weiterentwickelt. Sämtliche Modelle werden mit der objektorientierten Programmiersprache Python 3 implementiert.

Die Validierung wird mit öffentlich zugänglichen Daten für die österreichischen Bundesländer durchgeführt. Für alle umgesetzten Versionen bestätigt sie durchgängig gute Resultate. Insbesondere die zusätzliche Erweiterung mit regionsspezifischer externer Migration führt zu nochmals deutlich verbesserten Simulationsergebnissen. Abschließend wird der Versuch unternommen regionale Bevölkerungsentwicklungen zu prognostizieren.

Danksagung

An dieser Stelle möchte ich mich bei all jenen bedanken, die mich im Rahmen dieser Diplomarbeit und allgemein in meiner Studienzeit begleitet und unterstützt haben.

Zuallererst möchte ich mich bei Professor Felix Breitenecker für die Betreuung dieser Diplomarbeit bedanken. Zudem hat er mir als Teil seiner Forschungsgruppe die Teilnahme an interessanten Projekten, Konferenzen, Summerschools und nicht zuletzt sozialen Zusammenkünften im universitären Umfeld ermöglicht. Dadurch durfte ich einen anderen Aspekt der Institution Universität kennenlernen. Ein großer Dank gilt Martin Bicher für seine geduldige und unglaublich kompetente Unterstützung meiner Diplomarbeit. Seine präzise und schnelle Ortung von meiner Meinung nach unauffindbaren Fehlern in meinem Code hat mich immer wieder fasziniert. Ich möchte ihm des Weiteren für die vielen Stunden danken, die er in das Korrigieren meiner Arbeit investiert hat.

Der Weg bis zum Abschluss meines Studiums durch diese Arbeit wäre mir wohl nur schwer gelungen ohne die jahrelange Zusammenarbeit mit lieben Studienkollegen. Nicht nur die gemeinsame Bewältigung von Übungen und Prüfungen sondern auch das gemeinsame Kartenspielen in der Mensa sind unvergessliche Erinnerungen.

Ohne Familie und Freunde wäre der Großteil meiner Studienzeit ein sehr eintöniger Weg gewesen. Besonders bedanken möchte ich mich bei all jenen, die mir bei Rückschlägen zur Seite gestanden sind und mir wieder Motivation und Selbstvertrauen eingeflößt oder auch einfach nur zum richtigen Zeitpunkt ein offenes Ohr geschenkt haben.

Mein tiefster Dank gilt meinen Eltern, die mir mein Studium erst ermöglicht und mich immer unterstützt haben, sowie meinen Geschwistern Robert und Maria. In Bezug auf diese Arbeit gilt insbesondere meinem Vater und meinem Bruder Robert ein spezieller Dank, die ihr mit ihren fundierten Englischkenntnissen den letzten Schliff verpasst haben.

Zu guter Letzt gilt meiner Freundin Jennifer ein großer Dank, die mir jederzeit mit Rat und Tat zur Seite gestanden ist und nicht nur menschlich sondern auch fachlich eine wichtige Stütze war. Ich konnte mich immer auf ihre Unterstützung und ehrliche Kritik verlassen.

Contents

1 Introduction

Demography plays a very important role in many aspects of governmental and economic planning. Many essential features of our society, such as the health-care system, need reliable data on population development, in this special case e.g. to guarantee an unfailing provision of medical care.

The research field of demography has a very long history [10, p. 313-315]: In around 1700, Edmund Halley, famous for calculating the appearance of the Comet Halley, already created insurance tables based on registered births and deaths in the city of Wroclaw. Using this data he calculated death probabilities. His work was continued in the year 1744 by the Scottish clerics Robert Wallace and Alexander Webster, who developed an insurance for potential widows and orphans of priests. With the help of the mathematician Colin Maclaurin of the University of Edinburgh they evaluated statistical data of births, deaths and average live expectancy. Shortly before they started this research, Jacob Bernuoulli made discoveries concerning statistics and probability theory including his law of large numbers which turned out to be helpful. Equipped with this knowledge and the statistical data they were able to predict that at any given time their church would have 930 clerics, with 27 of them dying on average per year. Considering the emanating numbers of widows and orphans as well as widows getting wed again, they calculated necessary in-payments by the clerics into a fund for the widow's and orphans' pension. They further predicted that by the year 1765 there should be a capital of 58.348 pounds in the pension fund. In 1765 the actual capital turned out to be 58.347 pounds, which means their prediction was incredibly accurate. Wallace' and Webster's pensions office "Scottish Widows" is now one of the biggest insurance companies worldwide.

The impact of (internal) migration on demography came to attention in around 1885 when Ernst Georg Ravenstein gave a talk about "The Laws of Migration" [18]. He concentrated on the United Kingdom where the Industrial Revolution led to a fast growing individual mobility. He came to astonishing results, as for example regarding the higher mobility of women.

> Woman is a greater migrant than man. This may surprise those who associate women with domestic life, but the figures of the census clearly prove it. [. . .] the workshop is a formidable rival of the kitchen and scullery. [18]

Tom Wilson [25] has formulated three answers to the question, why in particular internal migration is worth a detailed examination. Firstly, for understanding the dynamics of a country's population geography, it is necessary to understand internal migration patterns. Especially migration age profiles can emblaze the redistribution of population according to employment, education and lifestyle factors and its age-specific occurrence. The analysis of these profiles is useful for comprehending human spatial mobility across the course of life. Possible intervention in this process by policies can be determined. Secondly, to compute population estimates in years where no census takes place, reliable (internal) migration data are essential. Wilson thus argued that "internal migration is often the most important demographic variable shaping regional population age structures." Thirdly, multi-regional demographic prospects by single age-classes of course need precisely modelled migration profiles by single age-classes.

Using an internal migration model for the European Union, Van Gaag and his colleges in 2000 compared a scenario with high economic and population growth and a strong convergence of regions, i.e. differences between countries and regions becoming smaller over time, with a scenario of low growth and constant or weakly converging regional differences. They discovered an impact of internal migration of up to 30% on the difference of the total regional population in the two scenarios. Therefore considering internal migration makes a difference in total population growth at the regional level [23]. Figure 1.1 illustrates this issue for the Austrian region of Vienna using actual results of this thesis.

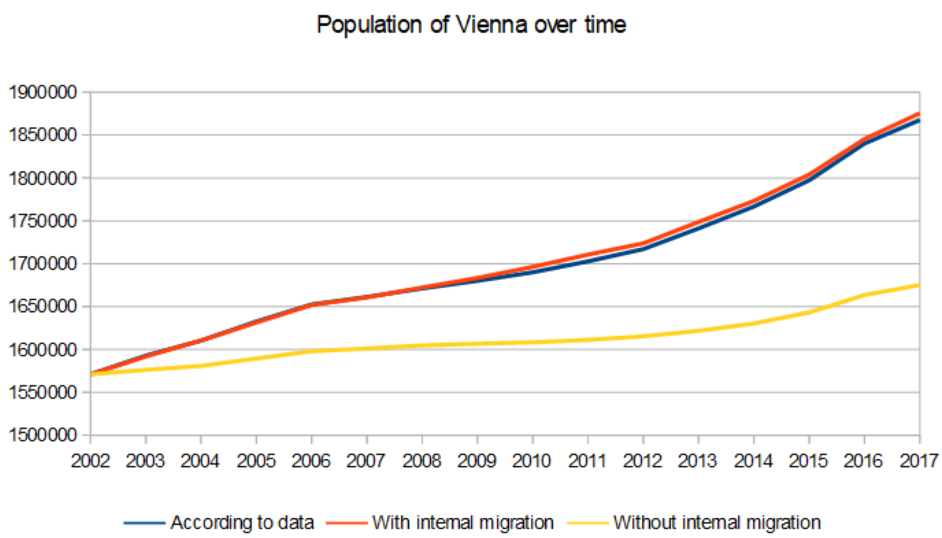

Figure 1.1: Population of Vienna as part of a simulation for Austria as a whole over the years. The modelling approach with respect to internal migration has a very good fit on the regional level. The simulation results presented were computed by the basic GEPOC and the *migration pool model with region-specific external migration* which will be introduced later on.

To embrace the influence of internal migration on regional population developments, this diploma thesis further investigates the impact internal migration has on a country's demography and how to model it in the most fitting way with respect to the data available for the regions in question. Therefore different model approaches will be presented and tested and their results compared. The models were implemented using the object-oriented programming language Python 3, altering the given basic model which will be presented in the next chapter. Among other advantages like free accessibility and an enormous pool of freely available Python packages, this language is capable of dealing with the high number of required agents.

2 Modelling of population development

By constructing a model, a usually simplified image of any point of interest is built. In 1975 Robert Shannon formulated a definition of a simulation model:

> Simulation is the process of designing a model of a real system and conducting experiments with this model for the purpose either of understanding the behaviour of the system and its underlying causes or of evaluating various designs of an artificial system or strategies for the operation of the system. [21]

A population model aims at a sufficiently accurate simulation of the development of a population group over a given period of time. Since the progress of time is an important part of such a model, naturally the ageing of the observed population is a point of interest. Three mathematical models for age-structured population developments have been introduced by Barbara Lee Keyfitz and Nathan Keyfitz [13] and will be presented in the following. Their basic assumptions are that on the one hand every person can either go on to the next year of age or die and on the other hand at certain ages a woman can give birth to a child aged zero. They suggest to express these postulates as "life-table data", where $l(a)$ is the probability of surviving from birth to age a, and $m(a)$ the probability for a woman aged a years, to give birth in a certain time unit. To keep the models simple, Keyfitz and Keyfitz assume a single-sex, female population model.

The matrix model

Based on so-called Lexis diagrams, where lifelines are represented two-dimensionally in a space with coordinate axes of age and time, the matrix formulation reflects the relation between two horizontal lines, representing time-steps. Let k be the number of age classes and $\mathbf{v}_t \in [0,1]^k$ a vector where $\mathbf{v}_t^j$ is the amount of women in age group j at time t. The population's age distribution at time t is pre-multiplied with $\mathbf{M}$ to generate the age distribution at time $t+1$.

$$\mathbf{v}_{t+1} = \mathbf{M}\mathbf{v}_t \tag{2.1}$$

The entries of the projection matrix $\mathbf{M}$ are determined from the former mentioned $l(a)$ and $m(a)$.

$$m_j := \int_0^1 m(j+s)\,\mathrm{d}s \quad \text{and} \quad L_j := \int_0^1 l(j+s)\,\mathrm{d}s \tag{2.2}$$

By defining m_j and L_j as discrete approximations to the continuous quantities like in equation (2.2), $\mathbf{M}$ can also be written as $\mathbf{M} = \mathbf{S} + \mathbf{B}$ where the sub-diagonal matrix $\mathbf{S} = (s_{ij})$ is called "survivor ship matrix" and calculated as shown in equation (2.3).

$$s_{j+1,j} = \frac{L_j}{L_{j-1}} \tag{2.3}$$

The non-zero entries $s_{j+1,j}$ represent the fractions of individuals surviving the time-step from $j-1$ to j. Therefore the first row of $\mathbf{M}$ equals the first row of the "birth matrix" $\mathbf{B} = (b_{ij})$ and is calculated as

$$b_{ij} = \frac{L_0}{2}\left(m_{j-1} + \frac{L_j}{L_{j-1} m_j}\right). \tag{2.4}$$

These are also the only non-zero entries of $\mathbf{B}$. They are calculated supposing that women, who are alive between t and $t+1$ and part of age group j, have a birth rate of m_j for the first half of the time interval and, if surviving, m_{j+1} for the second half.

The integral model

Be $m(a)$ a function with compact support in the interval $[\alpha, \beta]$, which represents the average fertility period of a woman. Therefore the number of births at a given time $B(t)$ depends on the amount of mothers being born α to β years earlier. Summing up results in the *Lotka formulation*

$$B(t) = \int_\alpha^\beta B(t-a) l(a) m(a) \,\mathrm{d}a + G(t). \tag{2.5}$$

$G(t)$ is the aggregation of given births at time t by women being part of the initial population. Examining the homogeneous equation leads to the exponential solution

$$B(t) = Qe^{rt} \tag{2.6}$$

and therefore equation (2.5) becomes

$$Qe^{rt} = \int_0^t Qe^{r(t-a)} l(a) m(a) \,\mathrm{d}a \tag{2.7}$$

or simplified

$$1 = \int_0^t e^{-ra} l(a) m(a) \,\mathrm{d}a. \tag{2.8}$$

A discrete a and growth rate $\lambda := e^r$, yield the discrete time equation known as *Euler-Lotka equation*

$$1 = \sum_{a=1}^{\omega} \lambda^{-a} l(a) m(a) \tag{2.9}$$

with ω being the maximum age. On the basis of the homogeneous solution, Lotka managed to solve the inhomogeneous equation, too, including several arbitrary constants and assigning them to fit the $G(t)$.

The partial differential equation model

A partial differential equation for an age-structured population can be formulated as

$$\frac{\delta P}{\delta t}(a,t)+\frac{\delta P}{\delta a}(a,t)+\mu(a)P(a,t)=0 \qquad a,t\geq 0 \tag{2.10}$$

with $P(a,t)$ representing the population density at age a and time t. This approach is called the *McKendrick form* and represents *local* effects of population dynamics.

> This equation can be seen as a representation of the local effects of population dynamics; it is derived from the fact that, since age and time are measured in the same units, the rate of change with respect to time of the size of the population of age a would exactly balance the derivative with respect to age, were it not for the [...] instantaneous death rate, $\mu(a)=-\frac{l'(a)}{l(a)}$. [13]

Equation (2.10) is solvable using the method of characteristics,

$$\frac{\mathrm{d}P}{\mathrm{d}s}(a,t)=\frac{\delta P}{\delta a}\frac{\mathrm{d}a}{\mathrm{d}s}+\frac{\delta P}{\delta t}\frac{\mathrm{d}t}{\mathrm{d}s}, \tag{2.11}$$

with $a=a(s)$ and $t=t(s)$. The detailed solution can be found in [13] and leads to the Lotka formulation (2.5). Therefore the approach via the McKendrick equation is rather conceptual than practical, with the advantage of being more transparent.

2.1 The Generic Population Concept (GEPOC)

In 2015 a generic population model, capable of producing a valid virtual image of Austria's population and feasible prognoses, only using public accessible initial population data, was completed as part of the Decision Support for Health Policy and Planning (DEXHELPP) health-care research project. Most information on this population model is found in the unpublished DEXHELPP project report "Generic Population Handbook" which is freely available on demand and in [3]. An up to date publication "GEPOC ABM: a generic agent-based population model for Austria" is currently in print and will be published in December 2018.

Amongst other things, the GEPOC model was implemented as an agent-based model which will be further developed as part of this thesis. A technical description of the GEPOC model will be given in Chapter 5. To facilitate the further understanding of the basic agent-based model and its developments concerning internal migration, an introduction into agent-based modelling will be given in the following chapter.

3 Agent-based modelling

Agent-based modelling (ABM) was developed in the eighties of the last century and got rather popular in the nineties. Thanks to its comparatively late development as a simulation method, ABM incorporates more modern features like highly potential computers or the internet [20]. While former models considering the macro-level only had homogeneous entities, ABM was the first to provide the opportunity of heterogeneous agents with individual internal models [9]. This is made possible trough the fusion of artificial intelligence and computer science. The result is a massive improvement of modelling, visualisation and implementation of computer systems in theory and practice. In Figure 3.1 the main points of an agent-based model are illustrated. One can see the environment in which the agents are moving and the upper level, where interactions take place, including the building of organisational relationships [12].

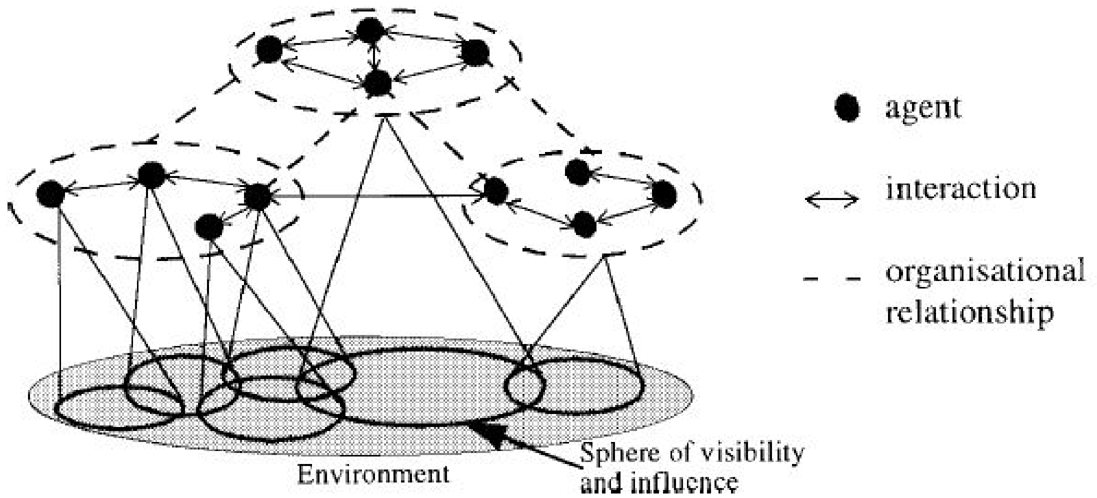

Figure 3.1: Schematic illustration of an agent-based system [12].

As visible in Figure 3.1, interaction between two of the groups is sufficient to get an information flow to all of the agents within them. Each agent can take its own decisions, depending on some rules it received on start-up and the situation it is in [4]. To reach such strategic decisions, an agent might consider whole internal models of other agents. The following actions can either take the form of information transfer or physical processes. The quality of the simulation is strongly influenced by size and content of these internal models and the amount of links an agent can handle. The development of game theory played an important role in enabling the decision process of agents [9]. An ABM usually contains more than one type of agent to simulate the problem stated. If, for example, a model of a railway system it intended, it might not be sufficient to use whole trains as agents but different types of wagons and locomotives or even workers and passengers. Information flow within the simulation can happen between all these completely different agent types.

3.1 Features of agent-based modelling

Some possible characteristics of agents are stated by Epstein [7]:

- Heterogenity: In ABM populations can be heterogeneous. Each agent is able to possess unique characteristics like culture, preferences or life expectancy. In addition each individual can have their own network. All features might change during the simulation.

- Autonomy: Individual behaviour is not controlled by a central or hierarchic authority. Of course there is feedback within the structures and new born agents will get some start-up information but afterwards they will be able to decide further actions on their own.

- Explicit space: As already seen in Figure 3.1 usually all events take place in one explicit environment.

- Local interactions: If agents interact, they typically do not exchange information with random agents but with neighbours in the environment.

- Bounded rationality: Many ABMs do not equip their individuals with global information or unlimited computational power. Therefore they often have to make decisions using simple rules, based on restricted, local information.

Therefore ABM is particularly suitable for systems with bounded space, where independent, heterogeneous persons with limited information live, move and interact locally [7]. Interaction can exceed simple information exchange by far. There can be requests from one agent to the other to fulfil certain actions, to collaborate, coordinate tasks and undertake business to reach common goals [12]. Repeated and occasionally competing interactions between the agents can be a particular advantage. During the simulation they are able to reflect dynamics which would not occur using purely mathematical methods. Therefore a simple ABM is able to illustrate comparable complex behaviour and give precious information about the dynamics of the simulated “real” system [4].

Advantages and disadvantages of using ABM

Compared to other simulation approaches Bonabeau states three major differences [4]:

ABM captures occurring phenomena. Newly emerged phenomena are the consequence of the interaction between the different entities. Since agents are constructed to solve problems flexibly and to move within an environment of which they only have partial grasp and control, interactions have to happen in a similarly flexible way. Therefore they can make decisions concerning their behaviour *during* the runtime of the simulation. Agents can execute or react to actions that were not even known when they were created [12]. In this simulation approach, the model cannot be reduced to its separate units. It has to be considered that the interactions between the agents are a very important part of the simulation. Out of the behaviour patterns of each individual, completely new situations can emerge. For example the behaviour and the interaction of the different road users might cause a traffic jam. ABM is a very practical modelling approach for arising phenomena. The behaviours of the parts (agents) of the system are modelled and simulated, including the interactions among them, while the resulting events are captured during the whole simulation [4]. In conclusion, trough the construction of agents with certain rules, emerging behaviour patterns, like a traffic jam, can be predicted.

ABM allows a natural description of a system. When a system of entities with certain behaviour patterns has to be described and simulated, ABM frequently turns out to be the most natural approach. Be it the structure of an organisation, movement sequences or a stock market, this simulation type matches reality best. It seems natural to simulate the motion of customers in a shopping mall rather than constructing equations that should display the dynamics of the shopper density. Since these equations result from the consumer's behaviour, with the agent-based approach the results of equation based models even can be generated out of the information gained on the manner. In addition ABM allows using the full potential of data a company has about its customers, as information about the actual behaviour of individuals can be gained out of recorded data and surveys [4]. Virtual agents can be generated with an individual shopping basket for example, representing an actual purchase, in contrast to baskets filled with products bought on average, like three quarters of a litre of milk.

ABM is flexible. The flexibility of this simulation approach emerges in multiple ways. For example, it is very simple to add additional agents to an agent-based model. Additionally there is a natural framework to extend the complexity of the agents. Behaviour, scale of rationality, ability to learn and develop, and the rules of information exchange and the handling of other agents can be enhanced any time. Flexibility can be found in the possibility of changing type and grouping of the agents. So whole groups, subgroups or individuals can be edited and exist next to and with each other in an existing model, despite different types and levels of complexity [4].

With these features, ABM is well suited for models where the depth and type of simulation which will finally be needed is not clear from the beginning and only becomes apparent during the development of the model [12]. According to Bonabeau, the feature of dealing with newly emerging phenomena is the biggest advantage offered by ABM [4].

Still, ABM has the disadvantage that for an actual model of e.g. human behaviour, factors like irrationality, subjectivity and complex psychology have to be considered. Thus, this quantification, calibration and validation of the model, as well as its construction, implementation and interpretation is complicated. As a consequence the computation can be very costly [8]. Clearly with increasing complexity of agents and their behaviour in a model, these processes tend to become more difficult. As the GEPOC ABM needs to simulate quantitatively reliable results, it uses rather simple agents with very limited behaviour (compare Section 3.3).

3.2 How to build an agent-based model

The following features stated by Hannapi in 2017 have to be considered for the implementation of an ABM by a computer program [9]:

- The program has to guarantee the following sequence in the agent's structure: perception, embedding in an internal model, choice of action (including communication).
- Since heterogeneity is an essential feature of ABM, the programs representing the agents usually have to be different in some way.
- A main program is required, where the environment is defined. Here the agents can execute their actions.

He also gives a "short recipe on how to cook" a classical agent-based model:

1. *Choose a topic.* Here the challenge consists of keeping the topic closed enough, concerning its relation to the environment, to be able to produce a valid model in regard to its internal dynamics. The information flow from outside the model should be much larger than the influence of the model on the environment. Therefore sticking to important elements and their interdependence is essential. In addition a time frame has to be set.
2. *Identify the major agents.* One has to consider equipping every agent introduced with a minimum of one goal variable and one instrument variable. That way its embedding within the model is defined. Other variables used might be auxiliary variables, to assist in the formulation of relationships, or exogenous variables, as a path to the outside world.
3. *Construct the internal models.* The smallest internal model of an agent would consist of a basic relationship between the obligatory goal and instrument variable

of step 2. Usually these models are more sophisticated, considering many different variables and connections. The internal model is the place where heterogeneity can be achieved. Here it is decided which information reaches the agent and how the agent handles it.

4. *Empirical data.* Theoretically (initial) data for all variables used has to be provided. For some models this might be impossible to achieve, or at least very difficult. Acquiring data concerning the internal decision process might be a problem in particular. Therefore in some cases, assumptions will have to be taken.

5. *Estimation and calibration.* The relationships introduced within the model now have to be estimated and calibrated based on the data found. The quality of this process strongly depends on the data provided. In extreme cases it might be necessary to guess certain parameters.

6. *Software implementation.* There are multiple programs and software tools available for ABM. Consider to implement the model in a way that allows easy modification, as now steps 1 - 6 might be iteratively repeated and improved. Especially using modules could give a big benefit in reviewing and editing the model.

7. *Systematic results generation.* Execute lots of simulation runs to collect the hereby generated data. Interpret and report the results.

3.3 GEPOC agent-based model

As an example for an agent-based model and further introduction into the GEPOC ABM its definition is explained in this section. The model is defined by its initial setup and time-dynamics:

Initial setup. Once a simulation start date is fixed, an agent-based model with $N + 1$ agents is initialized. N of them represent the inhabitants of Austria. Each of these agents can be imagined as representative of a real person and receives an individual birth-date and sex (male or female) as well as a unique ID. The empirical data used for the initial population is provided by the Austrian Bureau of Statistics [1]. The additional agent shall be referred to as *government-agent*, as it will play the role of the government.

Time-dynamics. The model is updated in a-priori defined time-steps which are not necessarily equidistant. For every time-step all agents representing individuals are iterated in random order. For each one the model decides if they die, emigrate or (in case of a female agent) give birth to a new agent, using an event-based strategy. Figure 3.2 illustrates such a simulation step where the decisions of each agent are made. In case of death or emigration the agent is removed from the model and possibly later scheduled events are skipped. The birth event leads to the construction of a new agent with a birth-date according to the schedule.

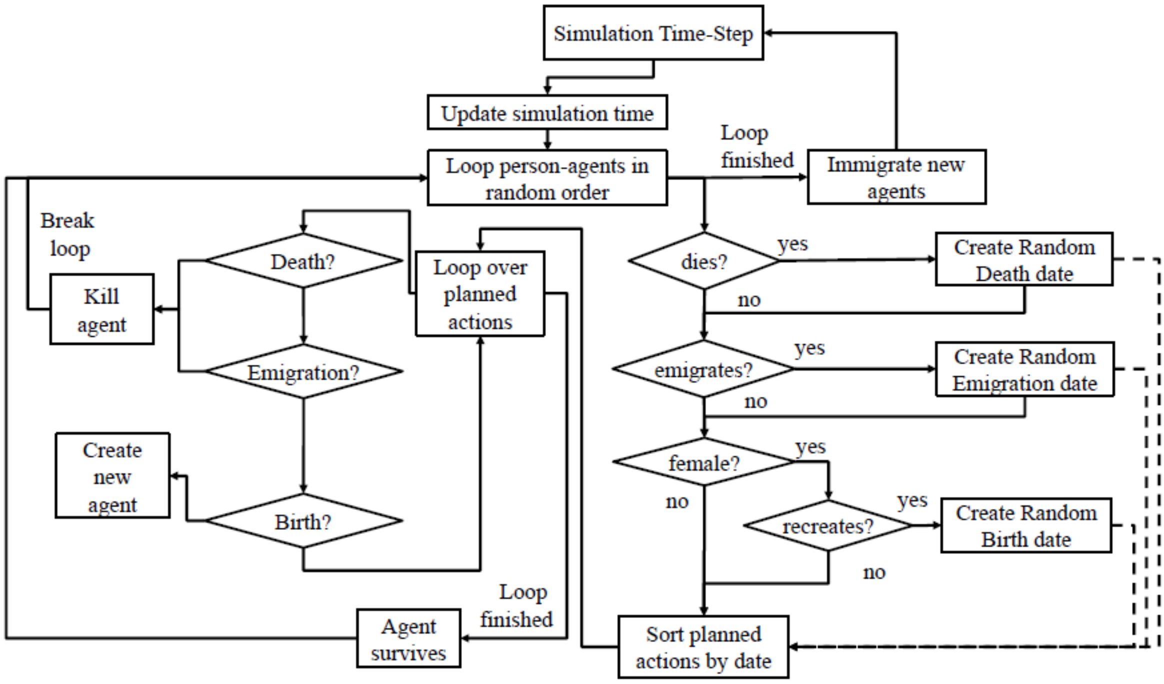

Figure 3.2: Discrete-event motivated strategy for updating person-agents in the GEPOC model during one time-step of the simulation.

> After the iteration is completed, the government-agent generates a given number of immigrants and adds them to the model as new agents.

The GEPOC ABM was implemented with Python 3. After the systematic generation of results, they have been validated using prognosis data of the Austrian Bureau of Statistics [1]

4 Mathematical models for internal migration

In this chapter different model approaches for handling internal migration will be presented. Table 4.1 lists the abbreviations used within the model equations. Henceforth (if not stated otherwise) migrant, immigrant and emigrant always refers to internal migration.

Various reasons and factors have to be considered when doing research on internal migration. For example Aude Bernard, Martin Bell and Elin Charles-Edwards concentrate on the life-course transitions that affect migration choices – especially of young adults – directly, calling them “proximate determinants.” They state that economic, social and other rather general factors, shape people’s plans for their lives and therefore lead to migration age profiles. The suggested framework can be found in Figure 4.1. Age profile differences between countries occur trough different timing of entry into education, labour market entry, partnership and childbearing [2]. The approach presented by them will be further explained in Chapter 4.1. Wilson and Bell state that multi-regional models with fixed migration rates tend towards dampening the net migration gains in fast growing regions and therefore causing convergent regional growth rates. The reason for this is that in such a fast growing region the potential number of emigrants is increasing at the same speed as the population, whereas the pool of possible immigrants cannot keep that pace. In addition, model approaches, where migration flows are influenced by both origin and destination, deliver better representations of migration behaviour compared to those depending merely on the original population size in the base year [26].

Petra De Jong and her colleges use age-dependent migration propensities, too. For their modelling approach of internal migration in the Netherlands, which is manly characterised by its urban communities, they use a residential density index. Here human activities regarding everyday life, work and leisure serve as indicators of urbanisation. The majority of residential mobility takes place within one region. In their paper they concentrate on migration flows between urban hierarchy levels. However, all interregional migrants in the Netherlands, except for the age group of $18-24$, tend to move towards less populated municipalities. With increasing age, people prefer to move away from highly populated areas. Despite these trends, parts of the age cohort between 65 and 74

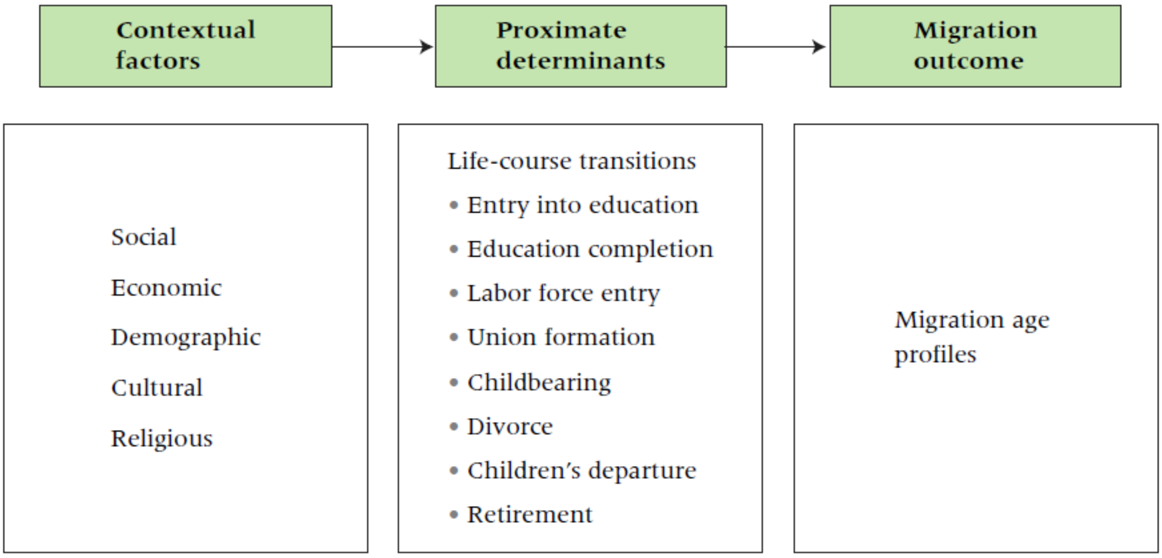

Figure 4.1: Important factors for generating age profiles of internal migration according to Bernard, Bell and Charles-Edwards [2].

have a tendency of moving upwards in the urban hierarchy, which might be explained by the retirement of the baby boom generation, which would not stick to traditional notions of retirement mobility. They conclude that therefore in the future a change in the patterns of residential mobility might have to be considered. The results found for the Netherlands are highly compatible to movement trends in the United States of America although the urban hierarchy levels and geographical conditions are very different [6].

Bruce Newbold [15] presents a "whole topology of return migration" for Canada considering individuals who move back to their initial region. Reasons for that are on the one hand planned returns after finishing education or employment and on the other hand "failed" migration. According to him little evidence suggests that return migration can be connected with the level of education. The reasons for return migration following a disappointing initial migration are rather obvious: the region of origin is usually well known, family and friends are available and therefore minimize the cost of return. The group of onward migrants introduced by Newbold usually is better educated and employed in more professional, skilled jobs. This migration type might be in search of better opportunities in different places. All forms of migration share the attraction by areas with higher per capita disposable incomes, with the restriction that return migrants might accept some economic drawback in exchange for support by family and friends and familiar surroundings. In addition, cultural similarity in the province of birth increases the trend of return migration whereas anonymity and distance have a negative influence. Interestingly, Newbold found no age effects in return migration decisions.

Nicole Van der Gaag and her team found a stable positive relationship between GDP per capita and internal migration. This relationship has turned out to be similar for different countries. In addition, the amount of new housing units built has an effect on internal migration [24]. However, including additional factors like welfare payments, unemployment insurance or government expenditures as a whole is not very helpful for forecasting studies, since forecasts have to be made for these variables, too. The improvement of the model by such additional information is often very limited or not happening at all. It may even happen that the results get worse. Andrew Isserman and his colleagues present a migration model based on the human capital theory, where internal migration is interpreted solely as an investment in monetary terms. Therefore, the expected income for each region has to be estimated. They suspect that this is the commonly used approach to migration analysis in the U.S. Still, during the examination of all interstate migration flows of the U.S. for 1975–1976 it turned out that in over 60% of the cases, net migration occurred in the direction of the state with lower wages (in contrast to the assumption made). They concluded that having employment trumps the expected income concept [11]. If economic prosperity is given, non-economic factors may increase in comparison to economic influences since higher incomes allow people to pay more attention to the quality of life when making decisions regarding migration [14].

Economic growth leads to higher mobility in all interregional migration flows. For the Netherlands the mobility curve closely resembles the economic growth rate [23]. On the other side the increase of female labour participation and as a consequence more dual-worker households reduce internal migration since it is more difficult to find two new appropriate jobs instead of just one. The ageing of the population in industrial states diminishes migration rates, too, since older workers change jobs more seldom [24]. The impact of (regional) governmental decisions must not be underestimated, although considering such actions in advance is hardly possible. The migration of children could be forecast by basing it on their parents behaviour. So an increase of people in the age group of 18–44 moving from i to j would lead to an increased movement of children [11].

As shown in the last paragraphs there are multiple factors for internal migration, of which many have to be examined critically with respect to the accuracy of their own predictions. The models implemented and presented below require demographic data available from most national (European) institutes for statistics or Eurostat, respectively [5].

Net migration versus multi-regional models.

If migration should be modelled, there are two possibilities to handle the task. The first option is to calculate the net migration for every region, whereas the second one is to model the whole emigration and immigration flows between all regions of interest. Net migration models have the big advantage of needing little data compared to the other approach. Nonetheless, especially from the agent-based point of view, they have major disadvantages. As Andrei Rogers states in *Requiem for the Net Migrant* [19] individuals that fit the pattern of so-called "*net migrants*" do not exist. In addition, there is the problem of fixed (out)flows of individuals from one region, which can lead to negative

Variable	Description
T	initial time of the simulation
t	actual time of the simulation
P	population
M	migration flow between two regions
EM	emigrants from one region
IM	immigrants to one region
Pool	migration "pool", sum of all migrants
m	migration rate between two regions
emr	emigration rate from one region
imr	immigration rate to one region
p	distribution proportion of immigrants
Subscripts	
i	origin
j	destination
s	sex of the population observed
a	age of the population observed

Table 4.1: Notation used in the model equations based on Wilson and Bell [26].

regional population figures in extreme cases.

Multi-regional models on the other hand allow migration streams in both directions. Here agents would be allowed to move to any region even if the net migration of the area chosen were negative. Changes in the sex/age structure, the size of the population or even geographical distribution of the agents can be considered due to the dependence of the inter-regional flows on the population [26]. Since one aim of modelling is to imitate reality as well as possible, this work only considers multi-regional models.

4.1 Migration age profile model (MAP)

Various scientists who have examined internal migration age profiles in different countries, have often found a high degree of regularity over space and time for migration age profiles , although huge variations in levels of migration occur. Migration probability age profiles have to be smoothed to achieve the age patterns without suffering from projecting noise and implausible population age-structures [25]. Therefore the big advantage of a MAP model is the small amount of additional input data for internal migration. It only consists of one matrix per gender, where migration flows between all regions (including the region of origin) for the required time are listed. They do not have to depend on age since a general age profile is assumed. Such a typical projection curve is shown in

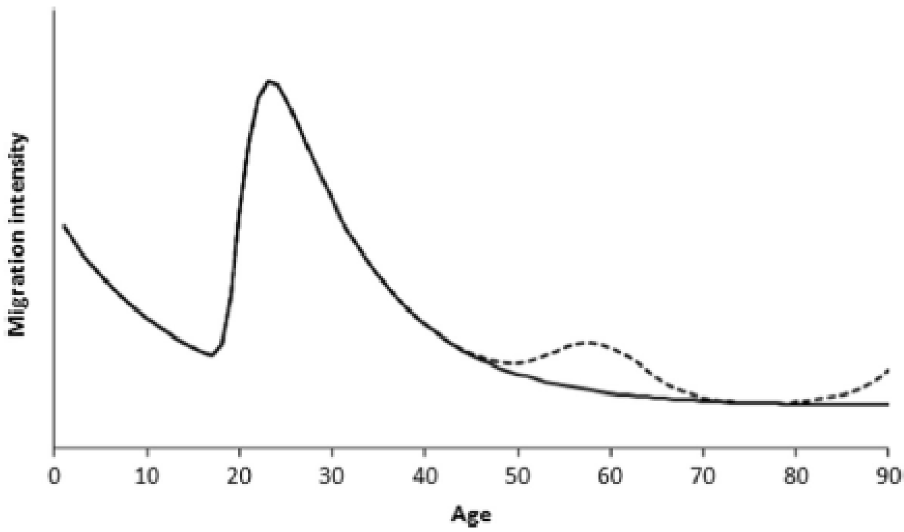

Figure 4.2: Typical age-dependent variation in migration intensities according to Wilson [25]. Note: The dashed lines represent migration intensities which occurred only in some migration age profiles.

Figure 4.2. One can see that the migration propensities per age undergo a high variation. Young adults form the most mobile group. Mobility peaks between twenty and thirty and declines steadily afterwards, with an eventual rise around retirement age and again shortly before death. Many authors consider this strong empirical regularity as almost universal, which is shown by the wide use of Tom Wilson's *"age schedule of migration"* [6].

Reasons for the peak at a young age are the comparable little (space-bound) commitments of people in that group. They can move to find employment, higher education, a partner or simply adventure. Getting older, a good job, mortgages, children who already entered school and social networks steadily reduce their mobility. Since they are dependent on their parents' residency, the reduced mobility of the children is explained. The reason for the sometimes occurring small peaks in Figure 4.2 is firstly that people around retirement do not have to consider their place of work or the education of their children when choosing a residence anymore. Therefore they might move to places they always dreamed of or lie closer to their family. Secondly, at the oldest ages people might move to retirement homes or places that offer support [25].

Aude Bernard, Martin Bell and Elin Charles-Edwards state a very similar age profile of migration including the key reasons for transitions in the life-course which can be seen in Figure 4.3. Even though especially Wilson's pattern is widely used and its regularities are undoubtedly persistent, cross-national studies have shown systematic variations in the age profile, especially regarding young adult ages. Although there has been a lot of work on reasons that trigger migration, Bernard, Bell and Charles-Edwards were the first to investigate how these causes interact in shaping the migration age profiles. According to them, it is necessary to distinguish between common factors

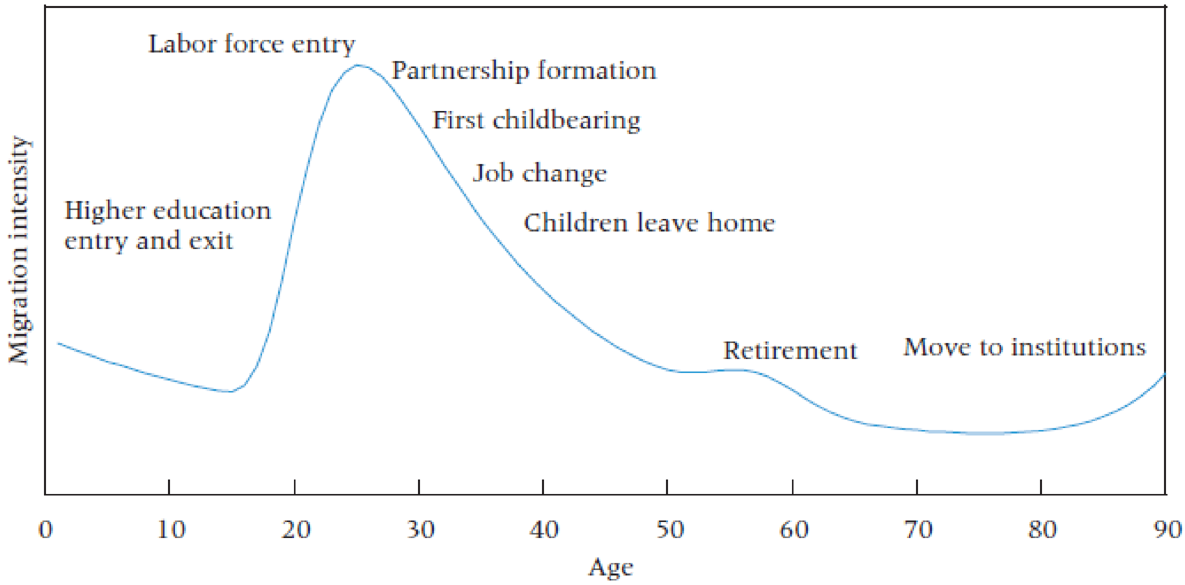

Figure 4.3: Typical age-dependent variation in migration intensities according to Bernard, Bell and Charles-Edwards including the main reasons for life-course transitions [2].

across countries and country-specific ones. They identified four key transitions at young adult ages for marking the passage to adulthood, namely finishing education, entering the labour force, forming a union and giving birth to a child. Taking the next step they tried to link these factors to the peak in age profiles of migration intensity occurring between twenty and thirty. With the study and comparison of the developed framework in 27 countries around the world, they concluded that in areas like Europe or Northern America, where these transitions happen at comparatively older ages, migration takes place later in life, too (compare Figure 4.4). In addition, in these countries the peak is more widely distributed across the age range [2].

A disadvantage of generalised migration age profiles is, according to David Plane, that fixed origin-destination migration rates cannot be justified on behavioural grounds, since true dynamics of regional demographic change are not reflected by them [17]. Figure 4.5 shows the existing regional differences in age-dependent migration intensities for Austria. Since this work concentrates especially on internal migration and the differences between the regions concerning a migration age profile, this approach has been rejected and replaced by improved models with higher accuracy.

4.2 Biregional model (BR)

The first well-functioning modelling approach for internal migration presented in this thesis is the construction of a so-called biregional model, based on the one presented

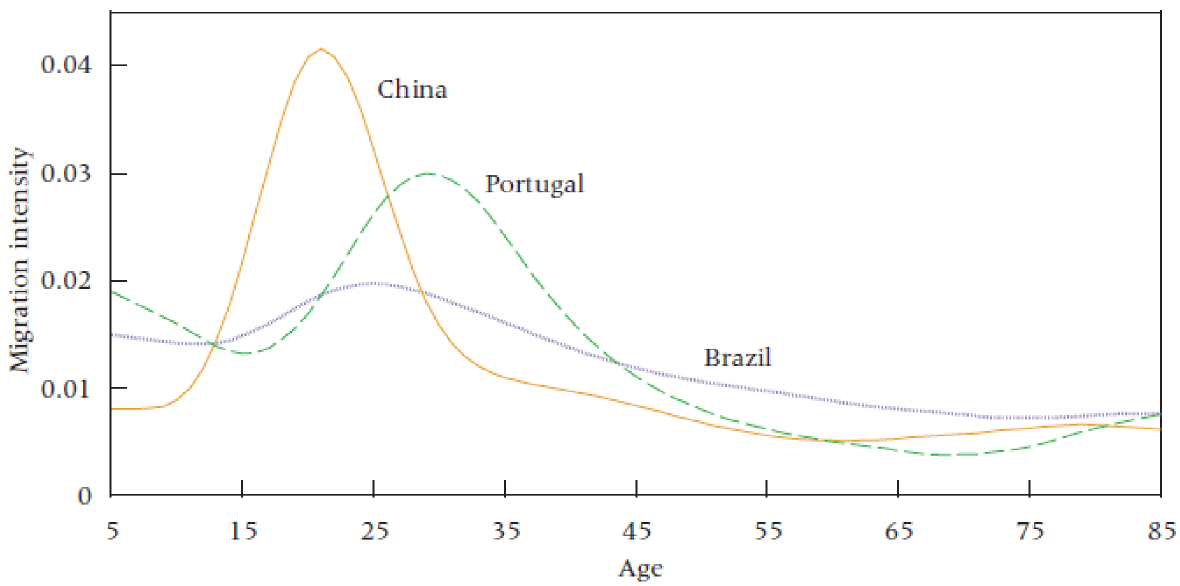

Figure 4.4: Nation-dependent variations in migration age profiles illustrated by the examples of China, Portugal and Brazil [2].

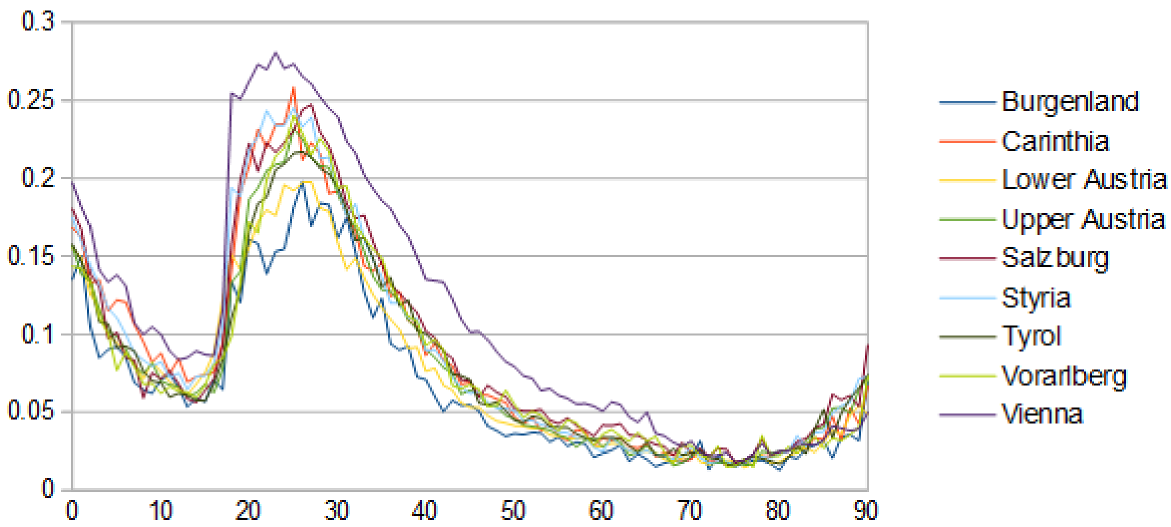

Figure 4.5: Differences in age-dependent migration intensities between the federal states of Austria in 2017 according to Statistik Austria for ages 0–90 [1]. Although a common tendency is visible, regional differences cannot be overseen. Especially Vienna shows higher internal migration rates due to its high population density.

by Wilson and Bell [26]. The additional data needed for the internal migration flows consists of initial population as well as emigration and immigration data for every region partitioned into sex and age. In fact, this approach is the combination of several biregional models instead of one whole multi-regional model. It is assumed that the country in question is divided into one region of interest and a combination of all other regions. Then emigration and immigration rates are calculated for this particular region. This process is repeated for every region in the simulation.

$$\mathrm{emr}_{i,s,a} = \frac{\mathrm{EM}_{i,s,a}(T)}{P_{i,s,a}(T)} \tag{4.1}$$

$$\mathrm{imr}_{j,s,a} = \frac{\mathrm{IM}_{j,s,a}(T)}{\sum\limits_{i \neq j} P_{i,s,a}(T)} \tag{4.2}$$

The emigration rate corresponds to the total emigration of the region divided by the total population of the region at start time T of the simulation. Wilson and Bell calculate the immigration rate as the total number of immigrations to a particular region from all other regions combined, divided by the total population of the country at start time T except for the one area of interest. Compare Table 4.1 for an explanation of the abbreviations used in the equations. Since in the constructed model internal migration within one region should be included, equation (4.2) is changed to:

$$\mathrm{imr}_{j,s,a} = \frac{\mathrm{IM}_{j,s,a}(T)}{\sum\limits_{i} P_{i,s,a}(T)} \tag{4.3}$$

Note that in equation (4.3) the whole population is used and therefore also immigration within one area is considered.

For the implementation of the biregional model – since the GEPOC model is agent-based – it has to be evaluated for every agent if they emigrate. With that knowledge the calculation of the total amount of emigrants, which naturally has to be the same as the total amount of immigrants for internal migration, is possible. To ensure that the overall net internal migration sums to zero, it is necessary to rescale the calculated immigration rate. Now all emigrating agents can be distributed to their destination. This process is repeated for every time-step, leading to a very costly procedure. Therefore in the second approach the internal simulation was implemented via a migration pool model.

4.3 Migration pool model (MP)

Another model based on the paper of Wilson and Bell [26] is the so-called migration pool model. It is very similar to the BR model. In theory it is executed in two steps: first the number of all emigrants from all regions is evaluated and the migrants concerned are put into a common "pool." Second they are divided into different regional destinations.

The big difference to the BR model is that this division is not based on the population but on the total number of migrants. The amount of emigrants per region has to be calculated according to equation (4.4).

$$\mathrm{EM}_{i,s,a}(t) = \mathrm{emr}_{i,s,a} \cdot \mathrm{P}_{i,s,a}(t) \tag{4.4}$$

Note that the emigration rate is calculated analogous to equation (4.1) in the BR model. Now the pool for every sex/age group at time-step t can be created out of all these emigrants

$$\mathrm{Pool}_{s,a}(t) = \sum_i \mathrm{EM}_{i,s,a}(t). \tag{4.5}$$

To obtain the total number of immigrations to each region, distribution proportions have to be calculated in the form of immigrants of one region separated by sex/age divided by total immigrants of the corresponding group

$$\mathrm{p}_{j,s,a} = \frac{\mathrm{IM}_{j,s,a}(T)}{\sum\limits_j \mathrm{IM}_{j,s,a}(T)}. \tag{4.6}$$

Now the amount of immigrants for each destination can be computed in equation (4.7)

$$\mathrm{IM}_{j,s,a}(t) = \mathrm{Pool}_{s,a}(t) \cdot \mathrm{p}_{j,s,a}. \tag{4.7}$$

Equation (4.8) depicting the migration flow between two regions, shows the major advantage the MP model has compared to the BR model in an agent-based approach.

$$\mathrm{M}_{i,j,s,a}(t) = \mathrm{emr}_{i,s,a} \cdot \mathrm{P}_{i,s,a}(t) \cdot \mathrm{p}_{j,s,a} \tag{4.8}$$

Since the migration flow is calculated as the product of two fixed rates and the group-specific population of one region, this results in an agent's probability m to migrate from region i to region j

$$\mathrm{m}_{i,j,s,a} = \mathrm{emr}_{i,s,a} \cdot \mathrm{p}_{j,s,a}. \tag{4.9}$$

Under these assumptions the migration probability for one agent is not dependent on time except for their increasing age pushing them into another group. In addition, it is not necessary to actually create a pool of emigrants in an intermediate step. Therefore during the simulation the whole migration process of one individual can be handled in one quick step. A point of criticism of this model is that migrants can return to their initial region during one migration process. Since an individual's migration within one region is a realistic scenario which is intended to be part of the simulations of this work, this objection is irrelevant.

To conclude, in the MP model immigration into regions is only dependent on the pool size and not the composition of the pool by region and origin [23].

4.4 Inter-regional migration model (IRM)

All models presented above have in common that emigration and immigration happen independently, thus there is no connection between origin and goal region. The fact that an individual emigrates from region A or B has no influence on their decision where to immigrate to afterwards. Although this has no effect on the number of migrants and the total population, from an agent-based point of view it should be considered. Since individuality is the unique feature of ABM, the agents in an internal migration model should decide where to immigrate based on their initial region. The inter-regional migration model reaches this goal by changing the internal migration process. While the decision to emigrate is still dependent on the emigration rate analogous to the BR and MP model, for immigration decisions age is not incorporated. Instead data which represents movement from every initial region to all goal regions is included. Thereby, while the agent's age is still important since it influences its emigration decision, the agent's destination is dependent on its origin.

Formalising this leads to an inter-regional migration rate for each sex which can be seen in equation (4.10):

$$\mathrm{p}_{i,j,s} = \frac{\mathrm{IM}_{i,j,s}(T)}{\sum\limits_{j} \mathrm{IM}_{i,j,s}(T)} \tag{4.10}$$

Note that now the distribution and number of immigrants to one region depend on their origin instead of their age. The new migration rate between two regions thus is:

$$\mathrm{m}_{i,j,s,a} = \mathrm{emr}_{i,s,a} \cdot \mathrm{p}_{i,j,s} \tag{4.11}$$

4.5 Enhancement with region-specific external migration (rem)

Until now all models have concentrated on internal migration only, thus external migration is handled for the conglomerate of all regions within the model. Emigration to a foreign country happens on the basis of a common probability for all agents, whereas immigrants from abroad are split proportionally according to the population of the regions.

With the additional region-specific external migration approach, every region has specifically adapted immigration and emigration rates for abroad. To implement this, foreign migration data for every region used has to be added, implying additional parametrisation data. Still, it has the big advantage of improving the simulation results significantly, which will be shown later on.

Table 4.2 offers an overview of the amount of data points necessary for the parametrisation of the modelling approaches. In the case presented, nine different regions containing agents of 112 age classes are assumed. The IRM model needs the least amount of data and the BRrem and MPrem the highest. The enhancement with region-specific external migration roughly equals a doubling of the required data points.

Model	Parameter	Sex	Age	Regions	Timespan	Datapoints
GEPOC	Emigration	2	112	1	16	3.584
	Immigration	2	112	1	16	3.584
	Birth	1	112	1	16	1.792
	Death	2	112	1	16	3.584
	TOTAL					12.544
BR	GEPOC					12.544
	Internal Emigration	2	112	9	16	32.256
	Internal Immigration	2	112	9	16	32.256
	TOTAL					77.056
MP	GEPOC					12.544
	Internal Emigration	2	112	9	16	32.256
	Internal Immigration	2	112	9	16	32.256
	TOTAL					77.056
IRM	GEPOC					12.544
	Internal Emigration	2	112	9	16	32.256
	Internal Immigration	2	1	81	16	2.592
	TOTAL					47.392
BRrem	Emigration	2	112	9	16	32.256
	Immigration	2	112	9	16	32.256
	Birth	1	112	1	16	1.792
	Death	2	112	1	16	3.584
	Internal Emigration	2	112	9	16	32.256
	Internal Immigration	2	112	9	16	32.256
	TOTAL					134.400
MPrem	Emigration	2	112	9	16	32.256
	Immigration	2	112	9	16	32.256
	Birth	1	112	1	16	1.792
	Death	2	112	1	16	3.584
	Internal Emigration	2	112	9	16	32.256
	Internal Immigration	2	112	9	16	32.256
	TOTAL					134.400
IRMrem	Emigration	2	112	9	16	32.256
	Immigration	2	112	9	16	32.256
	Birth	1	112	1	16	1.792
	Death	2	112	1	16	3.584
	Internal Emigration	2	112	9	16	32.256
	Internal Immigration	2	1	81	16	2.592
	TOTAL					104.736

Table 4.2: Parametrisation data points needed for the various models. Nine different regions with 112 age classes are assumed.

5 Implementation and parametrisation

Before the implementation of the different approaches presented in the last section could be realised, an adaption of the original GEPOC model had to be made to allow the modular adding of arbitrary regions. Thus the first section of this chapter outlines the implementation of the original GEPOC model. The subsequent sections specify the changes undertaken to allow internal migration and the construction of the different modelling approaches.

5.1 Implementation of the Generic Population Concept

In this section a short technical description of the original GEPOC model will be given. For a more detailed documentation the unpublished "Code-Documentation of the Agent-Based GEPOC Model implemented in Python 3" and the Generic Population Handbook are available on demand. The simulation of this process was realised with the following classes and modules:

- ABSimulation (module)
  - __init__ (constructor)
  - Simulation (class)
  - AgentFeatures (class)
  - Protocol (class)
  - dict_csv.py (collection of functions)
- Parameters (module)
  - __init__ (constructor)
  - Births, Deaths, Immigration, Emigration, InitialPopulation, Refugee (classes)
  - statics.py (script)
  - subfunctions.py (collection of functions)
- Presampling (module)
  - __init__ (constructor)

  - Presampler (class)
  - PopulationPresampler (class)
  - ImmigrationPresampler (class)
- Run.py (script)
- ExecutePart.py (script)
- ExecutionParameters.py (script)

The whole simulation can be started by executing the script *Run.py*. It is mainly responsible for the management of the parallelisation of the simulation, which guarantees a faster processing of a simulation run than without parallelisation. All parameters meant to be changed by the user before start-up (like country, timespan, population scaling factor and the number of simulation runs) can be set in *ExecutionParameters.py*. Each run is split into a number of simulation parts which are simulated independently and serially processed by a number of workers to allow multithreading. Each worker executes the script *ExecutePart.py* to perform one part of the simulation. To identify results and temporary files appearing during the simulation, *Run.py* generates a number of identifying time-stamps.

Since the creation of the initial population and the foreign immigrants requires some time, this process is decoupled from the rest of the simulation. If in *ExecutionParameters.py* re-sampling is allowed, *Run.py* tries to find existing population/immigration files in order to reuse them. Otherwise or if no files are found, new population/immigration samples are constructed, using the *Presampling* module. The pre-sampled lists are saved as csv files in a folder called *Presampled*. The separate simulation instances use these files to generate their agent population and immigration agents.

ExecutePart.py finally loads and uses the *ABSimulation* and *Parameters* modules to run the simulation.

The ABSimulation module

The *ABSimulation* module is mainly responsible for the execution of the agent-based model. It consists of three classes and a collection of additional functions for the loading and saving of Python-Dicts in csv format.

An instance of the *Simulation* class represents an executable object that executes an agent-based model of a population. This class requires the current simulation part as well as the start-up settings of *ExecutionParameters.py* and the paths to the pre-sampled population/immigration files. In addition, it requires and loads the *Parameter* module and its sampler classes.

Within this class the procedure `Simulation.Execute()` poses the most important routine. Here the simulation is initialised and the model time-steps are iteratively worked off. As visualised in Figure 5.1, `Simulation.Execute()` calls the initialise routine `Simulation.Initialise()` where all necessary variables are reset. Afterwards

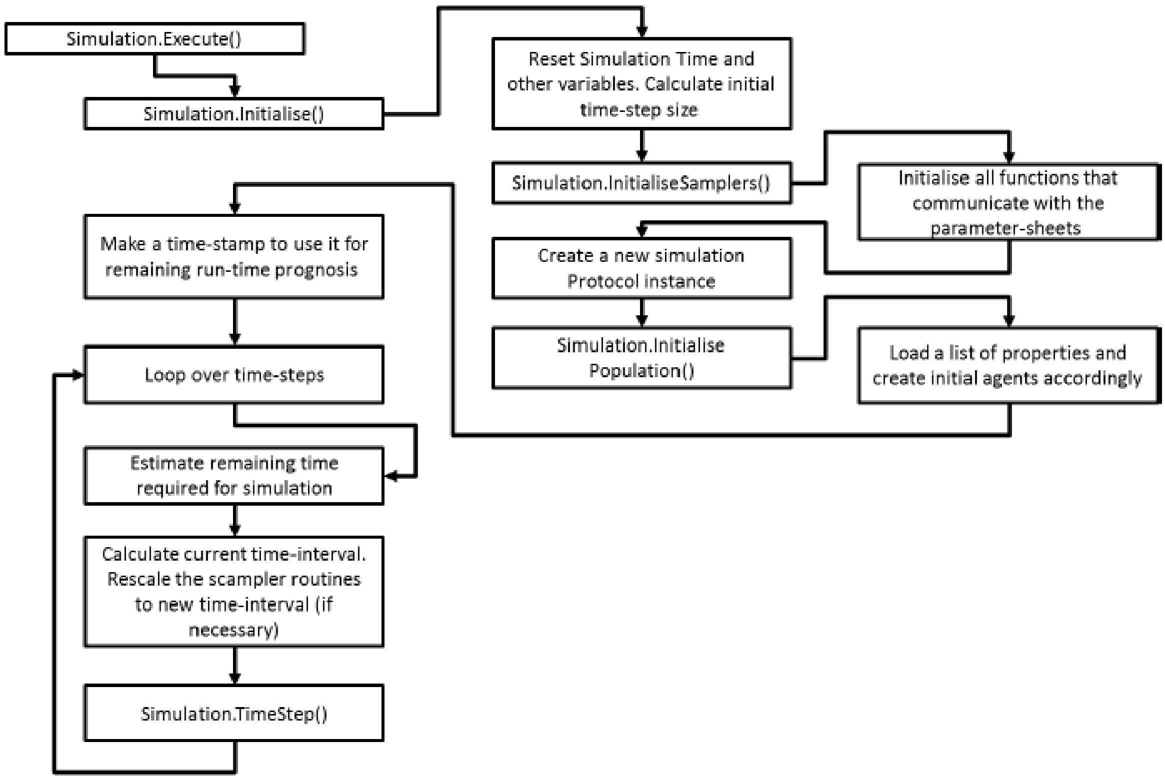

Figure 5.1: Flow chart of the `Simulation.Execute()` routine.

`Simulation.InitialiseSamplers()` creates one instance of all sampler classes offered by the *Parameter* module (*Births, Deaths* etc.). Now a loop over the given time-steps follows where, first, the simulation checks if it is necessary to adapt/rescale the sampler routines to a different time-step length and second, the `Simulation.TimeStep()` routine manages all events taking place in the current time-step. `Simulation.TimeStep()` is the heart of the agent-based simulation and has already been described in Section 3.3 under time-dynamics. At the end of each time-step, the `Simulation.Immigration()` routine (which was described as government-agent in Section 3.3) is called and creates newly immigrated agents.

Table 5.1 shows the features of an agent representing an individual. Since each one ages during the simulation, only its birth date is stored. To obtain the current age of the agent the `AgentFeatures.getAge()` method is used. Depending on the chosen format, it returns the age in days or years. Each agent possesses a unique index as an identifier, which is stored in the agent list of the simulation. Naturally the sex is part of each agent's information. It is either the variable `Parameters.MALE` or `Parameters.FEMALE`, which are defined in an extra file. The mother of an agent is stored as object pointer. Likewise, if an agent is female and has given birth, its children are stored into a list object. In case of immigration, the boolean `refugee` stores if the immigrant is a refugee or not. The protocol instance of the simulation is saved as a "pointer" to allow the agent

Agent
-birth_date : datetime -index : integer -sex : MALE/FEMALE -mother : Agent -childlist : list(Agent) -refugee : boolean -P : Protocol
+getAge(time : datetime, format : string) +Protocol(age : integer, message : string) +get/set-methods

Table 5.1: Structure of a person-agent according to the basic model.

to write on it. Therefore the routine *Protocol* is used whenever the agent has to protocol something in order to generate some simulation output. Some simple additional get- and set-methods are part of the agent class.

The *Protocol* class poses for the generator of the simulation output. While the *Simulation* class itself only takes care of the correct execution of the agent-based model, this class defines which parts of the model are documented for the simulation output. What should be documented depends on the research question and might vary a lot. On the one hand documentation consumes resources and too much output becomes unreadable. On the other hand if too little documentation is generated, any slight specification of the research question leads to a problem that requires modification of the *Protocol* class and new simulation executions. Therefore it is generally recommended to protocol as much as demanded by the research question plus some additional details.

The *dict_csv.py* file contains a collection of additional functions which are needed for the simulation. Among these are routines to save and load a standard `dict` object to and from a csv file, to aggregate all files that have been saved during one simulation run and one to deal with different results of different simulation runs and make some simple statistics with them. The main purpose of these statistics is to understand the level of fluctuation of single simulation runs and to automatically calculate a mean-value for all runs of a scenario.

The Parameters module

The *Parameters* module consists of a couple of classes, each responsible for the communication between simulation and parameter-sheets. Therefore it is necessary for applying parameter-values to simulation parameters and sampling random numbers.

Each of the *Births, Deaths, Emigration, Fertility* and *Refugee* classes communicates with one (or two in the case of female/male distinction) parameter-sheet. Its main routine is a `<...>.randomly_decide_<...>()` function that samples whether or not some event takes place during a time-step of the simulation. Therefore these classes behave like a sampler and are named accordingly in the simulation (e.g. `EmigrationSampler()`) Each of these classes contains the following three routines:

- `<...>.__init__(exeparamsfilename, timeDelta)`: During the class-initialisation the sampler loads the parameter-sheets. The procedure moreover creates a `dict` object of probabilities which is used by the aforementioned `<...>.randomly_decide_<...>()` method. As a matter of data, by default, parameter-sheets contain data in the form of rate per year, probability per year, absolute number per year, etc. Hence, in order to calculate probabilities per time-step-length, these numbers need to be pre-processed. While the calculation from rate/absolute number to probabilities works quite straightforward, the adaption to a different time-span is not as trivial. In this model, formula (5.1) gained from geometric distribution is used to scale the probabilities to an arbitrary time-span.

$$P(\Delta t = x[days]) = 1 - (1 - P(\Delta t = 365[days]))^{\frac{x}{365}} \tag{5.1}$$

- `<...>.rescale(timeDelta)`: As the simulation might possibly use dynamic time-step lengths, the parameters need to be rescaled to new `timeDelta` values. This is done by simply recalling the `<...>.__init__(...)` routine with a new time-step length.
- `<...>.randomly_decide_<...>()`: During this routine the saved probabilities are searched for a specific date (and age). A uniform random number is generated. If this number is smaller than the accessed probability, the routine returns `True`, otherwise `False`.

The *InitialPopulation* and *Immigration* classes mainly contain two routines: `<...>.__init__(exeparamsfilename, filename.csv,...)` loads a pre-sampled population/immigration sample from the *filename.csv* file, which was generated during the pre-sampling phase of the *Run.py* script via the *Presampler* module and contains a list of tuples. Each of these consists either of birth date and sex or birth date, sex and immigration date. The `<...>.sample()` class returns a specific subset of the loaded population/immigration sample. This sample is adjusted to the number of parts in which the simulation run is split and to the current-time interval (only for *Immigration*).

Static.py is a script to specify statics for the simulation. Basically it contains parameters that should not be changed. *Subfunctions.py* consists of two specific functions: `<...>.load_paramterfile(filepath)`, which loads a csv file at the given file path and generates a `dict` object of `dict` objects. The routine `<...>.get_indexyear_for_parameter(timeNow, timeDelta)` calculates which data values should be accessed for a given time-step.

The Presampling Module

The *Presampler* class basically offers one main routine: the `Presampler.ProvideSamples` routine. Hereby, the *Presampler* checks whether there already are pre-sampled populations and immigrants available that can are suitable for the simulation, according to the simulation settings (start-/enddate).

The *PopulationPresampler* class loads the parameter-file `Parameters.POPULATION(FE)MALEFILE` which matches *Results/COUNTRY/population_m(f).csv*. The start-year of the simulation is then used to access the correct initial sex/age distribution of the country's population. According to this 2D distribution the python module `PYMC` is used to set up a probability model. The `PopulationPresampler.sample(number)` routine finally samples an arbitrary number of individuals according to the set-up distribution. The output consists of a list of sampled tupels.

The *ImmigrationPresampler* class works analogously to the *PopulationPresampler*. Yet, immigrants for all years considered in the simulation are sampled here, which makes the use of the sampler a little more difficult.

5.2 General changes to enable internal migration

The following itemisation shows the structure of the original GEPOC model, with changed parts marked in blue, plus classes which have been added completely new in red. Common changes for all modelling approaches are described in the following.

- ABSimulation (module)
  - __init__ (constructor)
  - Simulation (class)
  - AgentFeatures (class)
  - Protocol (class)
  - dict_csv.py (collection of functions)
- Parameters (module)

- __init__ (constructor)
- Births, Deaths, Immigration[1], Emigration[2], InitialPopulation, Refugee (classes)
- statics.py (script)
- subfunctions.py (collection of functions)
- InternalMigration (class)

- Presampling (module)
  - __init__ (constructor)
  - Presampler (class)
  - PopulationPresampler (class)
  - ImmigrationPresampler (class)[2]
  - InternalMigrationrates (class)
  - Emigrationrates (class)[2]
- Run.py (script)
- ExecutePart.py (script)
- ExecutionParameters.py (script)

ExecutionParameters.py has an additional parameter `REGIONS`. Here the regions which should be part of the simulation and for which data sheets are provided, are inserted in string format. For each modelling approach the paths to the corresponding files are registered. The *Run.py* script calls the corresponding *InternalMigrationrates* class for each model in order to calculate internal migration rates if the data sheets provided by the user consist of absolute migration numbers. *ExecutePart.py* now receives a list of population files instead of one. To deal with that, a string separation had to be implemented.

The ABSimulation module

In *Simulation.py* the initialisation of the population now happens per region. `Simulation.TimeStep()` imports the current region for each agent. As a newly added action, it is decided if and when an individual decides to migrate internally and the result is added to the event-list. This decision process happens in a different way for the model approaches and will be described in the following sections. The new routine `Simulation.getShareOfTotalPopulation(...)` calculates the ratio of the regional and total population for every region. This is used in the `Simulation.Immigration(...)` routine to split external immigrants in proportion to the number of inhabitants between

[1]Only if no region-specific external migration is used.
[2]Only for region-specific external migration.

the regions. In `Simulation.GiveBirth(...)` every agent who has a mother inhabits the current region of its mother agent.

The class *AgentFeatures* now stores an agent's region as well as a list of its movements during the simulation. In addition its routine *Protocol* was altered for internal migration.

The Parameters Module

The constructor of the *Parameters* module additionally imports the *InternalMigration* class. For all models without region-specific external migration the *Immigration* class now has the `Immigration.getRegionToImmigrateTo(...)` routine. Here an initial region for each external immigrant is chosen according to the relative population of each region.

The new *InternalMigration* class handles the internal migration process. It contains methods to decide if an agent emigrates internally and where he moves to. It is constructed in a similar way as the other classes of the *Parameters* module. Since it differs for every modelling approach, it will be described in Sections 5.3–5.5. Changes in the *Emigration* class are only necessary for region-specific external migration and will be described in Section 5.6.

The Presampling Module

The constructor of the *Presampling* module additionally imports the *InternalMigrationrates* class. The `Presampler.ProvideSamples()` routine of the *Presampler* class now assembles the according parameter-files for every region. The pre-sampling process happening in the *PopulationPresampler* class is basically the same, but separately executed for every region.

The new *InternalMigrationrates* class is called during the pre-sampling process to generate data sheets which will be necessary during the simulation. This class differs for every modelling approach and therefore will be described in Sections 5.3–5.5. The *Emigrationrates* class is only needed for region-specific external migration and will be described in Section 5.6.

The data sheets necessary to run the model are listed in Table 5.2. As an alternative to the files concerning internal and external absolute migration numbers of the individual regions, it is possible to provide data sheets containing the corresponding migration probabilities. All files have to be in csv format with semicolon separators. Data sheets containing values separated into age classes do not necessarily have to be provided up to the maximum age an agent might reach. If e.g. the internal emigrations are documented up to 100 years but an agent turns 101, the emigration probability for centennial people will be reused.

Necessary general data		
Model	**Data sheet**	**Description**
all	fertility.csv	avg. number of children per 1000 women aged (row) in year (col)
all	death_probabilities_m(f).csv	death probability to die in year (col) with age (row)
no rem	immigrants_m(f).csv	immigrants per age (row) during specific year (col)
no rem	emigration_probabilities_m(f).csv	probability per age (row) to emigrate during specific year (col)
Necessary data for every region		
Model	**Data sheet**	**Description**
all	population_m(f).csv	population per age (row) at 01.01. of initial year (col)
all	internalEmigration_m(f).csv	internal emigrants per age (row) during specific year (col)
BR, MP	internalImmigration_m(f).csv	internal immigrants per age (row) during specific year (col)
rem	externalEmigration_m(f).csv	external emigrants per age (row) during specific year (col)
rem	externalImmigration_m(f).csv	external immigrants per age (row) during specific year (col)
Necessary data for every year		
Model	**Data sheet**	**Description**
IRM	migrationsBetweenRegions_m(f)_*year*.csv	migrations from one region (row) to another (col) for given year

Table 5.2: Necessary data sheets to execute a simulation. All files except for “fertility.csv” exist for male (m) and female (f) gender.

5.3 Biregional model

Since the BR model handles internal migration in two steps, in *Simulation.py* first it is determined for every individual if they emigrate internally. Afterwards a goal region is chosen for all emigrating agents. For this process the *InternalMigration* class had to be constructed.

Components of the InternalMigration class

- `<...>.__init__(exeparamsfilename, timeDelta)`: During the initialisation of the class, the parameter-sheets for internal migration probabilities and total population are loaded. The procedure moreover creates `dict` objects of probabilities, based on the imported files, which are used by the routines described in the following. Like with the other classes of the *Parameters* module, the data has to be pre-processed to fit the respective time-step-length.

- `<...>.randomly_decide_to_emigrate_internal(...)`: This routine chooses for every agent if it emigrates internally. A uniform random number is generated and the probabilities saved for a specific date, age and region are compared to it. If the number is smaller than the accessed probability, the routine returns `True` otherwise `False`.

- `<...>.decide_place_to_go(...)`: Here for every agent who emigrated internally at the current time-step t, a goal region is determined. For each region j, age a and sex s the share of immigrants per region (ipr) is calculated. P is the total population and em the amount of internal emigrants of a certain sex and age at time t. To provide the population numbers, in the *Simulation* class the `Simulation.getShareOfTotalPopulation(...)` routine additionally calculates the actual population per gender and age. Naturally, the population scaling factor sc needs to be included as well as the immigration probabilities ip, which are calculated during the class-initialisation.

$$\mathrm{ipr}_{j,s,a} = \frac{\mathrm{P}_{s,a}(t) \cdot \mathrm{sc} \cdot \mathrm{ip}_{j,s,a}(t)}{\mathrm{em}_{s,a}(t)} \tag{5.2}$$

  Equation (5.2) calculates the amount of agents in the simulation who actually immigrate to region j for every sex and age, divided by the number of total internal emigrants. To eliminate the risk of an emigrant getting no goal region because of calculating errors, the share of immigrants per region is rescaled to make sure the immigration rates add up to 1. Now every internal emigrant gets a new region assigned via the `<...>.getRegionToImmigrateTo(...)` routine.

- `<...>.getRegionToImmigrateTo(...)`: Based on an agent's age and sex, this routine compares a uniform random number with the share of immigrants per region calculated in <...>.decide_place_to_go(...). It returns the goal region for the particular agent.

- <...>.rescale(timeDelta): This routine rescales the parameters if necessary analogous to the other classes of the *Parameters* module.

- <...>.getMaxage(): This get-method returns the maximum age of all agents contained in the data.

Components of the InternalMigrationrates class

- <...>.__init__(): During the class-initialisation the sampler loads the files containing internal emigration and immigration numbers in absolute figures.

- <...>.checkAvailability(file): This routine checks if a certain file already exists and therefore does not need to be created again. It is used by the other routines of this class.

- <...>.ProvideAllEmigrationProbabilities(): Here it is evaluated for every region if for each sex a file containing the internal emigration probabilities for all ages already exists. Otherwise the necessary values are calculated using the <...>.calculateAllEmigrationProbabilities(...) routine and stored in the corresponding files.

- <...>.calculateAllEmigrationProbabilities(...): This routine calculates the regional emigration probabilities for male and female files for all age groups according to equation (4.1) of the BR model.

- <...>.ProvideAllImmigrationProbabilities(): Just as with emigration it is checked, for each sex and all age groups, if a file containing the probabilities for internal migration already exists. Otherwise the values are determined via <...>.calculateAllImmigrationProbabilities(...) and stored in the corresponding files.

- <...>.calculateAllImmigrationProbabilities(...): Here the regional immigration probabilities for male and female files for all age groups are calculated according to equation (4.3) of the BR model.

- <...>.ProvideTotalPopulation(): This routine checks if a file containing the total population of the simulation already exists. Otherwise such a file is created for each gender by adding the inhabitants of every region's age groups.

5.4 Migration pool model

Compared to the BR model the big advantage of the MP model is the handling of internal migration in one step. Thus, if during the "action phase" of a time-step in *Simulation.py* an agent emigrates internally, it is immediately decided via the *InternalMigration* class where it will move to.

Components of the InternalMigration class

- `<...>.__init__(exeparamsfilename, timeDelta)`: During the initialisation the sampler loads the parameter-sheets for internal emigration probabilities and distribution proportions. In addition, `dict` objects containing these probabilities, are created. They are used by the routines described in the following. Analogous to the other classes of the *Parameters* module, the data has to be pre-processed to fit the respective time-step-length.

- `<...>.getRegionToImmigrateTo_MP(...)`: Based on a uniform random number, this routine first decides if the particular agent emigrates internally. If that is the case, it decides (according to the distribution proportions) where the agent moves to, again using a uniform random number.

- The following routines are identical to the BR model:
  - `<...>.rescale(timeDelta)`
  - `<...>.getMaxage()`

Components of the InternalMigrationrates class

- `<...>.ProvideDistributionProportions()`: It is checked for each sex if a file containing the distribution proportions for all age groups already exists. Otherwise the necessary values are determined using the `<...>.calculateDistributionProportions(...)` routine and stored in the corresponding files.

- `<...>.calculateDistributionProportions(...)`: This routine calculates the regional distribution proportions for male and female files for all age groups according to equation (4.6) of the MP model.

- `<...>.ProvideTotalImmigrations()`: Similar to the `<...>.ProvideTotalPopulation()` routine of the BR model, this method checks if a file containing the total internal immigration numbers of the simulation exists. Otherwise such a file is created for each gender by adding the internal immigrants of every region's age groups.

- The following routines are identical to the BR model:
  - `<...>.__init__()`
  - `<...>.checkAvailability(file)`
  - `<...>.ProvideAllEmigrationProbabilities()`
  - `<...>.calculateAllEmigrationProbabilities(...)`

5.5 Inter-regional migration model

The IRM model differs in the internal immigration process, which is decoupled from the agent's age. Therefore, while the internal emigration still is decided as part of the "action phase", the goal region is chosen independently of age at the moment the emigration process actually happens. The difficulty of replacing the age component with the two dimensional migration flows between regions is solved by generating internal immigration probability files for each year.

Components of the InternalMigration class

- `<...>.__init__(exeparamsfilename, timeDelta)`: On initialisation this procedure loads the parameter-sheets for internal emigration probabilities and for migration rates between regions. According to them, `dict` objects containing these probabilities are created. The data has to be pre-processed to fit the respective time-step-length in the same way as the other classes of the *Parameters* module.

- `<...>.get_goal_region(...)`: Using a uniform random number, this routine chooses for every agent who emigrated internally a place to move to according to the interregional migration rates.

- Identical to the BR model are:
  - `<...>.randomly_decide_to_emigrate_internal(...)`
  - `<...>.rescale(timeDelta)`
  - `<...>.getMaxage()`

Components of the InternalMigrationrates class

- `<...>.__init__()`: During the initialisation of the class, the sampler loads files containing internal emigration and migration numbers between regions.

- `<...>.ProvideAllMigrationProbabilities()`: When this routine is called, it checks for each sex if a file containing the migration probabilities between regions already exists. If this is not the case, the necessary values are calculated via the `<...>.calculateAllMigrationProbabilities(...)` routine and stored in the corresponding files.

- `<...>.calculateAllMigrationProbabilities(...)`: Here the probabilities for migration from one region to every other single area (including the initial one) are calculated. This happens according to equation (4.10) of the IRM model.

- The following routines are identical to the BR model:
  - `<...>.checkAvailability(file)`
  - `<...>.ProvideAllEmigrationProbabilities()`

- `<...>.calculateAllEmigrationProbabilities(...)`

5.6 Enhancement with region-specific external migration

The enhancement with region-specific external migration works for every modelling approach presented in exactly the same way. Therefore the necessary code changes are only described once in the following.

To provide region-specific external migration, *Run.py* additionally calls the *Emigrationrates* class to provide the probabilities for external migration out of any region if the data sheets provided consist of absolute external migration numbers. Analogous to the population files, *ExecutePart.py* deals with a list of external immigration files instead of one. Since this amplification only deals with external migration, the *InternalMigration* and *InternalMigrationrates* classes remain identical to the respective modelling approach that is enhanced.

The ABSimulation module

In the *Simulation* class the sampling of external immigration takes place separately for each region. The refugee tool is deactivated. In `Simulation.TimeStep()` the decision of external emigration now considers the region of the agent. Since this enhancement splits external immigrants according to provided data, the `Simulation.getShareOfTotalPopulation(...)` routine used in the basic models is no longer needed. This routine is kept only for the BRrem model in order to calculate the total population per age and gender since for the BR approach this is needed for the internal immigration. In `Simulation.Immigration(...)` the new external immigrants are now separately sampled for every region.

The Parameters Module

The *Emigration* class is constructed completely new and similar to internal emigration parts of the *InternalMigration* classes of the BR and IRM models. During initialisation, files for external emigration probabilities per region are imported, pre-processed according to the time-step and saved to `dict` objects. Based on these probabilities the `Emigration.randomly_decide_to_emigrate(...)` routine decides if the agent moves to a foreign country. Again a routine to rescale the parameters if necessary is included in this class.

The Presampling Module

The constructor of the *Presampling* module additionally imports the *Emigrationrates* class. Despite the region-wise population pre-sampling, the *ImmigrationPresampler* class generates the external immigration files separately too.

The *Emigrationrates* class had to be newly constructed to calculate external emigration rates for each region based on data files containing absolute numbers. It is similar to the parts of the *InternalMigrationrates* classes concerning internal emigration. Thus it will not be described in detail.

The implementation and parametrisation of these approaches led to well fitting, similar results, which will be analysed in detail in the following section.

6 Simulation results

For all models presented, the initial data has been provided by the STATcube tool of the Austrian Bureau of Statistics [1]. STATcube is a statistical data base that allows the processing of large amounts of data combined with user-defined generation and exportation of tables. For benchmarking, the federal states of Austria are used as a partition of Austria. For the generation of all results presented, a time-step size of 365 days for the years 2002–2017 was set. In addition, a scale of 0.01 was chosen, leading to about $0.01 \cdot 8\text{m} \approx 80.000$ regarded agents. This time range corresponds with the internal migration data freely available for Austria via STATcube. Note that the accumulated total population of Austria remains unchanged according to the original GEPOC model for all approaches. The calculation happened with an Intel® Core™ i5-6200U CPU 2.30GHz/2.40GHz. The simulation results are multiplied with the factor $\frac{1}{\text{scale}} = \frac{1}{0.01} = 100$ to guarantee comparability with the actual data. Observing the real data for internal migrations in the federal states of Austria shows that although in most parts a tendency towards growing movement occurs, in others it stagnates. In addition, the amount of internal movements diminishes for some years and regions. For the validation of the model approaches in the years 2002–2017 real data for internal migration was used. A prognosis for the further population development regarding internal migrations will be attempted in Section 6.2.

Table 6.1 shows the average duration of one simulation run for the different modelling approaches when no partitioning of the agent population is chosen. The computation time is calculated for simulations using an already pre-sampled agent population. Comparing the runtimes shows that for every approach the enhancement with region-specific external migration seems to have no influence on the runtime. The reason for the different computation times of the three general approaches has to lie in the way internal immigration is handled. The most probable explanation for the longest computation duration of the BR approach is the separate handling of emigration and immigration. It is the only version where the latter takes place for all agents combined after the agent list is run trough, which leads to a second call of all internal emigrants. The most probable reason for the IRM model being the fastest enhancement is the much smaller amount of data needed for the internal immigration process, leading to a faster initialisation of the *InternalMigration* class. Overall, the small difference between the computation time of the original GEPOC model and the enlargements with internal migration confirms that all internal migration models can be regarded as valid.

Model	Avg. runtime	Model	Avg. runtime
BR	26.201	BRrem	26.412
MP	24.195	MPrem	24.188
IRM	22.690	IRMrem	22.779
GEPOC	20.212		

Table 6.1: Comparison of the average runtime in seconds of the different modelling approaches and the original GEPOC model without the pre-sampling process.

6.1 Results of the different modelling approaches

Pages 46–51 present the simulation results generated by the different modelling approaches. Figure 6.1 shows the total populations of each region for the years 2002–2017. First the data according to the census is plotted, followed by the values generated with the models implemented. Afterwards Figure 6.3 presents a comparison of the absolute internal emigrations per region taking place in this timespan, followed lastly by Figure 6.5 representing the census data and results of the absolute internal immigrations happening. These findings will be analysed in Section 6.3.

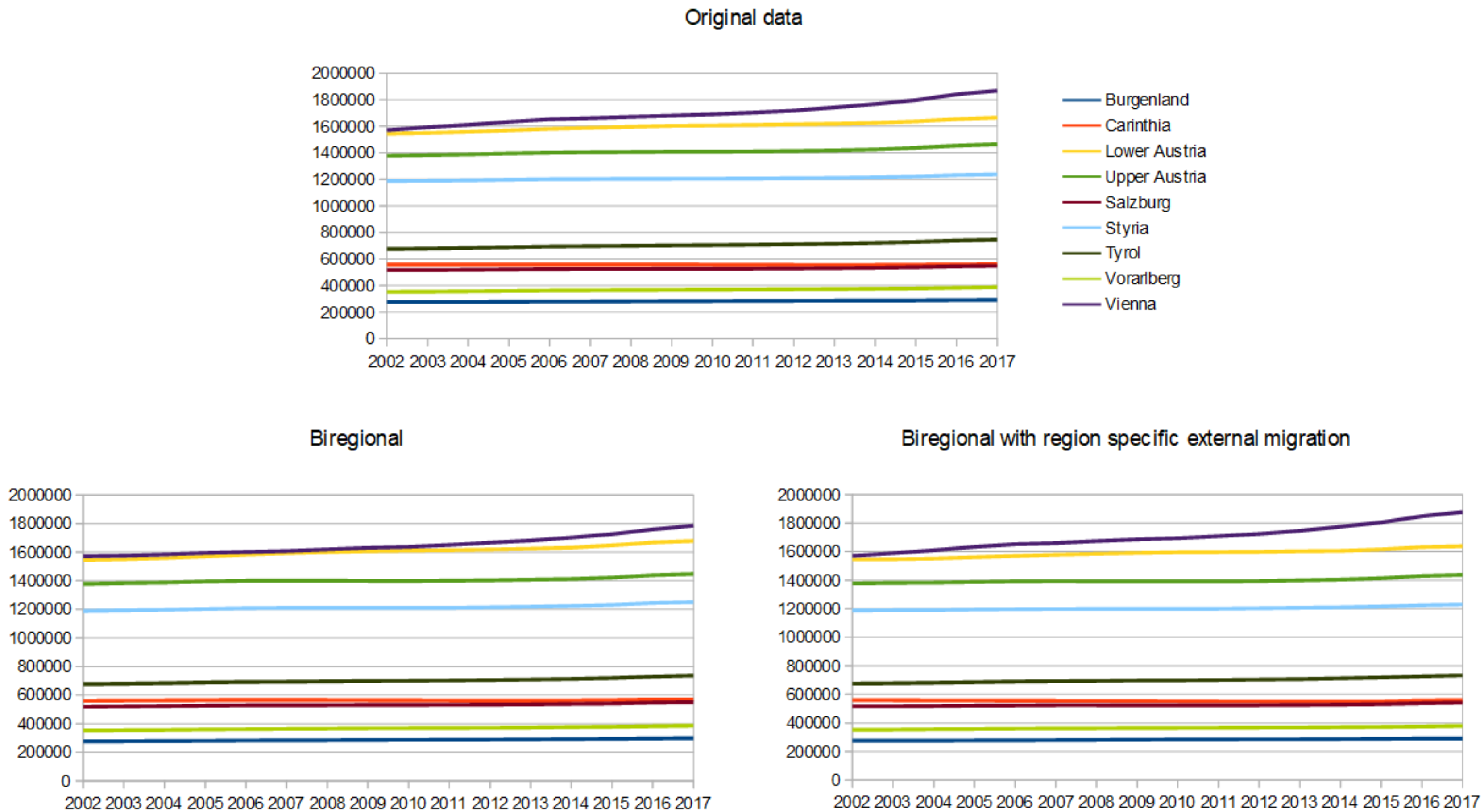

Figure 6.1: Comparison of the total populations per region of all models. Part 1.

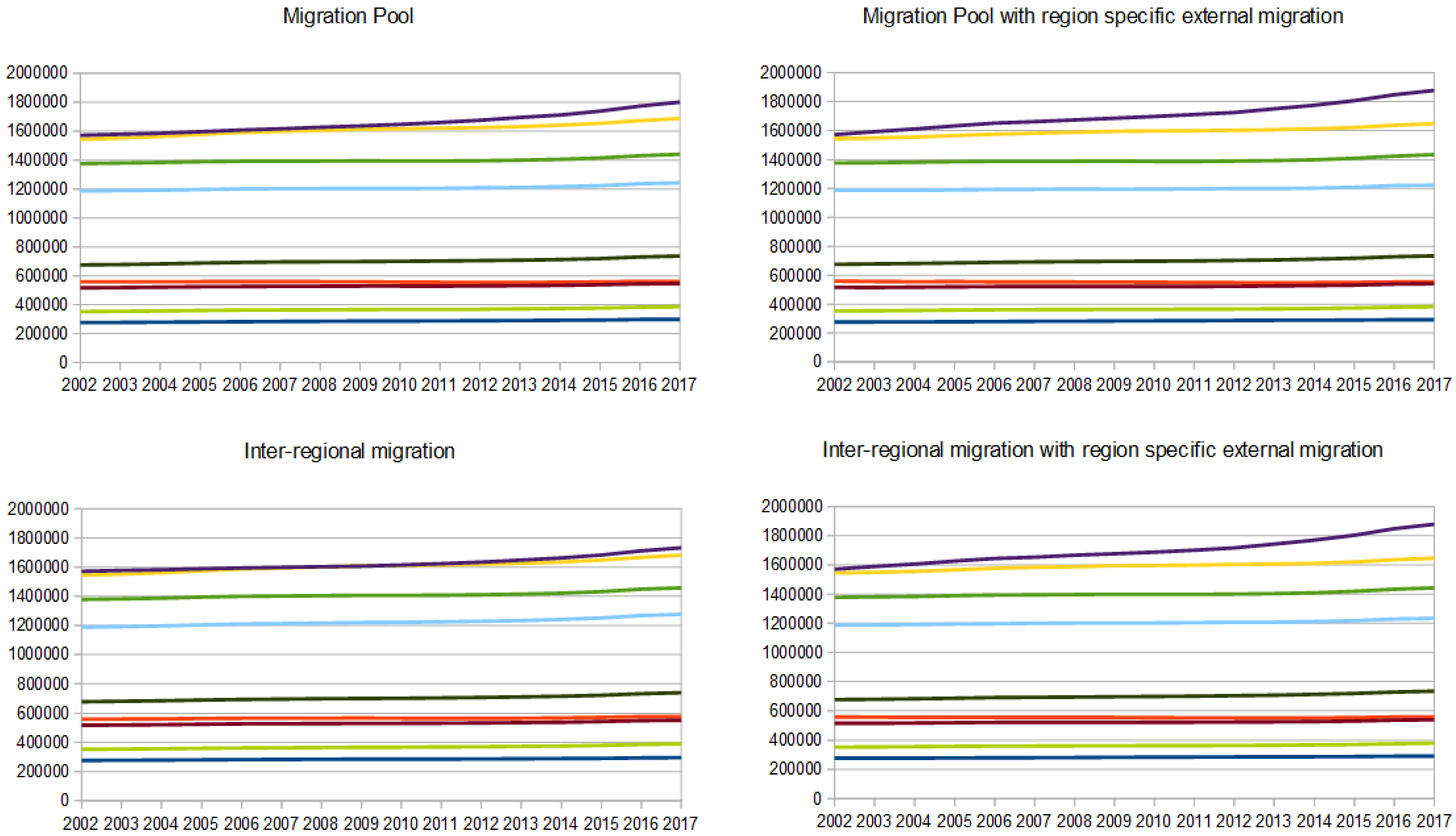

Figure 6.2: Comparison of the total populations per region of all models. Part 2.

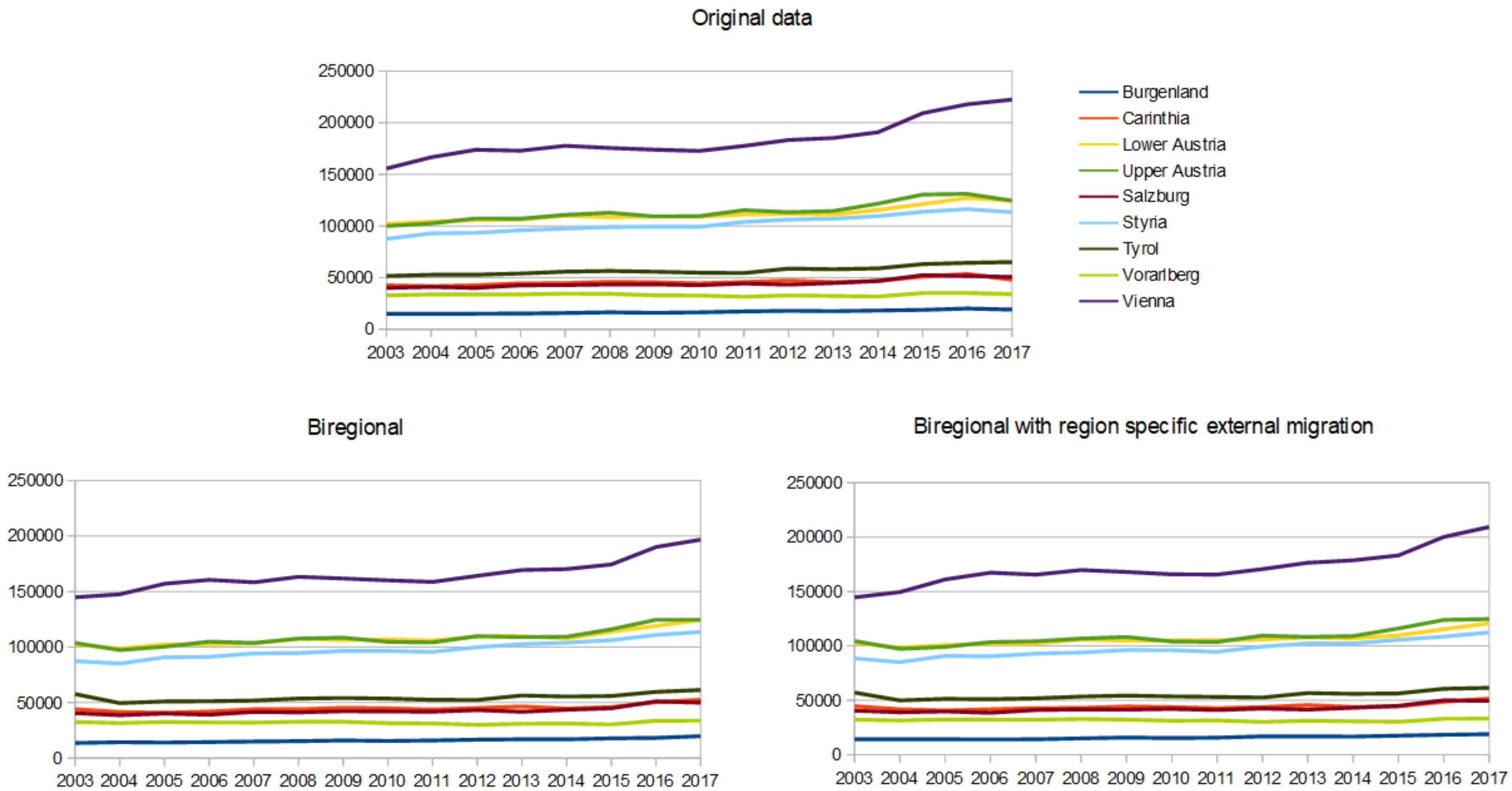

Figure 6.3: Comparison of the internal emigrations per region of all models. Part 1.

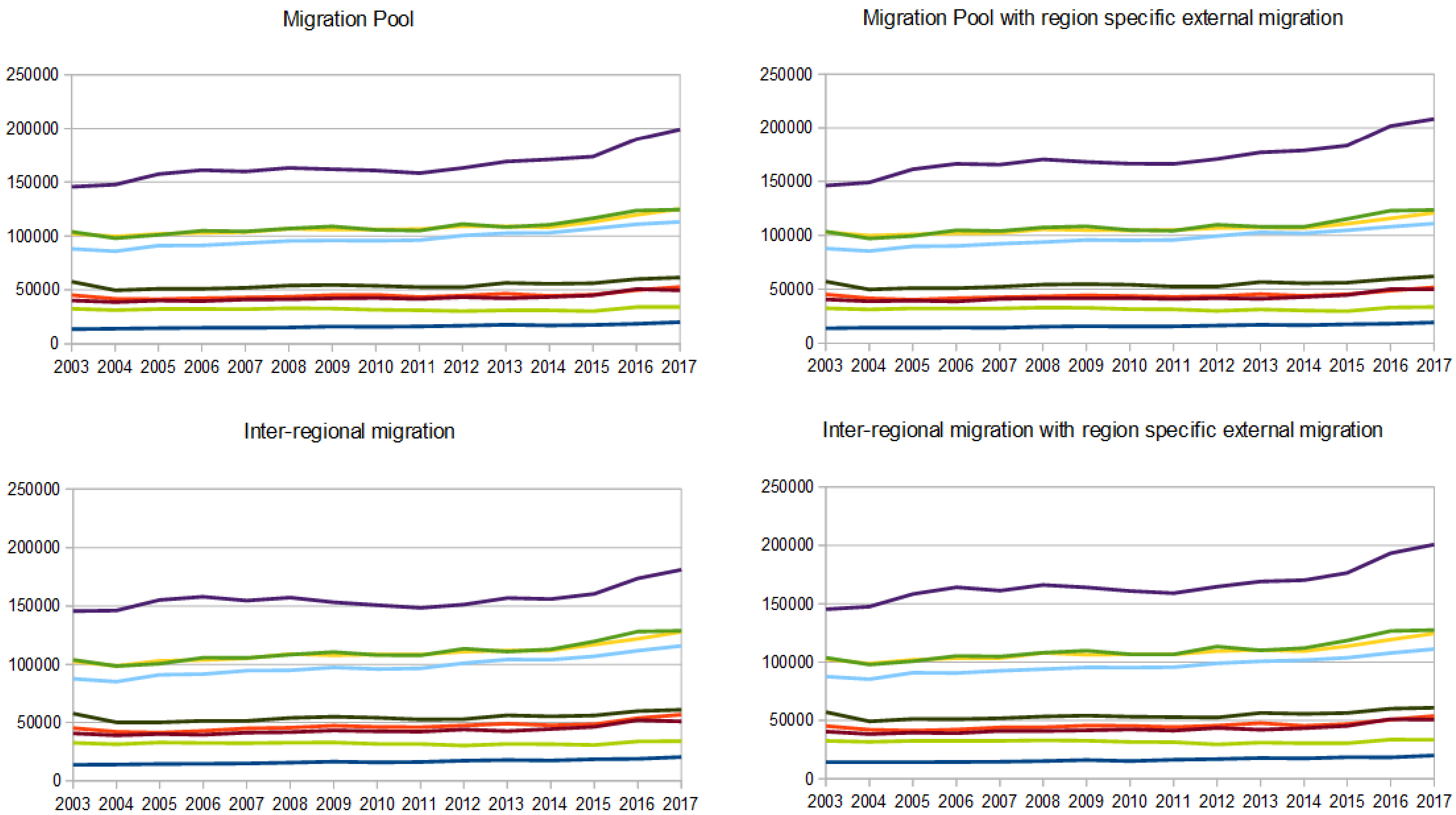

Figure 6.4: Comparison of the internal emigrations per region of all models. Part 2.

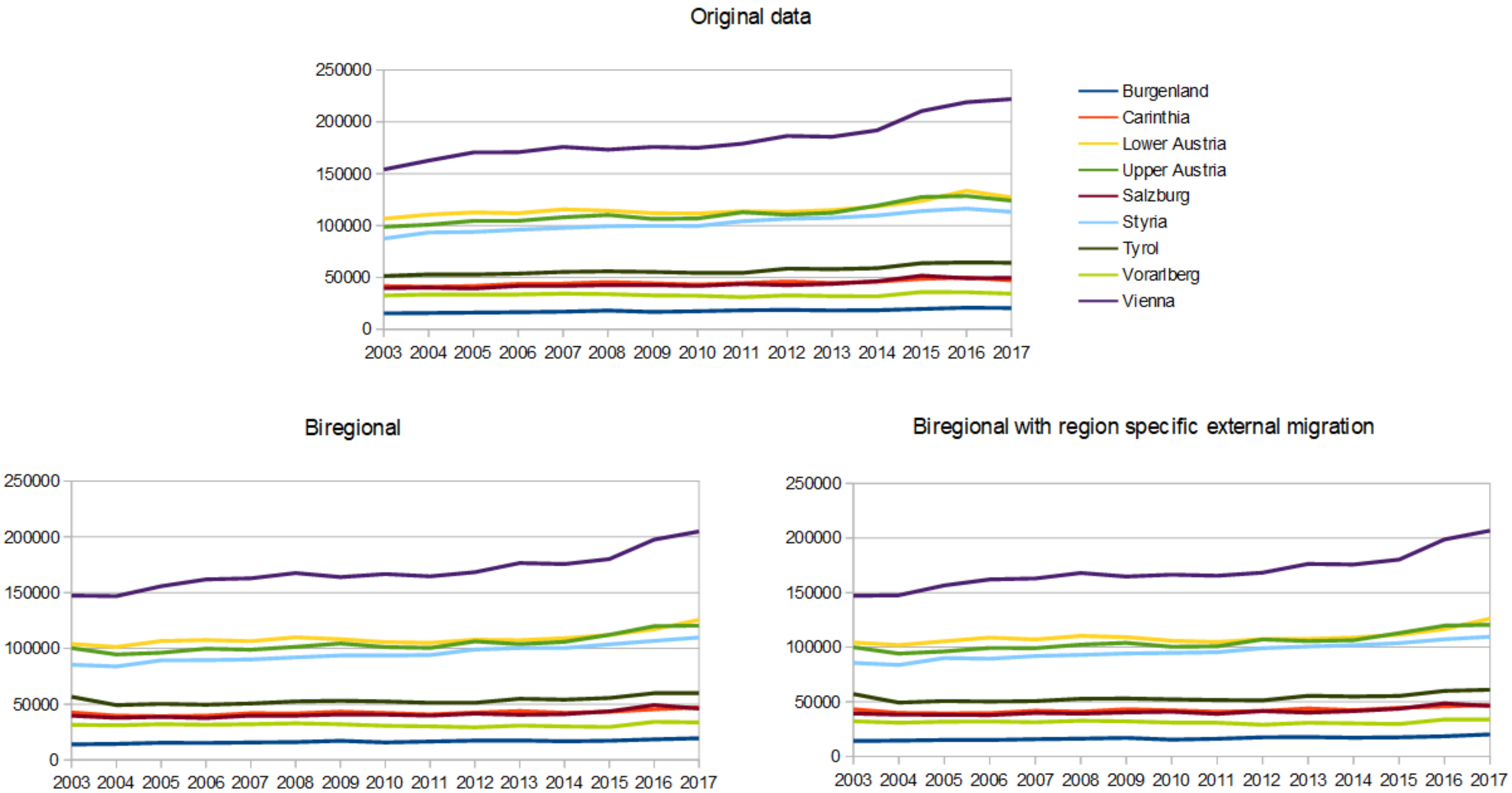

Figure 6.5: Comparison of the internal immigrations per region of all models. Part 1.

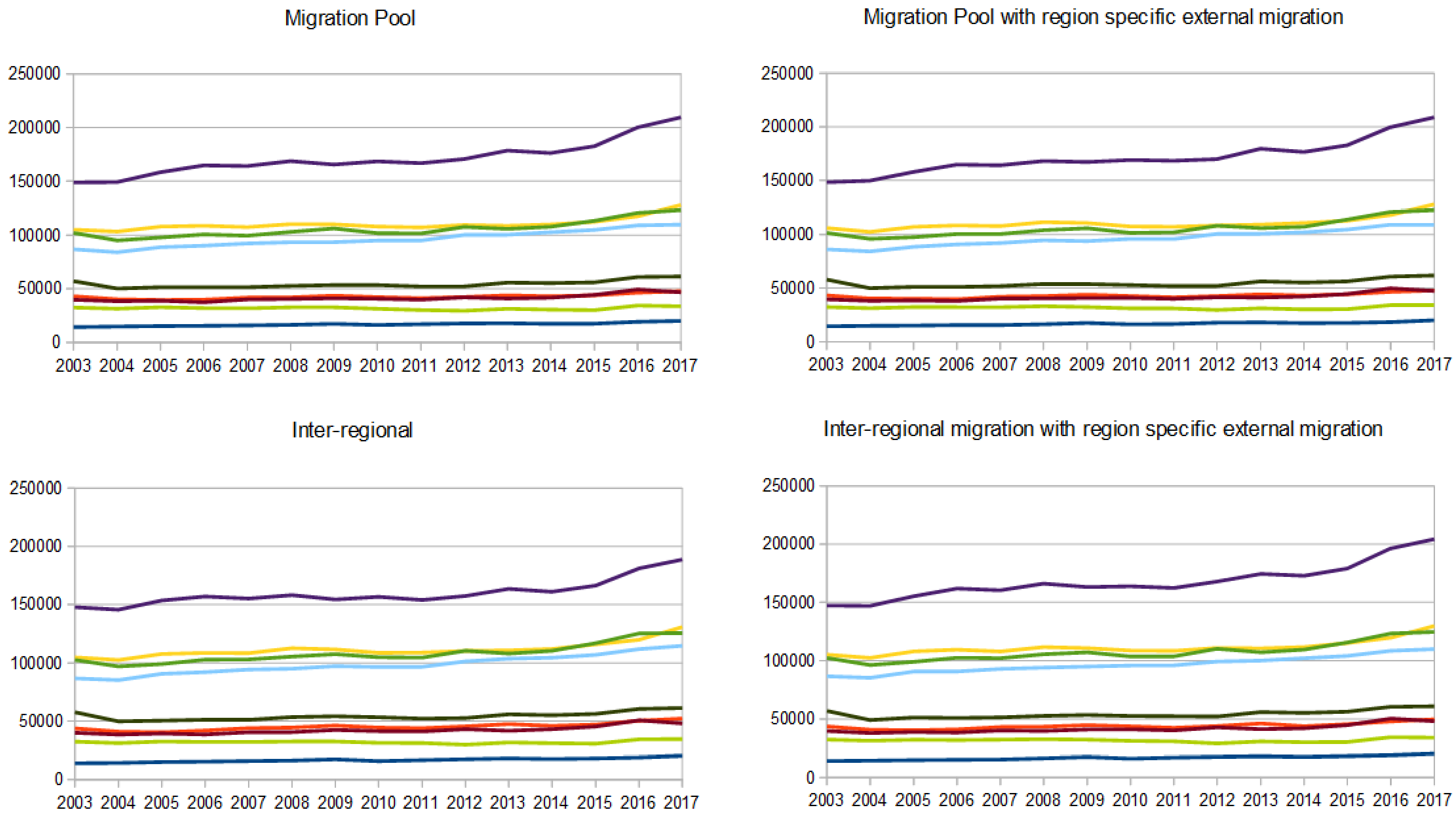

Figure 6.6: Comparison of the internal immigrations per region of all models. Part 2.

6.2 Forecast attempt

To allow simulation over a longer period than the data provided for internal migration ranges, extrapolation for some parametrisation data sheets had to be undertaken. To shorten the process and to avoid unnatural numbers for population as much as possible, the already calculated parametrisation data sheets for internal migration probabilities were extrapolated linearly. Still, for external immigrations the parameter-data in the "externalImmigration_m(f)" files containing absolute numbers of immigrants had to be amplified. For the BR model this way of generating data is sufficient. For the MP model the extrapolation process for internal emigration and external migration functions analogously. Since for internal immigrations the MP approach uses distribution functions, the extrapolation of the parametrisation data in files containing these distribution probabilities is not sufficient as it might destroy the total allocation of the migrants. Therefore the parametrisation data in all "internalImmigration_m(f)" files first has to be extrapolated and second rounded to natural numbers. Rounding is necessary to pre-sample the distribution probabilities accordingly. The extrapolation for the IRM approach is set aside since here the parametrisation data for migration is required per year, which makes the extrapolation process very difficult. Instead, while emigration was handled according to the other enhancements, for the calculation of internal immigration rates the last year available (2017) was reused for generating all forecast data.

The extrapolation process was executed with the function `FORECAST(value;data_Y;data_X)` provided by Apache OpenOffice™Calc version 4.1.5. This routine calculates future values by executing linear regression. As `value` the year of interest is inserted. The `data_X` block consists of all years for which the number of migrants is known, while the `data_Y` block contains these numbers. To avoid possible negative migration rates an `if` condition was included to set such values to zero.

Figure 6.7 on page 53 shows the population results generated for the period 2017–2030. Since no data for the years 2018–2030 is accessible, validation of these forecasts is not possible. Hence this part of the thesis mainly aims at showing the possibility of generating prognosis data and comparing the results of the different approaches.

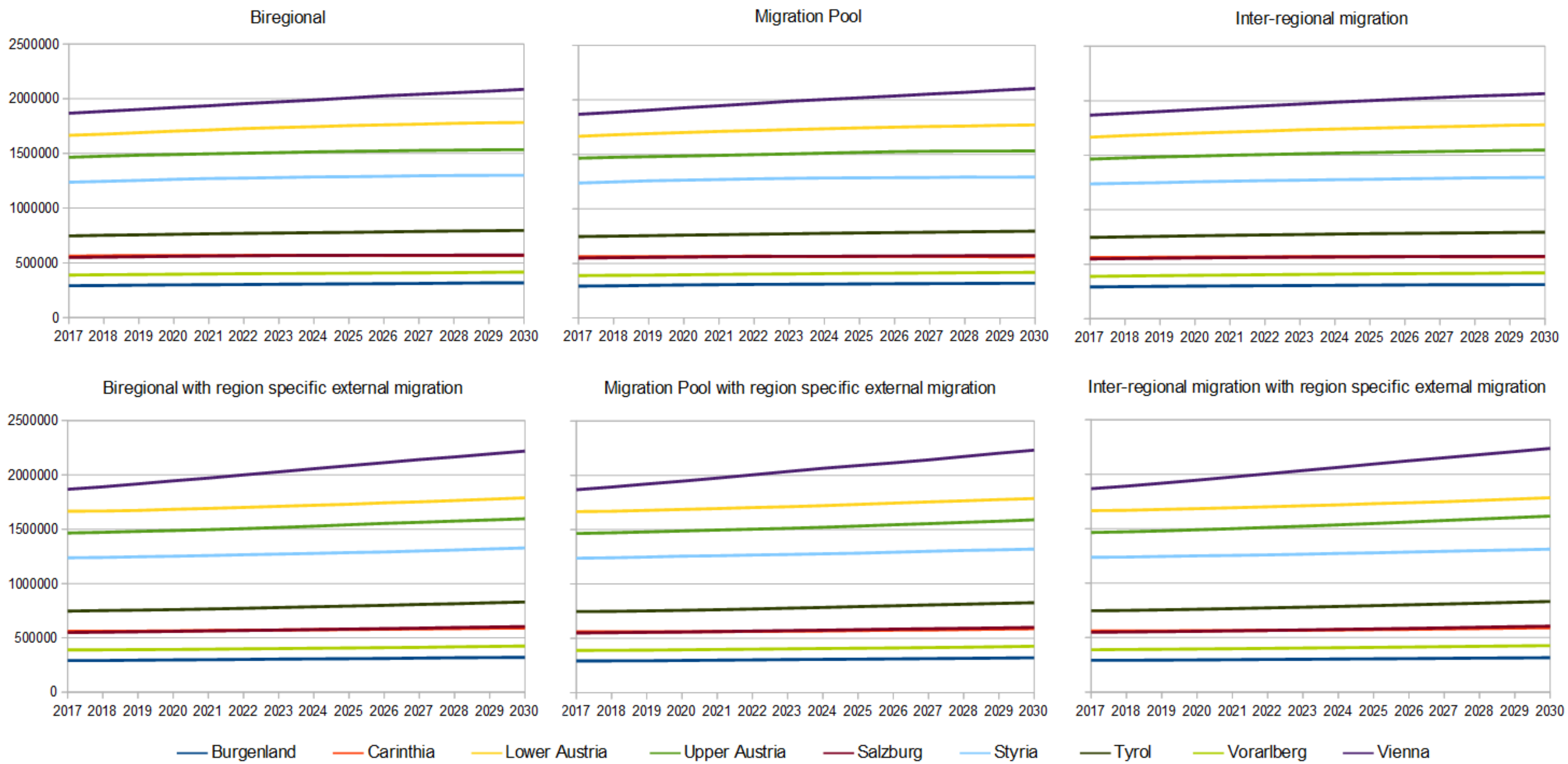

Figure 6.7: Possible population development for the years 2017–2030 according to the models.

6.3 Interpretation of the results

For the interpretation of the simulation results the differences and the relative differences of the generated data are used. They are calculated as

$$e_a = x_D - x_S \quad \text{and} \quad e_r = \frac{x_D - x_S}{|x_D|} \tag{6.1}$$

with x_S being the value sought after according to the simulation and x_D corresponding to the data. Calculating these differences is chosen, instead of the usually positive defined errors, to allow a better visualisation of the simulation results. Thus, the figures in this chapter showing the computed differences, provide a quick overview where too many and where not enough inhabitants are simulated.

The MP model has the highest accuracy of the standard models

The BR and MP models produce very similar results, with the MP approach being slightly more accurate, whereas the IRM model differs from them. The simulation results for the latter show the by far biggest deviation compared to the data. While for the first two approaches the relative difference overall constitutes a maximum deviation of 4.41% for the BR and 3.64% for MP model respectively (both Vienna, 2016), the difference for the IRM model reaches a deviation of up to 7.23% (Vienna, 2017). For all three approaches the growth of Vienna drags behind in favour of the other, less densely populated areas close to Vienna. Since the BR and MP simulations use the same data and only vary in how they handle internal immigration, with the MP approach simplifying the process, their similarity is not very surprising. With the MP model having a quicker computation time, it is satisfying to observe that it produces the best results of the basic implementations.

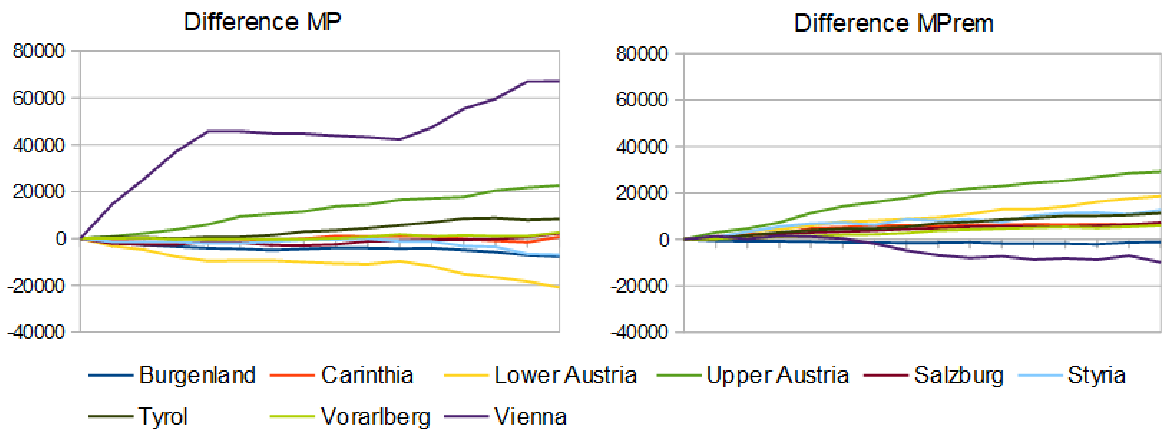

Figure 6.8: Comparison of the differences of the population sizes according to the MP and MPrem models for the years 2002–2017.

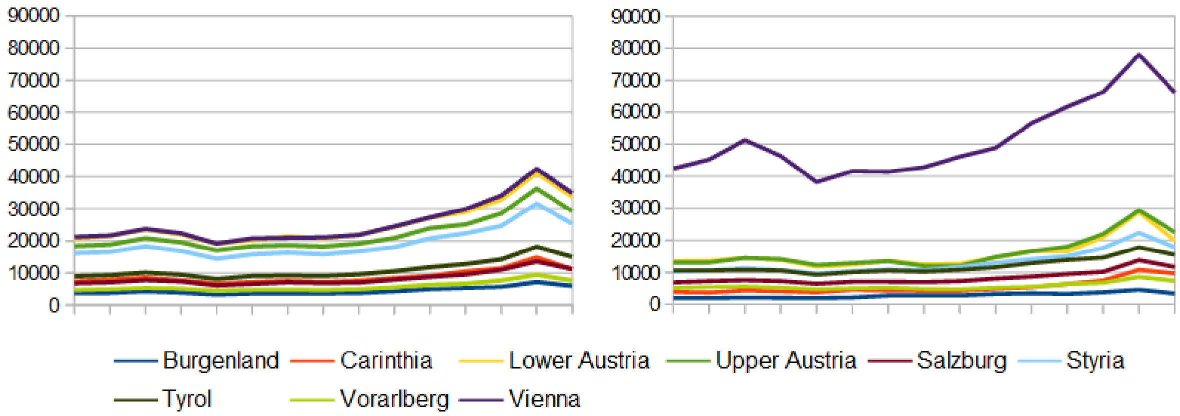

Figure 6.9: Change of the amount of external immigrants trough the enhancement with region-specific external migration rates for the IRM (left) and IRMrem (right) models in the years 2002–2017. These values are basically the same for all simulation approaches.

Enhancement with region-specific external migration pays off

While examining Figure 6.1, which presents the total population numbers, the much better match of all models with the data after the enhancement with region-specific external migration catches the eye. The differences of the MP and the MPrem approach for population numbers in the years 2002–2017 are plotted in Figure 6.8 as examples for the difference between the basic and enhanced modelling approaches. This considerable improvement of simulation results is similar for all three upgrades. While for Vienna the difference between data and outcome differ up to 81.952 people in 2017 for the MP approach, the upgraded MPrem reaches a maximum discrepancy of (minus) 9.802 for Vienna. Still, compared to the actual population of Vienna in 2017 of 1.867.582 persons, the results of both approaches are rather accurate. The relative difference of the MP model upgraded with region-specific external migration now diminishes to a maximum of 1.99% (Upper Austria, 2017). Figure 6.8 also shows the successful transition of the (too many) people living in regions like Lower Austria, according to the standard MP model, to Vienna trough the region-specific external migration enhancement. As illustrated in Figure 6.9, the distribution of external immigrants changes from the allocations solely depended on population density in the standard models, to a distribution according to the real data. The magnitude of this change is the same for all approaches and for external emigration.

The IRMrem model has the highest accuracy of all models implemented

Of all modelling approaches compared, the IRMrem model reaches the most accurate results for the years 2002–2017. As Figure 6.10 shows, the highest deviation reaches an absolute number of 22.437 persons, compared to 29.173 in the second-best performing

MPrem approach (both Upper Austria, 2017). Since the basic IRM model has the worst accuracy, the highly increased performance of the IRMrem model is very interesting. Apparently the internal immigration decision which is directly dependent on the origin of the agent adds up well with the region-specific external migration. This observation will be further investigated in the next paragraph. Considering the very good runtime performance of the IRMrem approach, this model is the best option to choose if the necessary data for the region-specific external migration is available.

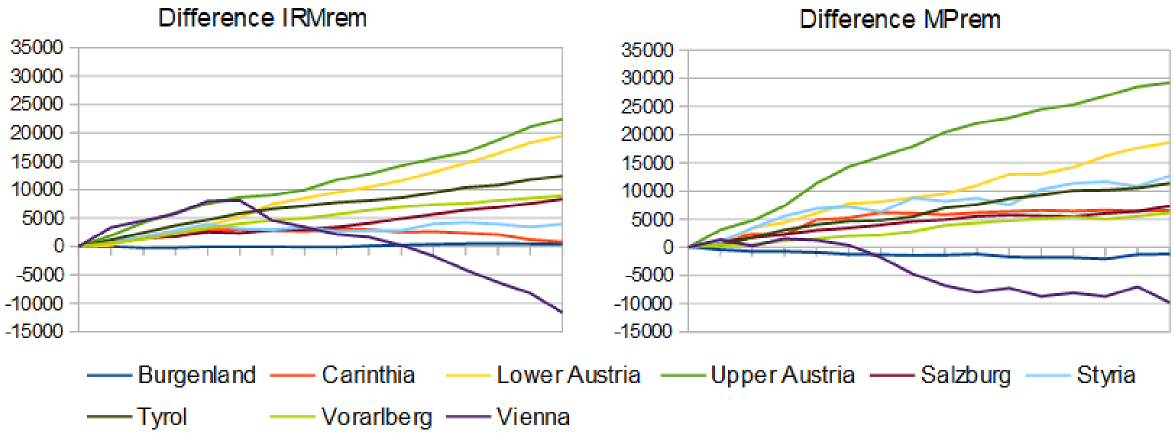

Figure 6.10: Comparison of the difference of the population size according to the IRMrem and MPrem models for the years 2002–2017.

Overall simulated migration rates are too low

A closer look at the internal migration data shown in Figures 6.3 and 6.5 shows that – with very few exceptions – the number of migrants calculated tends to be lower than the actual ones. The rates are too low, especially for Vienna. This finding is in line with the results concerning the population figures. Again the BR and MP models produce similar results with the IRM model dragging behind. The standard models have a maximal deviation between approximately 35.000 (BR and MP) and 50.000 (IRM) regarding the data concerning internal emigrants. The maximal absolute difference of 32.767 for the IRMrem approach compared to a maximum of approximately 25.000 for the BRrem and MPrem models (all Vienna, 2015), is striking. Altogether, the implementation of region-specific external migration has a positive impact on internal emigration rates, too. The similar behaviour patterns of the approaches derive from the use of the same data and the identical data processing for the internal emigration process. The difference between the IRM model and the other two develops over time and is a consequence of the significantly lower internal immigrations happening right from the beginning.

A comparison of the differences for internal migration illustrates the differences of the modelling approaches. While the BR model's difference is about 30.000 missing immigrants, the MP model has a higher accuracy of approximately 27.500. Enhancement with region-specific external migration has only a small compensatory impact on these internal immigration rates. In contrast, an improvement from a difference of about 50.000 to approximately 33.000 (all Vienna, 2015) for the IRM model is to be noticed. This does not keep the IRMrem approach from having a less accurate internal migration fit than the BRrem and MPrem models. Still, as shown above, the IRMrem model has the highest accuracy as regards population figures. The most probable explanation for this is the balance of the internal migration differences for emigration and immigration.

The simulation forecasts produce very similar results

The population forecasts produced by the different modelling approaches are, as visible in Figure 6.7, nearly akin. The differences between the basic models reach a maximum value of less than 40.000 inhabitants (Vienna, 2030 for MP and IRM). For the enhanced approaches the discrepancy is even lower with less then 25.000 people (Upper Austria, 2030 for MPrem and IRMrem). Comparing all approaches, the MP model predicts more people living in the city of Vienna, with accordingly less in the countryside. The IRM model mirrors the contrary development and the BR model is somewhere in between those two. These trends are confirmed when the internal migration rates of all simulation results are compared. Thus, the observations of the timespan 2002–2017 persist for the forecast, although, through the comparably simple extrapolation of the migration data by linear regression, the informative value of the predicted population development is questionable. The low impact of keeping the internal immigration rates fixed on those of 2017 for the IRM and IRMrem approach is an interesting result. A possible explanation is the fact that the determination if an agent emigrates internally (which is decisive for the total number of migrants) happens analogous to the other models. If, compared to the other simulation, too many people immigrated to a certain region, the greater local population would lead to comparably more emigrants in the next time-step. Hence this is a compensatory process. Figure 6.11 tries to illustrate this situation. In addition, the fact that here the internal immigration consists of an allocation formula independent of age, might be helpful.

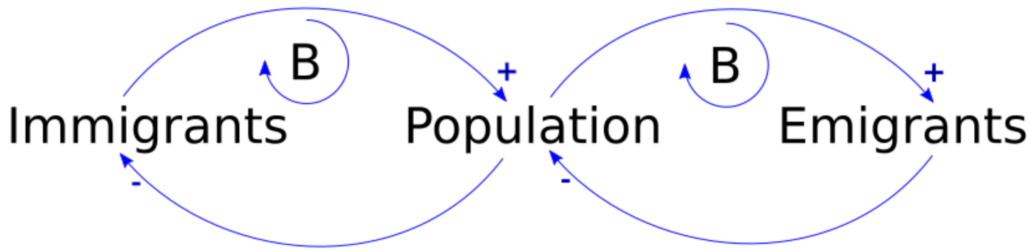

Figure 6.11: Illustration of the connection between immigration, population and emigration. The plus and minus symbols represent positive respectively negative influence and the letter “B” stands for balanced cycles.

For the validation of this forecast data, predictions by the Austrian Bureau of Statistics can be found exemplary for Vienna in [22]. It shows that the original assumption of 2 million people living in Vienna in the year 2022 matches the results of the models enhanced with region-specific external migration. However their prognosis is revised up to 2026 due to decreased external immigration. This suggests that especially the linear extrapolation of the external migration data is not sufficient to allow reliable forecasts. In general the Austrian Bureau of Statistics provides different scenarios of predictions, showing the difficulty of reliable future prospects [1].

7 Conclusion

After presenting various modelling approaches to simulate population development and an introduction into agent-based modelling, the main part of this thesis concentrated on the expansion of the existing demographic GEPOC model with internal migration to simulate regional population changes. Therefore three basic upgrades were implemented and afterwards enhanced with respect to region-specific external migration. Thus, six different agent-based modelling approaches to simulate regional population developments with respect to internal migration were implemented and analysed. All models are constructed to handle an arbitrary amount of regions, as long as data for the migration flows between them is provided. The analysis and validation was undertaken for the federal states of Austria. In conclusion the inter-regional migration model with region-specific external migration turned out to be the best performing one, regarding both computation time and accuracy of the results.

As it has been shown in this work, internal migration has a high impact on simulating regional population numbers accurately. Considering this, Ravenstein's [18] first law of migration stating that most migrations are characterized by short distances is still valid, although the range of such short distances might have grown. His fifth law "migrants proceeding long distances generally go by preference to one of the great centres of commerce or industry" corresponds with the findings of the improved models with region-specific external migrations, showing that regional population development cannot be simulated accordingly without region-specific external migration. Still the necessity of additional data for this enhancement has to be considered.

The generation of accurate population development forecasts is very difficult due to the unavailability of data concerning future migration trends. As long as quickly changeable federal political decisions can have great impacts on internal migration flows, reliable future migration data will be hard to generate. Exemplary, in [16] the story of an 86-year-old Styrian women needing permanent care can be found: She is not allowed to move to a nursing home in Vienna to be close to her daughter, because fostering and its costs are a matter of the federal states. This is only one example of the large amount of influences to be considered. Nonetheless the simulation of the future population development of the Austrian federal states was attempted using linear regression.

The developed agent-based models can be further enhanced and used for many different applications. A possible usage lies in the analysis of probable developments of local labour markets. Various other possibilities lie in planning regional infrastructure like schools, retirement homes, hospitals and public transport.

Nomenclature

ABM Agent-based Modelling

BR Biregional Model

DEXHELPP Decision Support for Health Policy and Planning

GEPOC Generic Population Concept

IRM Inter-regional Migration Model

MAP Migration Age Profile Model

MP Migration Pool Model

rem Region-specific External Migration

List of Figures

List of Tables

Bibliography

[1] Statistik Austria. Statcube. `http://statcube.at`. Accessed: 2018-07-18.

[2] Aude Bernard, Martin Bell, and Elin Charles-Edwards. Life-course transitions and the age profile of internal migration. *Population and Development Review*, 40(2):213–239, 2014.

[3] Martin Bicher, Barbara Glock, Florian Miksch, Günter Schneckenreither, and Niki Popper. Definition, validation and comparison of two population models for Austria. In *Book of Abstracts, 4th International Conference on Business, Technology and Innovation*, 2015.

[4] Eric Bonabeau. Agent-based modelling: Methods and techniques for simulation of human systems. *PNAS*, 99(3), May 2002.

[5] European Commission. Eurostat. `https://ec.europa.eu/eurostat/data/database`. Accessed: 2018-09-18.

[6] Petra A de Jong, Aleid E Brouwer, and Philip McCann. Moving up and down the urban hierarchy: age-articulated interregional migration flows in the Netherlands. *The Annals of Regional Science*, 57(1):145–164, 2016.

[7] Josuah M. Epstein. Agent-Based Computational Models And Generative Social Science. *John Wiley & Sons, Inc.*, 4(5), May 1999.

[8] Barbara Glock. Ein System Dynamics Modell zur Prävalenz von Adipositas in Österreich. Master's thesis, Technical University Vienna, 2014.

[9] Gerhard Hanappi. Agent-based modelling. history, essence, future. *PSL Quarterly Review*, 70(283):449–472, 2017.

[10] Yuval Noah Harari. *Eine kurze Geschichte der Menschheit.* DVA, 2013.

[11] Andrew M Isserman, David A Plane, Peter A Rogerson, and Paul M Beaumont. Forecasting interstate migration with limited data: A demographic-economic approach. *Journal of the American Statistical Association*, 80(390):277–285, 1985.

[12] Nicolas R. Jennings. On agent-based software engineering. *Artificial Intelligence*, 117, September 1999.

[13] Barbara Lee Keyfitz and Nathan Keyfitz. The McKendrick partial differential equation and its uses in epidemiology and population study. *Mathematical and Computer Modelling*, 26(6):1–9, 1997.

[14] Charles B Nam, William J Serow, and David Sly. *International handbook on internal migration*. Greenwood Publishing Group, 1990.

[15] K Bruce Newbold. Counting migrants and migrations: Comparing lifetime and fixed-interval return and onward migration. *Economic Geography*, 77(1):23–40, 2001.

[16] ORF.at. Tochter kämpft um Pflegeheimwechsel für Mutter. `https://wien.orf.at/news/stories/2923301/`. Accessed: 2018-07-07.

[17] David A Plane. Requiem for the fixed-transition-probability migrant. *Geographical Analysis*, 25(3):211–223, 1993.

[18] Ernst Georg Ravenstein. The laws of migration. *Journal of the statistical society of London*, 48(2):167–235, 1885.

[19] Andrei Rogers. Requiem for the net migrant. *Geographical analysis*, 22(4):283–300, 1990.

[20] Elena Romero and M. Carmen Ruiz. Proposal of an agent-based analytical model to convert industrial areas in industrial eco-systems. *Elsevier*, August 2013.

[21] Robert E Shannon. Systems simulation; the art and science. Technical report, 1975.

[22] Der Standard. Wien wird doch erst 2026 zur Zwei-Millionen-Metropole. `https://derstandard.at/2000068346185/Weniger-Wachstum-Wien-wird-erst-2026-zur-Zwei-Millionen-Metropole`. Accessed: 2018-10-11.

[23] Nicole Van der Gaag, Evert Van Imhoff, and Leo Van Wissen. Internal migration scenarios and regional population projections for the European Union. *International Journal of Population Geography*, 6(1):1–19, 2000.

[24] Nicole Van der Gaag and Leo Van Wissen. Economic determinants of internal migration rates: a comparison across five European countries. *Tijdschrift voor economische en sociale geografie*, 99(2):209–222, 2008.

[25] Tom Wilson. The impact of education-bound mobility on inter-regional migration age profiles in Australia. *Applied Spatial Analysis and Policy*, 8(4):371–391, 2015.

[26] Tom Wilson and Martin Bell. Comparative empirical evaluations of internal migration models in subnational population projections. *Journal of Population Research*, 21(2):127, 2004.