



dwh
technical solutions
simulation services



TECHNISCHE
UNIVERSITÄT
WIEN

GEPOC Parameters

Open Source Parametrisation and Validation for Austria

Version 2.0

November 4, 2025

Martin Bicher^{1,3,*}, Maximilian Viehauser^{2,3}, Daniele Giannandrea^{2,3}, Hannah Kastinger³,
Dominik Brunmeir^{1,3}, Claire Rippinger³, Christoph Urach^{2,3}, Niki Popper^{1,2,3}

¹ TU Wien, Institute of Information Systems Engineering, Favoritenstraße 9-11, 1040 Vienna, Austria

² TU Wien, Institute of Statistics and Mathematical Methods in Economics, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria

³ dwh GmbH, dwh simulation services, Neustiftgasse 57-59, 1070 Vienna, Austria

*Correspondence: martin.bicher@dwh.at; martin.bicher@tuwien.ac.at

Abstract

GEPOC, short for Generic Population Concept, is a collection of models and methods for analysing population-level research questions. For the valid application of the models for a specific country or region, stable and reproducible data processes are necessary, which provide valid and ready-to-use model parameters. This work contains a complete description of the data-processing methods for computation of model parameters for Austria, based exclusively on freely and publicly accessible data. In addition to the description of the source data used, this includes all algorithms used for aggregation, disaggregation, fusion, cleansing or scaling of the data, as well as a description of the resulting parameter files. The document places particular emphasis on the computation of parameters for the most important GEPOC model, GEPOC ABM, a continuous-time agent-based population model. An extensive validation study using this particular model was made and is presented at the end of this work.

Contents

1	Introduction	4
2	Model Parameters	4
2.1	GEPOC ABM	5
2.2	GEPOC ABM Geography	6
2.3	GEPOC ABM IM	7
3	Demographic Terminology and Relations	8
3.1	Regional-Level and Identifiers	8
3.2	Census	10
3.2.1	Population P	10
3.2.2	Births B , Deaths D , Emigrants E , Immigrants I	10
3.2.3	Births by Mother B_m	10
3.2.4	Internal Migrants M and Net Internal Migration ΔM	11
3.2.5	Aggregation	11
3.3	Event Probabilities and Rates	12
3.3.1	Event Probabilities X^p	12
3.3.2	Event Rates X^r	12
3.3.3	Probabilities from Rates and Census	13
3.3.4	Life Tables and Life-Expectancy LE	15
3.4	Demographic Balance Equations	17
4	Disaggregation Algorithms	18
4.1	One-Sided Disaggregation	18
4.1.1	Proportional Disaggregation	18
4.1.2	Integer-Valued Disaggregation	19
4.2	Two-Sided Disaggregation	20
4.2.1	Disaggregation of Marginal Distributions	20
4.2.2	Disaggregation of Integer-Valued Marginals (Concepts)	22
5	Source Data	23
5.1	Sources for GEPOC ABM Parametrisation	23
5.2	Sources Specific for GEPOC ABM Geography Parametrisation	28
5.3	Sources Specific for GEPOC ABM IM Parametrisation	29
6	GEPOC Parameter Calculation	31
6.1	Population $\hat{P}$	31
6.2	Births $\hat{B}_m, \hat{B}$ and Probability of a Male Child $\hat{\alpha}_m$	35
6.2.1	Births by Age of the Mother $\hat{B}_m$	36

6.2.2	Birth Probability of a Male Child $\hat{\alpha}_m$	38
6.2.3	Births $\hat{B}$	39
6.3	Deaths $\hat{D}$	39
6.4	Emigrants $\hat{E}$	43
6.5	Immigrants $\hat{I}$	46
6.6	Birth-, Death-, and Emigration probabilities $\hat{B}^p, \hat{D}^p, \hat{E}^p$	49
6.7	Region-Families (A_j^{rx})	51
6.7.1	Multi-Polygon Processing	51
6.7.2	Raster Image for $A^{r_{min}}$	53
6.8	Internal Migrants $\hat{O}D, \hat{I}E, \hat{I}I, \hat{M}$	53
6.8.1	Interregional Flows $\hat{O}D$	53
6.8.2	Internal Emigrants $\hat{I}E$	55
6.8.3	Internal Immigrants $\hat{I}I$	57
6.8.4	Internal Migrants by Age, Origin and Destination $\hat{M}$	59
6.9	Internal Emigration Probability $\hat{I}E^p$	60
7	Validation	62
7.1	Population	63
7.1.1	Comparison with Source 5.1 (Bevölkerungsstand)	63
7.1.2	Comparison with Source 5.3 (Bevölkerung zum Jahresanfang 1952 bis 2101)	65
7.2	Births and Deaths	73
7.2.1	Comparison with Source 5.9 (Demographische Zeitreihenindikatoren) - Births	73
7.2.2	Comparison with Source 5.9 (Demographische Zeitreihenindikatoren) - Deaths	73
7.2.3	Comparison with Source 5.7 (Bevölkerungsbewegung 1961 bis 2100)	76
7.3	Emigrants and Immigrants	78
7.3.1	Comparison with Source 5.4 (Wanderungen mit dem Ausland ab 2002 nach Alter, Geschlecht und Staatsangehörigkeit)	78
7.3.2	Comparison with Sources 5.5 and 5.6 (Wanderungen mit dem Ausland von 2002 bis 2014 (ab 2015) nach Altersgruppen, Gemeinde und Staatsangehörigkeit)	78
7.3.3	Comparison with Source 5.7 (Bevölkerungsbewegung 1961 bis 2100 nach Bundesland, Bewegungsarten und Szenarien)	81
7.4	Internal Migrants	82
7.4.1	Comparison with Source 5.1 (Bevölkerungsstand) and 5.3 (Bevölkerung zum Jahresanfang 1952 bis 2101)	82
7.4.2	Comparison with Source 5.15 (BWanderungen innerhalb Österreichs ab 2002)	82
7.4.3	Comparison with Source 5.14 (Wanderungen innerhalb Österreichs ab 2002)	84
7.4.4	Comparison with Source 5.7 (Bevölkerungsbewegung 1961 bis 2100)	87
7.5	Additional Validation Scenarios	91
7.5.1	Simulation Scenario SC5 (Long Run).	91
7.5.2	Simulation Scenarios SC6 (Monthly Steps) and SC7 (Scale 0.1).	91

A	Appendix	96
A.1	Validation Plots - Population	96
A.2	Validation Plots - Births	105
A.3	Validation Plots - Deaths	112
A.4	Validation Plots - Emigrants	119
A.5	Validation Plots - Internal Migration	126

1 Introduction

In the present document, we will describe relevant data processes of GEPOC (Generic Population Concept), a collection of models and methods for analysing population-level research questions. We will lay a special focus on the parametrisation of the agent-based model GEPOC ABM, in specific, GEPOC ABM Version 2.2 and the two modules GEPOC ABM Geography, necessary for spatial analysis, and GEPOC ABM IM, necessary for internal migration analysis. We refer to the openly available model documentation [3] for details on the model and to [6] for information about Version 1.0 of the data processes. The model will be parametrised for the country Austria, based exclusively on freely available data. The data source is primarily Statistics Austria and its different publication platforms for open data.

The document contains

1. the different source data sets, their origin, their content and their license,
2. methods for aggregation, disaggregation, fusion, cleansing or scaling of the data to be suited as model input,
3. the different derived parameter values, which can be used as model input, and
4. a rigorous quantitative validation of the model with the presented parameter values.

The document is structured as follows: First, in Section 2, we state the parameters required by GEPOC ABM, most of which are equally relevant to the other GEPOC models (e.g. GEPOC SD [4], GEPOC PDE [5]). Sections 3 and 4 include the methodological background of the parameter calculation. The prior contains specification of important demographic terminology and states the most important demographic relations, balance equations and formulas. The latter contains relevant own-developed generic algorithms, which are not limited to demography in application. Equipped with the relevant terminology, formulas and algorithms, and the source data, in Section 5, the parameter calculation process follows in Section 6, which we regard as the core part of the document. It displays how the different source data, which all have different levels of aggregation in temporal, spatial, age, and sex resolution, are processed to get one high-quality set of model parameters. Finally, the validation section 7 provides evidence that the derived parameter values lead to quantitatively valid simulations.

We want to emphasise that in GEPOC, as well as in the census data, **sex**, i.e. *female*, *male*, is solely interpreted from the biological point of view. In GEPOC, we regard a person as female if it is, according to their biological properties, capable of producing offspring. It is important to mention that this does not necessarily reflect the person's gender (which is not implemented in GEPOC ABM).

2 Model Parameters

In this section we quickly summarise the parameters required for successful parametrisation of GEPOC ABM. Note that these can be found in [3], Sections 3.3.3, 4.3.3, and 5.3.3.

To differentiate the model parameter from the actual demographic quantity (see Section 3), we mark the parameter with a hat symbol: $\hat{(\cdot)}$.

Since we will compute parameters for GEPOC ABM Geography for different regional-levels (see 3.1), the parameters for GEPOC ABM directly result for the country-level.

2.1 GEPOC ABM

Table 1: Parameters of GEPOC ABM. Note that the sex variable s in GEPOC ABM is interpreted from a perspective of reproduction. Agents which are in principle capable of reproduction are called *female*, all other are called *male*.

Parameter	Dimensions	Unit	P. Space	Interpretation
$\hat{\alpha}_m$	-	probability	$[0, 1]$	probability for male person-agent at birth
$\hat{a}_{max}$	-	years	$\mathbb{N}/\{0\}$	maximum age regarded in the parameters
$\hat{P}(y, s, a)$	$y \in \{y_0, \dots, y_e\}, a \in \{0, \dots, a_{max}\}, s \in \{\text{male}, \text{female}\}$	persons	$\mathbb{N} \cup \{0\}$	total population per age a , sex s at the start of year y .
$\hat{I}(y, s, a)$	$y \in \{y_0, \dots, y_e\}, a \in \{0, \dots, a_{max}\}, s \in \{\text{male}, \text{female}\}$	persons	$\mathbb{N} \cup \{0\}$	total immigrants with age a (at time of immigration), sex s in the course of year y .
$\hat{D}^p(y, s, a)$	$y \in \{y_0, \dots, y_e\}, a \in \{0, \dots, a_{max}\}, s \in \{\text{male}, \text{female}\}$	probability	$[0, 1]$	probability of a person with sex s , who has had its a -th birthday in year y , to die before its $a+1$ -st birthday.
$\hat{E}^p(y, s, a)$	$y \in \{y_0, \dots, y_e\}, a \in \{0, \dots, a_{max}\}, s \in \{\text{male}, \text{female}\}$	probability	$[0, 1]$	probability of a person with sex s , who has had its a -th birthday in year y , to emigrate before its $a+1$ -st birthday.
$\hat{B}^p(y, s, a)$	$y \in \{y_0, \dots, y_e\}, a \in \{0, \dots, a_{max}\}, s \in \{\text{male}, \text{female}\}$	probability	$[0, 1]$	probability of a female person, who has had her a -th birthday in year y , to give birth to a child before her $a+1$ -st birthday. This probability must compensate for multiple-births, which are not depicted in the model.

2.2 GEPOC ABM Geography

Table 2: Additional parameters of GEPOC ABM Geography extending Table 1.

Parameter	Dimensions	Unit	P. Space	Interpretation
r_x	$x \in \{0, d, e, b, i, min\}$	name	various	regional-levels used for initialisation, death, emigration, birth and immigration processes.
$A_j^{r_x}$	$x \in \{0, d, e, b, i, min\}, j \in \{1, \dots, q_x\}$	$\{(long, lat)\}$	$\subset \mathbb{R}^2$	Specification of the region-families matching to the specified regional-levels with a suitable area-status. Hereby, regional-level r_x has q_x regions.
$\hat{P}(y, i, s, a)$	$y \in \{y_0, \dots, y_e\}, i \in \{1, \dots, q_0\}, a \in \{0, \dots, a_{max}\}, s \in \{\text{male, female}\}$	persons	$\mathbb{N} \cup \{0\}$	total population per region $A_i^{r_0}$, age a , sex s at the start of year y .
$\hat{I}(y, i, s, a)$	$y \in \{y_0, \dots, y_e\}, i \in \{1, \dots, q_i\}, a \in \{0, \dots, a_{max}\}, s \in \{\text{male, female}\}$	persons	$\mathbb{N} \cup \{0\}$	total immigrants to region $A_i^{r_i}$ with age a (at time of immigration), sex s in the course of year y .
$\hat{D}^p(y, i, s, a)$	$y \in \{y_0, \dots, y_e\}, i \in \{1, \dots, q_d\}, a \in \{0, \dots, a_{max}\}, s \in \{\text{male, female}\}$	probability	$[0, 1]$	Probability of a person with sex s living in region $A_i^{r_d}$, who has had its a -th birthday in year y , to die before its $a+1$ -st birthday.
$\hat{E}^p(y, i, s, a)$	$y \in \{y_0, \dots, y_e\}, i \in \{1, \dots, q_e\}, a \in \{0, \dots, a_{max}\}, s \in \{\text{male, female}\}$	probability	$[0, 1]$	Probability of a person with sex s living in region $A_i^{r_e}$, who has had its a -th birthday in year y , to emigrate before its $a+1$ -st birthday.
$\hat{B}^p(y, i, s, a)$	$y \in \{y_0, \dots, y_e\}, i \in \{1, \dots, q_b\}, a \in \{0, \dots, a_{max}\}, s \in \{\text{male, female}\}$	probability	$[0, 1]$	Probability of a female person living in region $A_i^{r_b}$, who has had her a -th birthday in year y , to give birth to a child before her $a+1$ -st birthday. This probability must compensate for multiple-births which are not depicted in the model.

2.3 GEPOC ABM IM

Table 3: Additional parameters of GEPOC ABM IM extending Table 2.

Parameter	Dimensions	Unit	P. Space	Interpretation
r_{im}	-	name	various	regional-level used for internal migration.
A_j^{rim}	$j \in \{1, \dots, q_{im}\}$	$\{(lat, long)\}$	$\subset \mathbb{R}^2$	Specification of the regional set-families for internal migration.
$\hat{IE}(y, i, s, a)$	$y \in \{y_0, \dots, y_e\}, i \in \{1, \dots, q_0\},$ $a \in \{0, \dots, a_{max}\}, s \in \{\text{male}, \text{female}\}$	probability	$[0, 1]$	Probability of a person with sex s living in region i , who has had its a -th birthday in year y , to emigrate internally before its $a + 1$ -st birthday.
Interregional model				
$\hat{OD}(y, i, s, j)$	$y \in \{y_0, \dots, y_e\}, i \in \{1, \dots, q_{im}\}, s \in \{\text{male}, \text{female}\}, j \in \{1, \dots, q_{im}\}$	persons	$\mathbb{N} \cup \{0\}$	total migrants from region i to j with sex s in the course of year y .
Biregional model				
$\hat{II}(y, j, s, a)$	$y \in \{y_0, \dots, y_e\}, j \in \{1, \dots, q_{im}\}, s \in \{\text{male}, \text{female}\}, a \in \{0, \dots, a_{max}\}$	persons	$\mathbb{N} \cup \{0\}$	internal immigrants into region j with sex s and age a in the course of year y .
Full Regional model				
$\hat{M}(y, i, s, a, j)$	$y \in \{y_0, \dots, y_e\}, i \in \{1, \dots, q_{im}\}, s \in \{\text{male}, \text{female}\}, a \in \{0, \dots, a_{max}\}, j \in \{1, \dots, q_{im}\}$	persons	$\mathbb{N} \cup \{0\}$	internal migrants from region i into j with sex s and age a in the course of year y .

3 Demographic Terminology and Relations

3.1 Regional-Level and Identifiers

The common option to communicate a specific location in a given country is to specify the sub-region of the country in which the point lies in. The smaller the sub-region, the more accurately the statement describes the point. To standardise communication, there are various ways how the country is divided into multiple regions with comparable sizes.

Definition 3.1 (regional-level, region-id). *We denote the way how a country is divided into sub-regions as **regional-level**. Each region the country is divided into for a regional-level is identified by a specific **region-id**.*

In some situations it might be useful to compare regional-levels w.r. to how specific they describe a location:

Definition 3.2 (fine/coarse). *A regional-level is said to be more **fine** than the other, if every region of the latter can be split into regions of the prior. In this situation, the latter is also said to be more **coarse** than the prior.*

In Austria there are various well-known regional-levels used for different purposes, from division of legislative, executive, juristic competences up to sole statistical purposes. The most important ones are surely the nine federal-states. Their region-ids range from *AT-1* to *AT-9*, referring to the names of the states in alphabetical order. Internationally, the federal-states count to the NUTS-2 (Nomenclature des unités territoriales statistiques) regional-level. The Austrian contribution for the NUTS-1 level, which is coarser than NUTS-2, is not so often used. It refers to three combinations of federal-states: (*AT-1+AT-3+AT-9*, *AT-2+AT-6*, and *AT-4+AT-5+AT-7+AT-8*). The federal-state level can be refined into the 35 NUTS-3 regions or into the 95 political districts of Austria. Note that the district level is not finer than NUTS-3, since some districts are split into two different NUTS-3 regions. The 2000 Municipalities are finer than NUTS-3 and the district level and count to the internationally used LAU (local administrative unit) regions.

We furthermore explain, in detail, the regional-levels and corresponding region-ids used for the GEPOC parametrisation.

identifier	meaning	ID (ISO) structure
country	No regional resolution. Data is given for Austria as a total.	<i>AT</i>
federalstates	Data is given for the nine federal-states of Austria.	<i>AT-1</i> to <i>AT-9</i>
districts	Data is given for the roughly 100 (94 as of 2022) political districts in Austria. Note that Vienna, as a whole, is one of those.	Three digit ISO code. The first digit (i.e. the 100-digit) refers to the federalstate (<i>AT-1</i> to <i>AT-9</i>), the other two are (more less) ascending integers. E.g. <i>301</i> is the “first” district (Krems an der Donau (Stadt)) in AT-3 (Lower Austria). Due to created and removed districts, we find occasional gaps in the ascending order. E.g. district <i>613</i> (Mürzzuschlag) was removed and integrated into newly developed district <i>621</i> (Bruck-Mürzzuschlag) in 2013. Vienna, as a special case, has ISO code <i>900</i> .
districts_districts	Data is given for the 116 (as of 2022) NUTS3 regions in Austria. This is equivalent with the districts, yet Vienna is split into its 23 separate “Gemeindebezirke”.	Same as <i>districts</i> . ISO codes within Vienna are <i>901-923</i> .
municipalities	Data is given for the 1941 (as of 2022) municipalities in Austria. Note that Vienna, as a whole, is one of those.	Five digit ISO code whereas the first three refer to the political district. The last two digits are more less ascending integers analogous to the <i>district</i> -level. E.g. <i>30101</i> (Krems an der Donau) is (the only) municipality in district <i>301</i> ((Krems an der Donau (Stadt))). Following the scheme, Vienna has ISO code <i>90001</i>
municipalities_districts	Data is given for the $1940+23=1964$ (as of 2022) municipalities in Austria whereas Vienna is split into its 23 “Gemeindebezirke”.	Same as <i>municipalities</i> , yet Viennese “Gemeindebezirke” have codes <i>90101-92301</i> which also follows the scheme applied to the <i>districts_districts</i> level.
municipalities_registrationdistricts	To get a finer resolution within Vienna, data in the “Gemeindebezirke” are split into even finer regions - so called “Zaehlbezirke” (loosely translated, registration-districts). These are primarily used for statistical reasons. Anyway, this data includes $1940+247=2187$ (Gebietsstand 2022) regions.	Outside of Vienna: same as <i>municipalities</i> . Registration-districts within Vienna follow a different scheme: seven-digit ISO whereas the first five refer to the district (<i>municipalities_districts</i> level). The last two are ascending integers.

The statistical and political regions and their borders in Austria tend to change from time to time - the finer, the more often. E.g., municipalities change almost yearly, while the federal-states changed in 1987 for the last time, when Vienna was announced its own federal-state. As a result, any regional data must be accompanied with information for which time-span the data is valid w.r. to the political regions involved.

Definition 3.3 (area-status). *The term area-status (loosely translated from German “Gebietsstand”) refers to the year in which the geographical landscape of the country for a specific region-level is compatible with the given information.*

Note that the area-status of a data set does not necessarily imply that the data is also given only for this year. First, we have no breaches in the time-series if the geographical landscape remains unchanged for several years (e.g. time-series for the federal-states are consistent for over hundred years). Second, data which was collected under an outdated area-status could also be updated to a new one if the collection process allows it. This is currently done by Statistics Austria, which recomputes historic data to an updated area-status every year.

Before diving into the details of parameter calculation for GEPOC, we want to clarify details about the key quantities, our computations will be based on. Note that we still talk about (observing) the real system and not about a model.

3.2 Census

First of all, we discuss quantities which can be collected from the inhabitants by counting individuals.

3.2.1 Population P

For a given regional-level, $P(y, r, s, a)$ stands for the overall population with sex s (biological sex at birth, male/female) and age a (in years) in region-id r at day $y - 01 - 01$ (start of the year). For Austria, this refers to the state of the central register of residents (ZMR, Zentrales Melderegister) at the given date, which are all persons with an official primary place of residence in Austria. In the absence of better instruments to measure the inhabitants of Austria, and in absence of a better definition of “who counts as inhabitant of Austria” this data is denoted as ground-truth for parametrisation and validation of the simulation.

3.2.2 Births B , Deaths D , Emigrants E , Immigrants I .

For a given regional-level, $B(y, r, s, a)$, $D(y, r, s, a)$, $I(y, r, s, a)$, and $E(y, r, s, a)$ stand for the total number of new-born, died, immigrated, and emigrated persons with sex s and age a in the course of year y (i.e. between $y - 01 - 01$ and $y - 12 - 31$). Age a always refers to the age at the corresponding event, and r refers to the region the person lived before death and emigration, the person was born in, or the person migrated into according to the ZMR. Note that $B(y, r, s, a)$ is only nonzero for $a = 0$.

3.2.3 Births by Mother B_m .

For a given regional-level, $B_m(y, r, s, a)$ stands for the total number of newborns by sex s and age a of the mother in the course of year y (i.e. between $y - 01 - 01$ and $y - 12 - 31$). As before, age a refers to the age at the corresponding event, and r refers to the region the person lived at the time of birth. Note that $B_m(y, r, s, a)$ is only non-zero for $s = f$ (compare with the definition of sex, earlier in the text) and that

$$\sum_s B(y, r, s, 0) = \sum_a B_m(y, r, f, a).$$

3.2.4 Internal Migrants M and Net Internal Migration ΔM .

For a given regional-level, $M(y, r, s, a, r_2)$ refers to the number of persons with age a and sex s who migrated from region r into region r_2 in the course of year y . We furthermore define the net internal migration via

$$\Delta M(y, r, s, a) = \sum_{r_2} M(y, r_2, s, a, r) - M(y, r, s, a, r_2).$$

3.2.5 Aggregation

For all introduced quantities we use a very native notation for aggregation: $\forall X \in \{P, E, I, D, B, B_m, \Delta M\}$

$$\begin{aligned} X(y, r, s) &:= \sum_a X(y, r, s, a), \\ X(y, r, a) &:= \sum_s X(y, r, s, a), \\ X(y, s, a) &:= \sum_r X(y, r, s, a), \\ X(y, r) &:= \sum_s \sum_a X(y, r, s, a), \\ X(y, a) &:= \sum_s \sum_r X(y, r, s, a), \\ X(y, s) &:= \sum_r \sum_a X(y, r, s, a), \\ X(y) &:= \sum_r \sum_s \sum_a X(y, r, s, a). \end{aligned}$$

Note that this notation might not be mathematically sound, but it is highly useful for communication as long as we care about using the correct variable names. This notation is directly extended to internal-migrants, for which we also introduce new terms:

$$\begin{aligned} IE(y, r, s, a) &:= \sum_{r_2} M(y, r, s, a, r_2), \\ OD(y, r, s, r_2) &:= \sum_a M(y, r, s, a, r_2), \\ II(y, s, a, r_2) &:= \sum_r M(y, r, s, a, r_2), \end{aligned}$$

The terms stand for internal emigrants, migrants' origin-destination, and internal immigrants.

Finally, we also introduce an age cohort a_{max}^+ so that $\forall X \in \{P, E, I, D, B, \Delta M\}$

$$\begin{aligned} X(y, r, s, a_{max}^+) &:= \sum_{a \geq a_{max}} X(y, r, s, a), \\ M(y, r, s, a_{max}^+, r_2) &:= \sum_{a \geq a_{max}} M(y, r, s, a, r_2). \end{aligned}$$

This age class a_{max}^+ refers to all individuals with age higher or equal to a_{max} . Note that different quantities use different values of a_{max} .

3.3 Event Probabilities and Rates

Talking about total number of births, deaths, etc, we may think of shifting these quantities from the country-to the individual level in terms of rates and probabilities for the corresponding event.

3.3.1 Event Probabilities X^p

Defining individual event probabilities is not a straight forward task, since minor details may massively influence the outcome. In our work we define:

Definition 3.4 (probability of an event). *For a certain event, the quantity $X^p(y, r, s, a)$ stands for the probability that the event occurs to/for a person with sex s , which has its a^{th} birthday in the course of year y in region r , until the person turns $a + 1$. We will always indicate probabilities by a superscript “p”.*

This definition is in-line with the classic definition of the term *death-probability* as it is generally understood by statistics offices (compare with the online glossary of Statistics Finland [11]) and it matches the dynamic update concept of the GEPOC model.

In our studies, D^p , E^p , IE^p , and M^p refer to the probability that the person dies, emigrates, emigrates internally, and internally migrates to a certain destination. Probability B^p refers to the probability that a person gives birth to an offspring (can only be nonzero for $s = \text{female}$). Hereby it takes a special role, since the target persons for which the probability is specified, i.e. the female inhabitants, are not the ones who are recorded via B , namely the newborn children. Furthermore, the probability-concept does not make sense for immigration processes since the affected person is not part of the observed population before.

3.3.2 Event Rates X^r

Intuitively, dividing the number of events caused by a group of persons, by the size of the group, fulfils the requirements of a Laplace-space and can accordingly be interpreted as a probability. Therefore, we would expect that the term

$$\frac{X(y, r, s, a)}{P(y, r, s, a)}$$

provides proper insights into the likelihood of the event X per individual. Unfortunately, it is not that easy, because the group of persons responsible for the events recorded by $X(y, r, s, a)$ is not $P(y, r, s, a)$. A great part of the individuals responsible for the events already had their a -th birthday in year y before the event. Therefore, they have been members of the cohort $P(y - 1, r, s, a)$ and not $P(y, r, s, a)$, meaning that the actual denominator must have been larger. Nevertheless, the expression is meaningful as it describes an average rate of the event in year y , in particular if the average size of the cohort P_{avg} in the course of the year is used instead of the size P at the start of the year:

Definition 3.5 (average rate of an event). *For a certain event,*

$$X^r(y, r, s, a) := \frac{X(y, r, s, a)}{P_{avg}(y, r, s, a)} \tag{1}$$

defines the average rate of the event in the course of year y . Hereby P_{avg} describes the average size of the population over the course of the year.

In case the census does not provide average population information, it can be approximated by the arithmetic mean of the populations on new-year

$$P_{avg}(y, r, s, a) \approx \frac{P(y, r, s, a) + P(y + 1, r, s, a)}{2}. \quad (2)$$

In the following, we will apply this approximation for P_{avg} for all parameter computations.

Although X^r is not directly applicable as a probability, it is nevertheless often applied as demographic quantity. For example, the age-dependent rate of fertility $B_m^r(y, r, f, a)$ multiplied by 1000 is an important demographic indicator stating the average number of newborns per 1000 women with age a . It helps defining meaningful derived indicators:

Definition 3.6 (total fertility rate TFR and mean age at childbearing MAC). *With the average fertility rate B_m^r , the total fertility rate defines as*

$$TFR(y, r, f) = \sum_a B_m^r(y, r, f, a), \quad (3)$$

and the average fertility age defines as

$$MAC(y, r, f) = \frac{\sum_a a B_m^r(y, r, f, a)}{\sum_a B_m^r(y, r, f, a)}. \quad (4)$$

The prior estimates the number of newborns in a woman's lifetime, the latter the average age of a woman giving birth to a child.

3.3.3 Probabilities from Rates and Census

Due to mentioned population-group mismatch due to ageing, precise calculation of actual probabilities from census data alone is usually impossible. However, there are well established methods which give good estimates.

One of the most important ones dates back to life-science pioneer William Farr. In a paper published in 1859 [10], he presented relevant concepts for the computation of death-tables for Great Britain, including an important method for the estimation of death probabilities. The formula, found on page 848, is based on the idea that the average rate of mortality D^r , as defined in the section before, computes the rate of death per lived year of life, but not per individual. Farr assumed that the population of the cohort with age a remains constant and that, therefore, any of the D individuals lost by death must have been replaced by an equivalent new individual in the course of the year to maintain the total cohort size. Therefore, the total number of individuals under observation is larger than the overall lived years of life. Farr found the formula

$$D^p(y, r, s, a) = 1 - \frac{1 - \frac{1}{2}D^r(y, r, s, a)}{1 + \frac{1}{2}D^r(y, r, s, a)} = \frac{D^r(y, r, s, a)}{1 + \frac{1}{2}D^r(y, r, s, a)}. \quad (5)$$

to compensate for this bias. The formula can be reasoned by comparing an individual-level ratio with a population-level ratio for “deaths-per-person-year”:

When an individual turns a , its chance for surviving the upcoming year is $(1 - D_a^p)$, and its probability of death is D_a^p . Assuming that the time of death is uniformly distributed in the course of the individual's life-year, the average proportion of the year spent by the individual is

$$(1 - D^p(y, r, s, a)) + \frac{1}{2}D^p(y, r, s, a) = 1 - \frac{1}{2}D^p(y, r, s, a).$$

Therefore, we observe

$$\frac{D^p(y, r, s, a)}{1 - \frac{1}{2}D^p(y, r, s, a)}$$

deaths-per-person-year for the individual.

On the population scope, the age band of individuals with age a is assumed to be of constant size P_a . Therefore, the total years of life spent by persons of age a over the observed year is precisely P_a . With D_a recorded deaths of persons with age a , there are $\frac{D_a}{P_a}$ deaths per person year on the population-scope. Since the scopes must be identical, we have

$$\frac{D_a}{P_a} = \frac{D^p(y, r, s, a)}{1 - \frac{1}{2}D^p(y, r, s, a)}$$

which can be transformed to

$$D^p(y, r, s, a) = \frac{\frac{D_a}{P_a}}{1 + \frac{1}{2}\frac{D_a}{P_a}}.$$

Approximating $\frac{D_a}{P_a} \approx D^r(y, r, s, a)$ gives the stated formula.

Clearly, the formula incorporates several inaccuracies, some of which can be made more accurate without major computations. One of the most important ones refers to the uniformity of the time of death in the course of the year. In particular for newborn infants, the rate of mortality in the first months of life is a lot larger than later on. Therefore, the average time an individual that died with age 0 spent in the cohort is usually a lot smaller than 1/2. Leaving this average time free for parametrisation leads to the modern-time formulation of the death-rate formula:

Theorem 3.1 (Farr's Death Rate Formula (modern version)). *Let $1 - \alpha(a)$ stand for the expected year-of-life spent by a person in its age cohort, given that the person is going to die, then*

$$\frac{D^r(y, r, s, a)}{1 + \alpha(a)D^r(y, r, s, a)} = \frac{D(y, r, s, a)}{P_{avg}(y, r, s, a) + \alpha(a)D(y, r, s, a)}, \quad (6)$$

is a good model for the probability of death $D^p(y, r, s, a)$ as given by Definition 3.4. Hereby, P_{avg} stands for the average population over the course of year y .

As mentioned, it is legitimate to define $\alpha(a) = 1/2$ for $a > 0$, however, since infants are way more likely to die within the first few weeks of their life than in the rest of their first life-year, typically $\alpha(0) > 0.9$ is applied. Moreover, it is worth mentioning that the formula is equally meaningful when applied to the a_{max}^+ cohort to get a constant death probability $D^p(y, r, s, a_{max}^+)$ for any individual with age $\geq a_{max}$.

Analogous to the mentioned fertility rates the formula in Theorem 3.1 is an internationally recognized concept for the computation of a demographic indicator. Besides being a good approximation, it is also the one used by Statistics Austria to compute mortality- and life-tables [12]. As a result, it will be used for any parametrisation strategy involving statistical indicators like death probabilities and mortality tables. For model parametrisation from census data, the formula can be improved:

One of the key advantages of the formula is that the probabilities of death can be computed from census information from one single year. This became possible by the assuming

$$\frac{D_a}{P_a} = \frac{D(y, r, s, a)}{P_{avg}(y, r, s, a)} = D^r(y, r, s, a).$$

However, the year over which the deaths-per-life-year are computed on the individual-scope, namely the individual's $a + 1$ -st year of life, is not equivalent with the one over which the rate of mortality is computed,

namely Jan 1st to Jan 1st. In particular, if the probabilities are under subject of high dynamics, this results in a half-year time-lag. In case one is not bound to apply data from a single year, we can compensate the bias by taking the arithmetic mean

$$\frac{1}{2} \frac{D(y, r, s, a)}{P_{avg}(y, r, s, a) + \alpha(a)D(y, r, s, a)} + \frac{1}{2} \frac{D(y+1, r, s, a)}{P_{avg}(y+1, r, s, a) + \alpha(a)D(y+1, r, s, a)}. \quad (7)$$

Since all probabilities in GEPOC ABM are defined in the same manner as the death probability, we may also apply the formula for parametrisation of any probability occurring in the update in a slightly modified form:

Corollary 3.1 (Farr Formula (model parametrisation)). *Let $X \in \{B, D, E, IE\}$, and $Q(y, r, s, a) := E(y, r, s, a) + D(y, r, s, a)$ as the total number of individuals leaving the age cohort in the course of year y , and let y_N refer to the last available year in the population census. With*

$$P_{avg}(y, r, s, a) := \frac{1}{2} (P(y, r, s, a) + P(\max(y+1, y_N), r, s, a)), \quad (8)$$

$$X^{p,-}(y, r, s, a) := \frac{X(y, r, s, a)}{P_{avg}(y, r, s, a) + \frac{1}{2}Q(y, r, s, a)}, \text{ and} \quad (9)$$

$$X^{p,+}(y, r, s, a) := \frac{X(\max(y+1, y_N), r, s, a)}{P_{avg}(\max(y+1, y_N), r, s, a) + \frac{1}{2}Q(\max(y+1, y_N), r, s, a)}, \quad (10)$$

we get

$$\frac{1}{2} X^{p,-}(y, r, s, a) + \frac{1}{2} X^{p,+}(y, r, s, a) \quad (11)$$

as a good approximation for the probability $X^p(y, r, s, a)$ of the event according to Definition 3.4.

The formula can be reasoned for $X \neq D$ using the same ideas as presented before, namely comparing the individual occurrences of the event X per individual life-year with the corresponding population-scope observation. Since the model does not regard skewed infant-deaths or emigrations, usage of any $\alpha(a)$ different from 1/2 is not necessary. Finally, the model does, mechanistically, not differentiate between deaths and emigrants - in both cases, agents leave the model and future events are cancelled. As a result, joining D and E as a common census for individuals leaving the model boundaries is useful.

3.3.4 Life Tables and Life-Expectancy LE

Under $LE(y, r, s, a)$ we understand the life expectancy of a person with sex s , living in region r , at the point of its a -th birthday in year y . Note that this is usually a rising function when increasing a and y at the same time, i.e. $LE(y, r, s, a) < LE(y+1, r, s, a+1)$, since already surviving for a years reduces the chances for dying young. The function is not necessarily increasing when y is fixed, e.g. due to medical advancements.

Alike the computation of death/event probabilities, creating a mapping between deaths/death rates (probabilities) and the life expectancy is not straightforward either and requires assumptions. The most common way to approach this problem is by applying the concept of life-/death-/mortality-tables from survival analysis.

These tables are solely based on the death probabilities D^p computed from overall deaths with Farr's Death Rate Formula, Theorem 3.1. For given sex s , region r , and year y , the table is filled as follows (Sullivan method [14]):

1. First of all, we define the series of death probabilities

$$q_i := \begin{cases} D^p(y, r, s, i), & 0 \leq i \leq a_{max} \\ D^p(y, r, s, a_{max}^+), & i > a_{max} \end{cases} \quad (12)$$

2. In the next step, we compute population series l and absolute death series d recursively. We start with an arbitrary fictional start population l_0 (typically $l_0 = 100000$ is chosen):

$$d_i := l_i \cdot q_i \quad (13)$$

$$l_{i+1} := l_i - d_i. \quad (14)$$

This leads $l_i = l_0 \cdot \prod_{j=0}^{i-1} (1 - q_j)$ and $d_i = l_0 \cdot q_i \cdot \prod_{j=0}^{i-1} (1 - q_j)$.

3. Furthermore, the population at-risk L is computed. Again, we find α , as defined in Theorem 3.1, in this formula:

$$L_i := l_i - \alpha(i) d_i. \quad (15)$$

This leads to the explicit formula

$$L_i = l_0 (1 - \alpha(i) q_i) \prod_{j=0}^{i-1} (1 - q_j).$$

4. We furthermore compute the cumulative number of population years at risk T :

$$T_i := \sum_{j=i}^{\infty} L_j, \quad (16)$$

which can be written explicitly as

$$T_i := l_0 \sum_{j=i}^{\infty} (1 - \alpha(j) q_j) \prod_{k=0}^{j-1} (1 - q_k).$$

Note that the infinite sum can be computed analytically, since $q_k = q_{a_{max}}$, $\alpha(k) = 0.5$ for $k \geq a_{max}$. Thus, for $k > a_{max}$, L_i forms a geometric series with

$$\begin{aligned} \frac{l_i}{l_{i-1}} &= (1 - q_{a_{max}}) \Rightarrow \sum_{i=a_{max}}^{\infty} l_i = l_{a_{max}} \frac{1}{q_{a_{max}}} \\ L_i &= l_i - \alpha(i) l_i q_i = l_i (1 - \alpha(i) q_i) \Rightarrow \sum_{i=a_{max}}^{\infty} L_i = l_{a_{max}} \frac{1 - \alpha(a_{max}) q_{a_{max}}}{q_{a_{max}}} = l_{a_{max}} \frac{1 - \frac{q_{a_{max}}}{2}}{q_{a_{max}}}. \end{aligned}$$

5. Finally, we compute the life expectancy vector e by

$$e_i := \frac{T_i}{l_i}. \quad (17)$$

Putting it all together, we find the following closed formula

Definition 3.7 (Life Expectancy Formula). *The life expectancy of a person with sex s , living in region r at the point of its a -th birthday in year y can be approximated with*

$$LE(D^p, y, r, s, a) = \frac{\sum_{j=a}^{\infty} (1 - \alpha(j)D^p(y, r, s, j)) \prod_{k=0}^{j-1} (1 - D^p(y, r, s, k))}{\prod_{j=0}^{a-1} (1 - D^p(y, r, s, j))}. \quad (18)$$

Most importantly, l_0 cancels out which renders its choice irrelevant for computing the life expectancy.

3.4 Demographic Balance Equations

In the next step, we will state some balance equations, which can be used to calculate relations between the different quantities. The first balance refers to the overall population.

Corollary 3.2 (Overall Population Balance). *For all y and s ,*

$$P(y + 1, s) = P(y, s) + B(y, s) + I(y, s) - D(y, s) - E(y, s). \quad (19)$$

This equation does not require additional explanation, but we want to emphasise that it does not at all hold without full summation over age and region.

Corollary 3.3 (Region-Specific Population Balance). *For all y , r and s ,*

$$P(y + 1, r, s) = P(y, r, s) + B(y, r, s) + I(y, r, s) - D(y, r, s) - E(y, r, s) + \Delta M(y, r, s). \quad (20)$$

With introduction of the spatial component, the (net) internal migration becomes relevant.

Development of age-dependent balance equations is possible, however, additional assumptions need to be made (e.g. how demographic events are distributed over the life-year). Moreover, these formulas turned out to be numerically unstable with respect to data errors. As a result, we will not state any of those.

4 Disaggregation Algorithms

In the last section, several formulas have been introduced, which can be used to compute probability parameters for the model when census information is provided. With these formulas, GEPOC should be configurable for as long a period as possible, extending not only as far back into the past as possible, but also into the future. This process, in turn, brings with it other difficulties that have nothing to do with demographics per-se: Since both historical and forecast census information are not available in the same resolution as current data, they must be harmonised. If the coarsest common level of detail was chosen for harmonisation, an enormous loss of information for the current data would result, making the model unnecessarily inaccurate. Accordingly, the opposite approach is chosen: data with lower resolution is **disaggregated** to the finest level of resolution using assumptions about the distribution.

In the following, we distinguish between two problem statements. In the *one-sided disaggregation* problem, one data-set is strictly finer than the other, and the goal of the procedure is to elevate the resolution of the coarser data-set to that of the finer one. In the *two-sided disaggregation* problem, both data-sets each have resolution deficiencies in different dimensions, and the final result harmonises the data on the joint finest resolution. The latter can be regarded as estimating a distribution given its marginals, and occurs in the computation of internal migration parameters.

4.1 One-Sided Disaggregation

In the one-sided problem, one data-set is strictly finer than another one. Let $\Psi_1 := (Y_1, R_1, S_1, A_1)$ be the set of all year-region-sex-age tuples of the data entries $X_1(y, r, s, a), (y, r, s, a) \in \Psi_1$ of the fine-grained data and $\Psi_2 := (Y_2, R_2, S_2, A_2)$ be the indices to the coarse data X_2 , then there is an aggregation mapping $f : \Psi_1 \rightarrow \Psi_2$ so that X_1 can be aggregated to the coarse level by summation over all of the inputs of f with the same image. That means, let

$$A(y, r, s, a) := \{(y', r', s', a') \in \Psi_1 : f((y', r', s', a')) = (y, r, s, a)\},$$

then

$$X_{1,agg}(y, r, s, a) = \sum_{(y', r', s', a') \in A(y, r, s, a)} X_1(y', r', s', a').$$

While this computation is uniquely defined, the inverse problem, i.e. bringing X_2 to the fine level $X_{2,disagg}$, requires additional assumptions. Clearly,

$$X_2(y, r, s, a) = \sum_{(y', r', s', a') \in A(y, r, s, a)} X_{2,disagg}(y', r', s', a'), \quad (21)$$

must be fulfilled, apart from that, however, the computation of $X_{2,disagg}$ has no further constraints. To solve this problem, we usually define a distribution which the disaggregated data should follow, i.e. for every $(y, r, s, a) \in \Psi_2$:

$$\begin{aligned} \vec{X}_{2,disagg} &:= (X_{2,disagg}(y', r', s', a'))_{(y', r', s', a') \in A(y, r, s, a)} \sim \\ &\quad (p(y', r', s', a'))_{(y', r', s', a') \in A(y, r, s, a)} =: \vec{p}(y, r, s, a) \in (\mathbb{R}^+)^{|A(y, r, s, a)|} \end{aligned} \quad (22)$$

In most cases, this distribution is gained from the fine-resolution data-set.

4.1.1 Proportional Disaggregation

The most straightforward strategy to disaggregate the value will be called *proportional disaggregation*:

Algorithm 4.1 (Proportional Disaggregation). *Let $X_2(y, r, s, a)$ be a value to be disaggregated and $\vec{p}(y, r, s, a)$ a vector which the disaggregated values should follow, then*

$$\vec{X}_{2,disagg} \leftarrow \vec{p}(y, r, s, a) \frac{X_2(y, r, s, a)}{\sum \vec{p}(y, r, s, a)}. \quad (23)$$

It is clear that the disaggregated vector fulfils conditions (21) and (22) perfectly.

Unfortunately, there are still cases where a different algorithm is required, namely when it comes to conserve whole numbers. Both the initial population and the immigration process require integer-valued parameters.

4.1.2 Integer-Valued Disaggregation

The problem of disaggregating a given natural number of elements $X_2(y, r, s, a) \in \mathbb{N}$ onto a number of parties according to a distribution vector $\vec{p}$ into an integer-valued vector $\vec{X}_{2,disagg}$ is usually called an apportionment problem. This problem has a wide range of applications, most famously in politics, to match the result of demographic elections onto a finite number of seats in the parliament. Interestingly, the problem is well known (proven, [1]) for having no fully fair solution and that corresponding apportionment algorithms for disaggregation will always generate paradoxical and/or biased situations, such as the famous Alabama Paradox: The apportionment strategy used by that time in the US for assigning parliament seats was based on ranking division remainders (Hamilton-method). However, an employee of the census office found out that the method would assign the state of Alabama 8 seats if 299 total seats were available, but only 7 if 300 total seats were available [15].

For GEPOC, we decided to apply the Huntington-Hill apportionment strategy [2]. The method is known to be considerably fair and is based on iterative disaggregation of a value by drawing indices:

Algorithm 4.2 (Huntington-Hill Disaggregation). *Let $X_2(y, r, s, a) \in \mathbb{N}$ be a value to be disaggregated and $\vec{p}(y, r, s, a) \in (\mathbb{R}^+)^n$ a vector which the disaggregated values should follow, then the Huntington-Hill method is given as follows:*

1. *Initialise two state vectors $\vec{v} \leftarrow \vec{p}(y, r, s, a)$ and $\vec{w} \leftarrow (0)_{i=1}^n$.*
2. *Perform the following step $X_2(y, r, s, a)$ -times:*
3. *Let $j \leftarrow \operatorname{argmax}(\vec{v})$, then $\vec{w}_j \leftarrow \vec{w}_j + 1$ and*

$$\vec{v}_j \leftarrow \frac{\vec{p}_j(y, r, s, a)}{\sqrt{\vec{w}_j(\vec{w}_j + 1)}}.$$

Hereby, highest $\vec{p}(y, r, s, a)$ and lowest j are used as tiebreakers if multiple highest values occur.

Set $\vec{X}_{2,disagg} \leftarrow \vec{w}$.

It must be noted that $\vec{p}(y, r, s, a)$ is not normed in this process. In general, $\vec{p}(y, r, s, a) \in (\mathbb{N}^0)^n$ can even be helpful for computation. With $N := \sum \vec{p}(y, r, s, a)$, then, for any $X_2(y, r, s, a) = kN, k \in \mathbb{N}$, the algorithm will return the values $k\vec{p}(y, r, s, a)$. That means, if $X_2(y, r, s, a) > N$, we may compute $k := \lfloor X_2(y, r, s, a)/N \rfloor$, apply the algorithm only on $X_2(y, r, s, a) - Nk$, and add $k\vec{p}$ at the end.

As an alternative to an apportionment algorithm, the strategy can also be interpreted as a drawing process from a probability distribution. Yet, in contrast to using random numbers for drawing, the resulting sequence

is deterministic and creates fully reproducible results without stochastic fluctuations. Quite the opposite is true: convergence properties to the distributions have been thoroughly investigated [2].

4.2 Two-Sided Disaggregation

Finally, we require algorithms to match two data-sets, both of which are coarse on different dimensions. To be specific, we will regard datasets X_1 and X_2 so that the corresponding index sets Ψ_1 and Ψ_2 fulfil the following three properties:

- there is a dimension j along which X_1 is fully aggregated, i.e. $|\{x_j, x \in \Psi_1\}| = 1$,
- there is a different dimension $j \neq k$ along which X_2 is fully aggregated, i.e. $|\{x_k, x \in \Psi_2\}| = 1$,
- for all other dimensions, both datasets have the same resolution.

Let, without loss of generality, be $j = 2$ and $k = 4$, then dataset X_1 differentiates age but is aggregated overall regions, and dataset X_2 differentiates regions but is fully aggregated over all age-classes. For every year y and sex s , the problem states as follows: find a data-set $X_{1,2}(y, r, s, a)$ so that

$$\sum_{r'} X_{1,2}(y, r', s, a) = X_1(y, s, a), \text{ and} \quad (24)$$

$$\sum_{a'} X_{1,2}(y, r, s, a') = X_2(y, r, s), \quad (25)$$

assuming that

$$\sum_{a'} X_1(y, s, a') = \sum_{r'} X_2(y, r', s) \quad (26)$$

holds for all y and s .

The main field of application of this problem, in terms of the parametrisation of GEPOC, is internal migration. Data which includes origin and destination information often does not include any other dimensions such as age. Internal emigration or immigration data, on the other hand, does not include any information about the individual's origin/destination. As a result, there are different data available, each showing one side of the medal, but not all. They can essentially be interpreted as the marginals of an unknown distribution.

4.2.1 Disaggregation of Marginal Distributions

For any values of the codimensions, problems (24) and (25) translate to a very fundamental mathematical problem. Find a positive matrix $M \in (\mathbb{R}^+)^{m \times n}$ so that the sum over all columns matches $\vec{a} \in (\mathbb{R}^+)^m$ and the sum over all rows matches $\vec{b} \in (\mathbb{R}^+)^n$, i.e.

$$\sum_{j=1}^n M_{i,j} = \vec{a}_i, \quad \sum_{i=1}^m M_{i,j} = \vec{b}_j.$$

For any $n \geq m > 2$, the problem is well underdetermined ($m \times n$ vs. $m + n$ degrees of freedom) and potentially has various solutions. One well-known method to find one of them is *Iterative Proportional Fitting* (IPF), sometimes also called RAS algorithm or biproportional fitting. It iteratively divides a given initial estimate by the target row- and column sums, leading to convergence.

Algorithm 4.3 (Iterative Proportional Fitting (IPF)). Let $\vec{a} \in (\mathbb{R}^+)^m$ and $\vec{b} \in (\mathbb{R}^+)^n$ be row- and column-sums of an unknown matrix $M \in (\mathbb{R}^+)^{m \times n}$, where $\sum_{i=1}^m \vec{a} = \sum_{j=1}^n \vec{b}$ is given. Furthermore, let $M_0 \in (\mathbb{R}/\{0\})^{m \times n}$ be an initial guess for M with positive values. Define $X := M_0$ as the state, and

$$\text{res}(X) := \left\| \sum_{i=1}^m M_{i,\cdot} - \vec{b} \right\| + \left\| \sum_{j=1}^n M_{\cdot,j} - \vec{a} \right\| \quad (27)$$

as the residual of the current state X . Until the residual is smaller than a defined tolerance or does not improve any more, perform the following steps:

1. Update:

$$\forall i, j : X_{i,j} \leftarrow X_{i,j} \frac{a_i}{\sum_{j=1}^n X_{i,j}}. \quad (28)$$

2. Update:

$$\forall i, j : X_{i,j} \leftarrow X_{i,j} \frac{b_j}{\sum_{i=1}^m X_{i,j}}. \quad (29)$$

The algorithm has been proven to converge as long as a solution exists and has been shown to have minimum distance (maximum likelihood) to the initial guess [9]. It is worth mentioning that the same concept can also be used in three dimensions, as it is required to fully entangle internal-emigration data and internal-immigration data, both with high age resolution, with origin-destination data without age information.

Algorithm 4.4 (Iterative Proportional Fitting 3D (IPF-3D)). Let $A \in (\mathbb{R}^+)^{m \times n}$, $B \in (\mathbb{R}^+)^{n \times r}$ and $C \in (\mathbb{R}^+)^{m \times r}$ be the marginals of an unknown tensor $M \in (\mathbb{R}^+)^{m \times n \times r}$, i.e.

$$\sum_{k=1}^r M_{\cdot,\cdot,k} = A_{\cdot,\cdot}, \quad \sum_{i=1}^m M_{i,\cdot,\cdot} = B_{\cdot,\cdot}, \quad \sum_{j=1}^n M_{\cdot,j,\cdot} = C_{\cdot,\cdot} \quad (30)$$

whereas we assume that

$$\sum_{i=1}^m A_{i,\cdot} = \sum_{j=1}^r B_{\cdot,j}, \quad \sum_{j=1}^n A_{\cdot,j} = \sum_{j=1}^r C_{\cdot,j}, \quad \sum_{i=1}^m B_{i,\cdot} = \sum_{i=1}^m C_{i,\cdot} \quad (31)$$

is given. Furthermore, let $M_0 \in (\mathbb{R}/\{0\})^{m \times n \times r}$ be an initial guess for M with positive values. Define $X := M_0$ as the state, and

$$\text{res}(X) := \left\| \sum_{k=1}^r M_{\cdot,\cdot,k} - A_{\cdot,\cdot} \right\| + \left\| \sum_{i=1}^m M_{i,\cdot,\cdot} - B_{\cdot,\cdot} \right\| + \left\| \sum_{j=1}^n M_{\cdot,j,\cdot} - C_{\cdot,\cdot} \right\| \quad (32)$$

as the residual of the current state X (hereby $\|\cdot\|$ refers to a suitable matrix-norm). Until the residual is smaller than a defined tolerance or does not improve any more, perform the following steps:

1. Update:

$$\forall i, j, k : X_{i,j,k} \leftarrow X_{i,j,k} \frac{A_{i,j}}{\sum_{k=1}^r X_{i,j,k}}. \quad (33)$$

2. *Update:*

$$\forall i, j, k : X_{i,j,k} \leftarrow X_{i,j,k} \frac{B_{j,k}}{\sum_{i=1}^m X_{i,j,k}}. \quad (34)$$

3. *Update:*

$$\forall i, j, k : X_{i,j,k} \leftarrow X_{i,j,k} \frac{C_{i,k}}{\sum_{j=1}^n X_{i,j,k}}. \quad (35)$$

4.2.2 Disaggregation of Integer-Valued Marginals (Concepts)

Clearly, the ideas from the last section can also be adapted to sole integer-valued disaggregation. For this purpose, we developed two algorithms ourselves, one for 2D and one for 3D. We will sketch the ideas:

In the first step of the 2D algorithm, marginal $\vec{a}$ is disaggregated to the second dimension using the Huntington-Hill Strategy, Algorithm 4.2, to get an initial estimate $M_0 =: X \in \mathbb{N}^{m \times n}$. We may use the IPF algorithm to get a good initial distribution for the Huntington-Hill Disaggregation algorithm. This way, we can be sure that the equations for the first marginal are fulfilled right from the start.

At this point, the actual algorithm starts, which can be motivated by the idea of connecting pins on a board by strings. Every unit-value in X corresponds to an individual (e.g. $X_{i,j} = 4$ refers to four individuals) and we imagine the individual to be a string on a wall of pins. The wall has one column of pins for every dimension and one pin for every possible index in this dimension. An individual with the indices i, j (e.g. origin i and destination j) connects the pin i in the first column and j in the second column.

If we take a string (i_0, j_1) and change its second pin to j_2 , then the first marginal equation would remain unchanged. The operation would correspond to $X_{i_0, j_1} - = 1$ and $X_{i_0, j_2} + = 1$. That means if j_1 was chosen so that $\sum_{i=1}^m X_{i, j_1} - \vec{b}_{j_1} > 0$ and j_2 so that $\sum_{i=1}^m X_{i, j_2} - \vec{b}_{j_2} < 0$, the algorithm reduces the marginal. By design, the algorithm will converge if a solution exists.

For the three dimensional variant, we use the same strategy, however, with three columns of pins on the wall (see Figure 1). In the first step, we use the 2D version of the algorithm to compute an initial guess $M_0 = X \in \mathbb{N}^{m \times n \times r}$ which fulfils the A and C marginal equations.

The main part of the algorithm is similar to the 2D variant: we look for two strings with same first index (i_0, j_1, k_1) and (i_0, j_2, k_2) . A swap of the second index $j_1 \leftrightarrow j_2$ will conserve all connections between the first and second index. As a result, the A marginal equation will be conserved. Also, the swap will conserve the connections between the first and third dimension, thereby leaving the C marginal untouched. The B -marginal, however, changes, since the number of (j_1, k_1) and (j_2, k_1) connections will be diminished by one, whereas (j_1, k_2) and (j_2, k_2) will grow by one. When suitable pairs are chosen, the algorithm makes progress in the right direction. However, convergence is not guaranteed (and usually not reached).

Currently, no integer-valued two-sided disaggregation is applied in the GEPOC parametrisation. So we do not give a reproducible statement of the two mentioned algorithms.

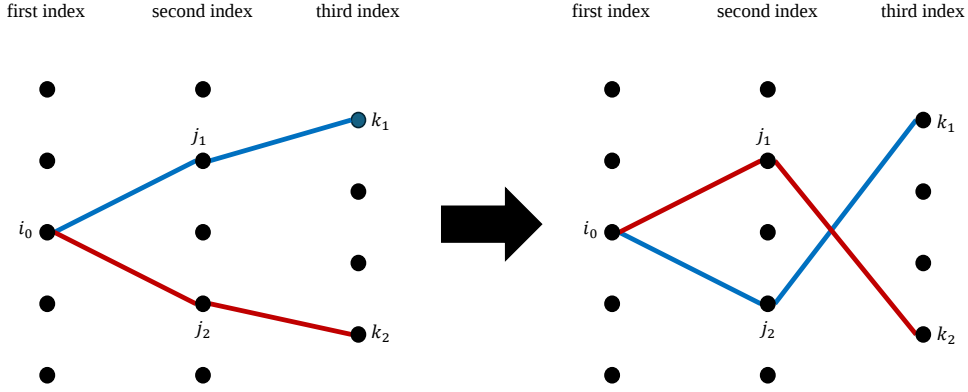


Figure 1: Concept of the Advanced Migration Matcher Algorithm via strings connecting pins on a wall.

5 Source Data

In this section we describe all source data used for parametrisation of GEPOC as of September 2025. Note that the parameter calculation might change as soon as data changes its format, is not updated anymore, or better data becomes freely available.

5.1 Sources for GEPOC ABM Parametrisation

The following tables show the source files for the parametrisation of GEPOC. Most of the files are updated yearly, therefore we will indicate the most recent year for which population information on Jan 1st is available as y_0 .

Source 5.1 (Bevölkerungsstand). *Bevölkerung zu Jahresbeginn (einheitlicher Gebietsstand y_0)*

Contents	
<i>Multiple files with population status for specific year (1.1.) for single age classes, sex and municipalities.</i>	
Resolution	
Time-Frame	<i>2002-y_0</i>
Regional-Level (Status)	<i>municipalities _ districts (y_0)</i>
Sex-Resolution	<i>{m, f}</i>
Age-Resolution	<i>[0, 100⁺]</i>
Source Information	
Filename	<i>OGD _ bevstandjbab2002 _</i>
Source (URL)	<i>https://data.statistik.gv.at</i>
Licence	<i>CC Namensnennung 4.0 International</i>

Source 5.2 (Bevölkerungsprognose 2014 bis 2024 - Zählbezirke (1) Wien). *VIE-Bevölkerungsprognose für 250 Wiener Zählbezirke nach Altersgruppen und Geschlecht 2014 bis 2024*

Contents	
<i>Population (and forecast) for Viennese registration-districts.</i>	
Resolution	
Time-Frame	2014-2024
Regional-Level (Status)	municipalities_registrationdistricts (2014)
Sex-Resolution	{m, f}
Age-Resolution	{0,3,6,10,15,20,25,30,45,60,75+}
Source Information	
Filename	vie303
Source (URL)	https://www.data.gv.at/
Licence	CC Namensnennung 4.0 International

Source 5.3 (Bevölkerung zum Jahresanfang 1952 bis 2101). *Prognose zur Bevölkerung zum Jahresanfang 1952 bis 2101 nach Alter in Einzeljahren, Geschlecht und Hauptszenario*

Contents	
<i>Population (forecast) for Austria for single age-classes, sex and federalstate</i>	
Resolution	
Time-Frame	1952-2101
Regional-Level (Status)	federalstates (1987)
Sex-Resolution	{m, f}
Age-Resolution	[0, 100+]
Source Information	
Filename	OGD_bevjahresanf_PR_BEVJA_4
Source (URL)	https://data.statistik.gv.at/
Licence	CC Namensnennung 4.0 International

Source 5.4 (Wanderungen mit dem Ausland ab 2002 nach Alter, Geschlecht und Staatsangehörigkeit). *Außenwanderungen ab 2002 nach Jahr, Alter in Einzeljahren, Geschlecht, Staatsangehörigkeit*

Contents	
<i>Immigrants and Emigrants since 2002 with respect to age and sex. No regional resolution!</i>	
Resolution	
Time-Frame	2002-($y_0 - 1$)
Regional-Level (Status)	country (1945)
Sex-Resolution	{m, f}
Age-Resolution	[0, 100+]
Source Information	
Filename	OGD_bevwan020_AUSSENWAND_100
Source (URL)	https://data.statistik.gv.at/...
Licence	CC Namensnennung 4.0 International

Source 5.5 (Wanderungen mit dem Ausland von 2002 bis 2014 nach Altersgruppen, Gemeinde und Staatsangehörigkeit). *Außenwanderungen 2002-2014 nach Jahr, 5jährige Altersgruppen, Gemeinde, Staatsangehörigkeit (Ländergruppen)*

Contents	
<i>Immigrants and Emigrants w.r. to 5year age groups and municipality. Additional info w.r. to origin countries. Gebietsstand 2022 not mentioned but used.</i>	
Resolution	
Time-Frame	2002-2014
Regional-Level (Status)	municipalities_districts (y_0)
Sex-Resolution	none
Age-Resolution	{0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100 ⁺ }
Source Information	
Filename	OGD_bevwan020_AUSSENWAND_201
Source (URL)	https://data.statistik.gv.at/...
Licence	CC Namensnennung 4.0 International

Source 5.6 (Wanderungen mit dem Ausland ab 2015 nach Altersgruppen, Gemeinde und Staatsangehörigkeit). *Außenwanderungen ab 2015 nach Jahr, 5jährige Altersgruppen, Gemeinde, Staatsangehörigkeit (Ländergruppen)*

Contents	
<i>Same as Source 5.5 but for years from 2015. No obvious reason found, why these two files are separated.</i>	
Resolution	
Time-Frame	2015-($y_0 - 1$)
Regional-Level (Status)	municipalities_districts (y_0)
Sex-Resolution	none
Age-Resolution	{0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100 ⁺ }
Source Information	
Filename	OGD_bevwan020_AUSSENWAND_202
Source (URL)	https://data.statistik.gv.at/...
Licence	CC Namensnennung 4.0 International

Source 5.7 (Bevölkerungsbewegung 1961 bis 2100 nach Bundesland, Bewegungsarten und Szenarien). *Bevölkerungsbewegung 1961 bis 2100, Bundesland, Bewegungsarten, Hauptszenario, Wachstum, Alterung, Fertilität, Wanderung*

Contents	
<i>Forecast of total number of died, newborn, internal and external migrants by federalstate.</i>	
Resolution	
Time-Frame	1961-2100 (births, deaths) 2002-2100 (migrants)
Regional-Level (Status)	federalstates (1987)
Sex-Resolution	none
Age-Resolution	none
Source Information	
Filename	OGD_bevbewegung_BEV_BEW_2
Source (URL)	https://data.statistik.gv.at/...
Licence	CC Namensnennung 4.0 International

Source 5.8 (Tabellensammlung Außenwanderung). *Wanderungen mit dem Ausland (Außenwanderungen) nach Bundesländern 1996- y_0*

Contents	
<i>Total number of emigrants and immigrants per federalstate.</i>	
Resolution	
Time-Frame	1996- $(y_0 - 1)$
Regional-Level (Status)	federalstates (1987)
Sex-Resolution	none
Age-Resolution	none
Source Information	
Filename	Tabellensammlung_Aussenwanderung_...
Source (URL)	https://data.statistik.at/statistiken/...
Licence	CC Namensnennung 4.0 International

Source 5.9 (Demographische Zeitreihenindikatoren). *Demographische Indikatoren sind international anerkannte und gebräuchliche Kennzahlen zur Beschreibung der Bevölkerungsstruktur und Bevölkerungsbe-
wegung...*

Contents	
<i>Various sources (one for each federalstate) with various indicators related to demography. This includes an age-specific death census, age-specific fertility, population with single age classes, and with reduced resolution internal and external migration counts per age-class.</i>	
Resolution	
Time-Frame	1961- y_0 (fertility, death) 2002- y_0 (migration)
Regional-Level (Status)	federalstates (1987)
Sex-Resolution	{m, f}
Age-Resolution	varies
Source Information	
Filename	xxx_ Zeitreihenindikatoren_ 1961_ y_0 .ods
Source (URL)	https://www.statistik.at/statistiken/...
Licence	CC Namensnennung 4.0 International

Source 5.10 (Demographische Indikatoren 1961 bis 2100). *Demografische Indikatoren nach Zeit, Bundesland und Szenarien*

Contents	
<i>Forecast for various fertility and death related parameters: Total fertility rate, gross reproduction rate, net reproduction rate, average fertility age, and life expectancy of men and women at birth and at the age of 65. The quality of the dataset is limited by the fact that Statistics Austria rounded every indicator to the nearest integer.</i>	
Resolution	
Time-Frame	1961-2100
Regional-Level (Status)	federalstates (1987)
Sex-Resolution	{m, f}
Age-Resolution	none
Source Information	
Filename	OGD_ demoind_ DEM_ IND_ 2
Source (URL)	https://data.statistik.gv.at/...
Licence	CC Namensnennung 4.0 International

5.2 Sources Specific for GEPOC ABM Geography Parametrisation

Source 5.11 (GeoJSON/TopoJSON Austria (2016-2017)). *Geo- and TopoJSON files of municipalities, districts and states in Austria, as of January 2017.*

Contents	
<i>GeoJSON file containing all borders of all municipalities in Austria including districts within Vienna as of region-status 2017. Since it was sufficient for GEPOC, a 95% simplified version was used.</i>	
Resolution	
Time-Frame	2017
Regional-Level (Status)	municipalities_districts (2017)
Sex-Resolution	-
Age-Resolution	-
Source Information	
Filename	gemeinden_95_geo.json
Source (URL)	https://github.com/ginseng666/GeoJSON-TopoJSON-Austria
Licence	CC BY 4.0, Flooh Perlot

Source 5.12 (Zählbezirksgrenzen Wien). *Zählbezirke sind statistische Definitionen. Die 23 Bezirke werden in 250 Zählbezirke unterteilt.*

Contents	
<i>GeoJSON file containing all borders of the Viennese registration-districts. Although internally consistent, it does not seamlessly fit with Source 5.11</i>	
Resolution	
Time-Frame	2019
Regional-Level (Status)	municipalities_registrationdistricts (2019)
Sex-Resolution	-
Age-Resolution	-
Source Information	
Filename	ZAEHLBEZIRKOGD
Source (URL)	https://www.data.gv.at/...
Licence	CC Namensnennung 4.0 International

Source 5.13 (GHS resident population grid). -

Contents	
<i>Raster image of Europe. Value per raster estimates the number of inhabitants.</i>	
Resolution	
Time-Frame	2016
Regional-Level (Status)	100[m] × 100[m] raster map
Sex-Resolution	-
Age-Resolution	-
Source Information	
Filename	Austria2016_100m.json
Source (URL)	https://www.eea.europa.eu/...
Licence	Creative Commons Attribution 4.0 International

5.3 Sources Specific for GEPOC ABM IM Parametrisation

Source 5.14 (Wanderungen innerhalb Österreichs). *Wanderungen innerhalb Österreichs ab 2002 (einheitlicher Gebietsstand y_0)*

Contents	
<i>Number of internal migrants between (origin-destination) municipalities/districts in Austria w.r. to migration year and sex.</i>	
Resolution	
Time-Frame	2002- $(y_0 - 1)$
Regional-Level (Status)	municipalities_ districts (y_0)
Sex-Resolution	{m, f}
Age-Resolution	none
Source Information	
Filename	OGDEXT_BINNENWAND_1
Source (URL)	https://data.statistik.gv.at/...
Licence	CC Namensnennung 4.0 International

Source 5.15 (Binnenwanderungen innerhalb Österreichs ab 2002). *Jahr und Alter in Einzeljahren nach Wanderungen innerhalb Österreichs und Politischer Bezirk / Wiener Gemeindebezirk - Herkunftsort nach Geschlecht*

Contents	
<i>Total number of internal emigrants and immigrants per district collected manually from STATcube. Vienna and remaining Austria are collected in different files due to download limitations.</i>	
Resolution	
Time-Frame	$2002-(y_0 - 1)$
Regional-Level (Status)	$districts_districts (y_0)$
Sex-Resolution	$\{m, f\}$
Age-Resolution	$[0, 95^+]$
Source Information	
Filename	-
Source (URL)	https://statcube.at/...
Licence	CC Namensnennung 4.0 International as long as the data is downloaded manually

Source 5.16 (Tabellensammlung Binnenwanderungen 2024). *Wanderungen innerhalb Österreichs (Binnenwanderungen) zwischen und innerhalb der Bundesländer 1996–2024*

Contents	
<i>Total number of internal emigrants and immigrants per federalstate without age-resolution.</i>	
Resolution	
Time-Frame	1996-2024
Regional-Level (Status)	$federalstates (1987)$
Sex-Resolution	none
Age-Resolution	none
Source Information	
Filename	Tabellensammlung_Binnenwanderungen_2024
Source (URL)	https://www.statistik.at/statistiken/...
Licence	CC Namensnennung 4.0 International

6 GEPOC Parameter Calculation

In the following we describe how we calculate the GEPOC parameters as introduced in Section 2 from the source data introduced in Section 5 in a reproducible way. We will refer to the year for which the last population-census data is available as y_0 . This will also correspond to the used area-status and we assume that the updated data for population change (migration, births, deaths, etc.) is available for this area-status and up to year $(y_0 - 1)$.

In general, calculation of parameters involves two steps. In the first and most complicated step, we compute harmonized census data (population, births, deaths, ...) for as long a period as possible. Our goal is to have a continuous time-series of census numbers ranging from the past into the far future with the same spatial and age-resolution. Browsing over the multiple stated source-files from Section 5, it becomes clear that the age-, sex-, and regional resolutions of the files are inconsistent. In particular, forecasts are usually given on coarser regional and age resolution than historic data. This is visualised in Figure 2. As mentioned in Section 4, we will harmonize all data on the finest level of resolution using one of the presented disaggregation algorithms.

In the second step we will use the harmonized data series to compute probabilities using the adapted Farr formula 3.1.

We will henceforth use the following notation for a disaggregation procedure with either of the two one-sided algorithms (Proportional Disaggregation method, Algorithm 4.1, and Huntington-Hill Disaggregation method, Algorithm 4.2). The expression

$$X_1 : (Y_1, R_1, S_1, A_1) \rightarrow (Y_2, R_2, S_2, A_2) \text{ via } X_2 \text{ key } (K_1, K_2, \dots). \quad (36)$$

means that dataset X_1 is disaggregated from year/region/sex/age resolution (Y_1, R_1, S_1, A_1) to the finer resolution (Y_2, R_2, S_2, A_2) using a distribution from dataset X_2 . Hereby, $(K_1, K_2, \dots)$ must be a sub-vector of (Y_1, R_1, S_1, A_1) and refers to all dimensions along which an individual distribution is used for disaggregation. To gain a better understanding, we refer to the first application of the notation below, since the concept is best explained on an example.

6.1 Population $\hat{P}$

Since providing a static micro-census of the population is one of the key objectives of GEPOC, our goal is to get a parameter set on the finest-possible age and regional resolution. Therefore, several data sources have to be disaggregated and joined. Note that in this processing strategy we will apply the disaggregation method of Huntington-Hill, Algorithm 4.2, since we require whole numbers for initialisation of the agent population.

The following files pose the basis for the population processing.

name of source	Bevölkerungsstand	Bevölkerung zum Jahresanfang 1952 bis 2101	Demographische Zeitreihenindikatoren	Bevölkerungsprognose 2014 bis 2024 Zählbezirke (1) Wien
source	Source 5.1	Source 5.3	Source 5.9 (Table 19)	Source 5.2
variable	$S_p(y, r, s, a)$	$S_{pf}(y, r, s, a)$	$S_i^p(y, r, s, a)$	$S_{pv}(y, r, s, a)$
$y \in$	[2002, y_0]	[1952, 2101]	[1962, 2021]	[2014, 2024]
$r \in$	municipalities _ districts	federalstates	federalstates	registrationdistricts (only Vienna)
$s \in$	$\{m, f\}$	$\{m, f\}$	$\{m, f\}$	$\{m, f\}$
$a \in$	[0, 100 ⁺]	[0, 100 ⁺]	[0, 95 ⁺]	{0, 3, 6, ..., 75 ⁺ }

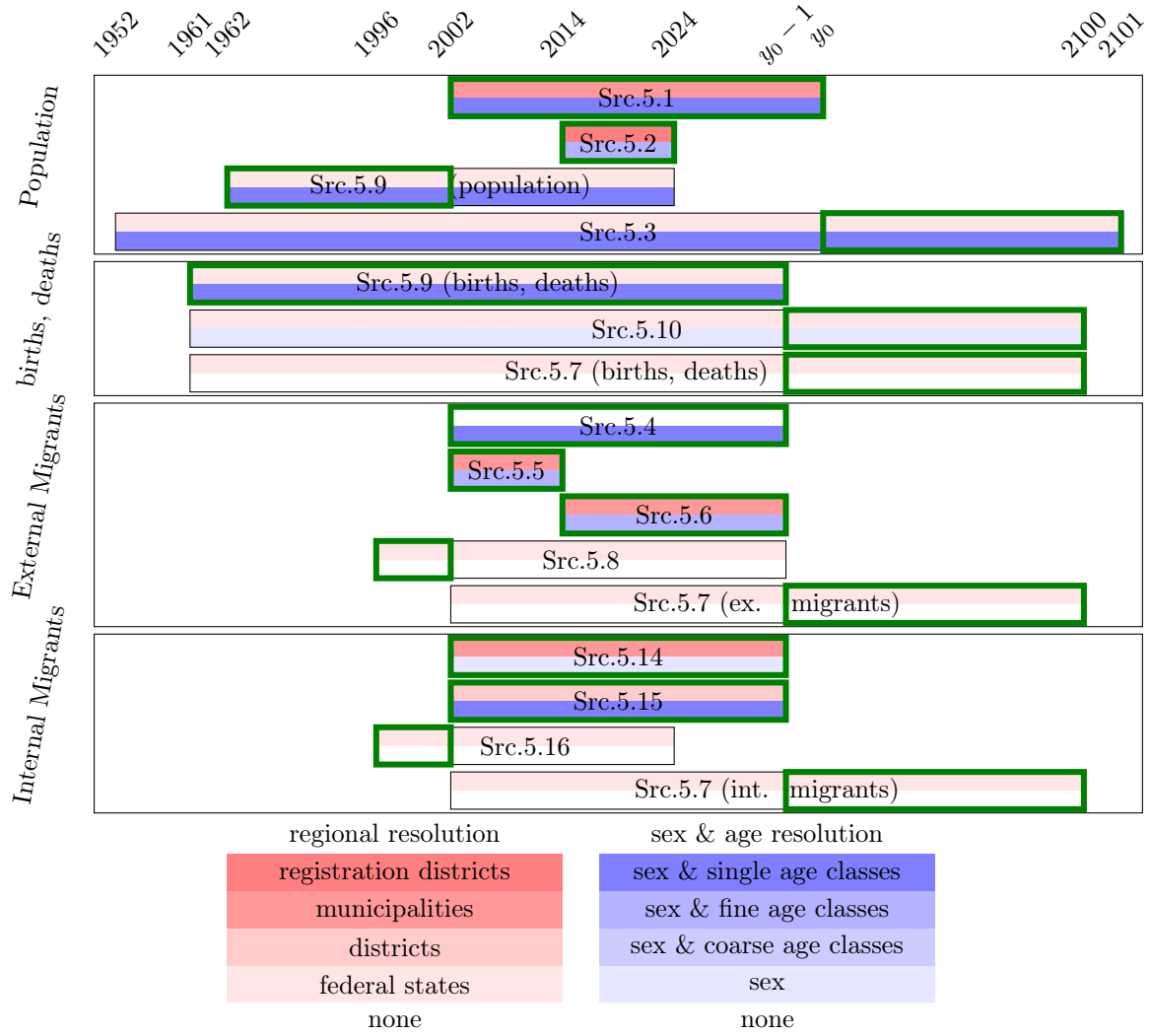


Figure 2: Overview of the time periods covered by the different data-sources. The darker the colour the higher the resolution of the corresponding data, red for spatial resolution, blue for age & sex resolution. Several data-sources are only used for a part of the time-frame which they cover, indicated by a green rectangle.

Note that we will use Source 5.9 between 1962 and 2001 to generate a population census earlier than 2002 and not Source 5.3, even though the latter would provide data until 1952 (see also, Figure 2). The reason for this is that we observed that Source 5.9 perfectly aligns with the high-resolution population data from Source 5.1 for the overlapping years 2002 to 2021 whereas the other does not. Apparently, Source 5.3 should be interpreted as a forecast over its whole time period and incorporates corresponding uncertainty.

1. We start processing with the population forecast (pf) $S_{pf}(y, r, s, a)$ (Source 5.3) restricted to $y \in \{y_0 + 1, \dots, 2101\}$ (earlier data will be ignored). To be compatible with S_p (Source 5.1), we need to disaggregate it w.r. to a municipality distribution.
2. We aggregate the last three years of the historic population (p) census S_p :

$$X_1(r, s, a) = \sum_{y=y_0-2}^{y_0} S_p(y, r, s, a).$$

3. We will use X_1 for disaggregation of S_{pf} using the Huntington-Hill Disaggregation method:

$$\begin{aligned} S_{pf} : ([y_0 + 1, 2101], \{\text{federalstates}\}, \{m, f\}, [0, 100^+]) \\ \rightarrow ([y_0 + 1, 2101], \{\text{municipalities_districts}\}, \{m, f\}, [0, 100^+]) \\ \text{via } X_1 \text{ key } (\{\text{federalstates}\}, \{m, f\}, [0, 100^+]) \end{aligned}$$

Since this is the first time we use the disaggregation notation introduced earlier, we describe the process in detail:

- (a) Disaggregation will be applied along the regional dimensions. So, for every federal-state id $r \in \{AT-1, \dots, AT-9\}$, we compute the list $\vec{z}(r)$ of all municipality ids which lie within r . E.g.

$$\vec{z}(AT-1) = (10101, 10201, \dots, 10932)$$

- (b) According to the key-tuple $(K_1, K_2, \dots) = (\{\text{federalstates}\}, \{m, f\}, [0, 100^+])$, we iterate over all federal-state ids r , sex $s \in \{m, f\}$, and age classes $a \in [0, 100^+]$, and define

$$\vec{P}(r, s, a) = (X_1(x, s, a))_{x \in \vec{z}(r)}$$

- (c) We initialise a data-set $X_2(\cdot, \cdot, \cdot, \cdot)$ with zero entries and same spatial resolution as X_1 and sex and age resolution as S_{pf} .
- (d) For every year $y \in [y_0 + 1, 2101]$, we iterate over all federal-states r , sex $s \in \{m, f\}$, and age classes $a \in [0, 100^+]$. We use $\vec{P}(r, s, a)$ as the distribution vector of Algorithm 4.2, disaggregate $S_{pf}(y, r, s, a)$ into a vector $\vec{q}$, and integrate it into X_2 via $X_2(y, \vec{z}(r)_i, s, a) := \vec{q}_i$.

The final dataset X_2 is a refined forecast until 2101:

$$X_2(y, r, s, a) : y \in [y_0 + 1, 2101], r \in \text{municipalities_districts}, s \in \{m, f\}, a \in [0, 100^+]$$

It has the following properties:

- The municipalities_districts-distribution per federal-state matches, within the numerical accuracy of the algorithm, the distribution of X_1 from $(y_0 - 2)$ to y_0 .
- It is integer-valued, which is reasonable for a population census and important for agent-based model parametrisation.

- Aggregation of X_2 from municipalities_districts to federalstates would perfectly result in S_{pf} .
4. In the next step, we extend the population data-set beyond the lower time-frame bound (2002) from Source 5.1. using S_i^p between 1962 and 2001.
 5. Since S_i^p has $a_{max} = 95$, whereas $a_{max} = 100$ is needed to match with S_p and X_2 , we first need to disaggregate the final age-cohort.
 6. For all federal-state ids r , sex $s \in \{m, f\}$ and age $a \geq 95$ define $X_3(r, s, a) := \sum_{r' \in r} \sum_{y=2002}^{2004} S_p(y, r', s, a)$. This way we receive an age distribution for persons ≥ 95 . We use it to disaggregate the 95^+ cohort from S_i^p as follows:

$$\begin{aligned}
S_i^p : ([1962, 2001], \{\text{federalstates}\}, \{m, f\}, 95^+) \\
\rightarrow ([1962, 2001], \{\text{federalstates}\}, \{m, f\}, [95, 100^+]) \\
\text{via } X_3 \text{ key } (\{\text{federalstates}\}, \{m, f\})
\end{aligned}$$

The resulting dataset X_4 has now the required $a_{max} = 100$:

$$X_4(y, r, s, a) : y \in [1962, 2001], r \in \{\text{federalstates}\}, s \in \{m, f\}, a \in [0, 100^+]$$

7. Furthermore we extrapolate from federalstates to municipalities_districts. Analogous to steps 2 and 3, we define

$$X_5(r, s, a) = \sum_{y=2002}^{2004} S_p(y, r, s, a)$$

and distribute

$$\begin{aligned}
X_4 : ([1962, 2001], \{\text{federalstates}\}, \{m, f\}, [0, 100^+]) \\
\rightarrow ([1962, 2001], \{\text{municipalities_districts}\}, \{m, f\}, [0, 100^+]) \\
\text{via } X_5 \text{ key } (\{\text{federalstates}\}, \{m, f\}, [0, 100^+])
\end{aligned}$$

using the Huntington-Hill Disaggregation method. The resulting data-set X_6 has the same resolutions as S_p and X_2 :

$$X_6(y, r, s, a) : y \in [1962, 2001], r \in \{\text{municipalities_districts}\}, s \in \{m, f\}, a \in [0, 100^+]$$

Since their time-frames $[1962, 2001], [2002, y_0]$ and $[y_0 + 1, 2101]$ fit seamlessly, we merge them into a combined dataset X_7 .

$$X_7(y, r, s, a) : y \in [1962, 2101], r \in \{\text{municipalities_districts}\}, s \in \{m, f\}, a \in [0, 100^+]$$

8. It remains to disaggregate Viennese districts into registration-districts using the corresponding population census for Vienna S_{pv} (Source 5.2). This is challenging due to (a) the unusual age-classes and (b) the limited time-frame of S_{pv} .

We define the auxiliary data-set

$$X_8(y, r, s, a) = S_{pv}(\min(\max(y, 2014), 2023), r, s, a).$$

to extend the time-frame of S_{pv} . We use it for the disaggregation process of X_7 for the Viennese regions using the Huntington-Hill Disaggregation method:

$$\begin{aligned} X_7 : ([1962, 2101], \{\text{districts (Vienna)}\}, \{m, f\}, [0, 100^+]) \\ \rightarrow ([1962, 2101], \{\text{registrationdistricts (Vienna)}\}, \{m, f\}, [0, 100^+]) \\ \text{via } X_8 \text{ key } ([1962, 2101], \{\text{districts (Vienna)}\}, \{m, f\}, \{0, 3, 6, \dots, 75^+\}) \end{aligned}$$

The resulting data-set $\hat{P}$ finally has the target resolution

$$\hat{P}(y, r, s, a) : y \in [1962, 2101], r \in \{\text{municipalities_registrationdistricts}\}, s \in \{m, f\}, a \in [0, 100^+]$$

and is the final result of the population data processing.

9. To be more flexible, we aggregated the data to coarser regional-levels as well.

Parameter Value 6.1 (Population $\hat{P}$).	
<i>Contents</i>	
<i>Population of Austria per age, region, sex at the start of the years between 1962 and 2101</i>	
<i>Resolution</i>	
<i>Time-Frame</i>	[1962, 2101]
<i>Regional-Level</i>	<i>municipalities_registrationdistricts</i>
<i>Age-Resolution</i>	[0, 100 ⁺]
<i>Other Regional-Levels</i>	<i>municipalities_districts / municipalities / districts_districts / districts / federalstates / country</i>

In comparison with the ground-truth provided by Statistics-Austria, we would score the validity of the parameter values (0 invalid, 10 fully valid) as follows:

years	regions	sex	age	score	description
[2002, y_0]	actual municipalities	all	all	10	raw data
[2002, y_0]	Viennese registrationdistricts	all	all	9.5	statistical linkage with population data from Vienna
$[y_0 + 1, 2101]$	all	all	[0, 100 ⁺]	8	federalstates extrapolated to municipalities and registrationdistricts using the distribution in years 2022-2024.
[1990, 2001]	all	all	[0, 94]	8	federalstates extrapolated to municipalities and registrationdistricts using the distribution in years 2002-2004.
[1990, 2001]	all	all	[95, 100 ⁺]	7	age cohorts extrapolated using the distribution in years 2002-2004.

6.2 Births $\hat{B}_m, \hat{B}$ and Probability of a Male Child $\hat{\alpha}_m$

In this section we compute a harmonized census for births (by sex/age of children and by sex/age of corresponding mothers).

6.2.1 Births by Age of the Mother $\hat{B}_m$

We start by computing births by the age of the mother. Like the population data, we need to disaggregate and join a series of different data with different resolutions. The situation becomes even more complicated, since forecast information for the age distribution of births is only available in form of fertility rates and mean fertility age (see Source 5.9). Thus, additional assumptions need to be made.

Moreover, we want to emphasise that computed births are no longer required to be integers, since we will only use them in form of probabilities in the model (see later). Therefore, we will use the Proportional Disaggregation algorithm, Algorithm 4.1, because it perfectly conserves the distribution.

We will use the following sources for computing births:

name of source	Demographische Zeitreihen -indikatoren	Bevölkerungs- bewegung 1961 bis 2100 nach Bundesland, Bewegungsarten und Szenarien	Demographische Indikatoren 1961 bis 2100	Population $\hat{P}$
source	S. 5.9 (Table 1)	S. 5.7 (births)	S. 5.10 (fertility age)	Parameter Value 6.1
variable	$S_i^b(y, r, s, a)$	$S_{mf}^b(y, r)$	$S_{if}^b(y, r)$	$\hat{P}(y, r, s, a)$
$y \in$	$[1961, y_0 - 1]$	$[1961, 2100]$	$[1961, 2100]$	$[1962, 2101]$
$r \in$	federalstates	federalstates	federalstates	federalstates
$s \in$	$\{m, f\}$	none	-	$\{m, f\}$
$a \in$	$[15, 49]$	none	-	$[0, 100^+]$

The population-parameters $\hat{P}$ derived in the previous section are required to compute birth rates from births and vice versa.

1. First of all, we use the age specific births from the time-series indicators source (Table 1 “Lebendgeborene” from Source 5.9) by age of the corresponding mother. We refer to this data as $S_i^b(y, r, s, a)$.
2. We will use this data in combination with the female population $\hat{P}(y, r, f, a)$ to compute the age-dependent average fertility rate (see Definition 3.5) by using the classic approximation for the average population. The data suggests that the rate for women outside of the given age-range $[15, 49]$ can be set to zero. So we define:

$$X_1(y, r, f, a) = \begin{cases} \frac{S_i^b(y, r, f, a)}{0.5(\hat{P}(y, r, f, a) + \hat{P}(y+1, r, f, a))}, & a \in [15, 49] \\ 0, & a \in [0, 14] \vee a \in [50, 100^+]. \end{cases}$$

At this point, X_1 has the following resolution:

$$X_1(y, r, s, a) : y \in [1962, y_0 - 1], r \in \{\text{federalstates}\}, s \in \{f\}, a \in [0, 100^+].$$

Note that we will drop data for year 1961, since no population data is available.

3. For computing the forecast, we make use of the migration forecast $S_{mf}^b(y, r)$ (Source 5.7), which also includes births and deaths, and the fertility related fields from the indicator forecast $S_{if}^b(y, r)$ (Source 5.10). The prior contains the *total number of births* for a given year and federalstate without further specifying age and sex, neither of the child nor the mother. The latter contains an estimate about the dynamics of the *mean-fertility age* (MAC, see Definition 3.6) for every federalstate. That means, for each year and federalstate, the forecast of 101 birth-probabilities (one for each age-class) will be based on two scalar variables.

4. For each year $y \in [y_0, 2100]$ and federalstate $r \in \{\text{AT-1}, \dots, \text{AT-9}\}$ we define the following optimisation problem:

Let $\vec{b} \in (\mathbb{R}^+)^{101}$ stand for the average age-dependent birth-rate vector for year y and federalstate r then

$$f_1 : (\mathbb{R}^+)^{101} \rightarrow \mathbb{R}^+ : \vec{b} \mapsto \sum_{i=0}^{100} \vec{b}_i \frac{\hat{P}(y, r, f, i) + \hat{P}(y+1, r, f, i)}{2}$$

maps the rates to the *total number of births* and

$$f_2 : (\mathbb{R}^+)^{101} \rightarrow \mathbb{R}^+ : \vec{b} \mapsto \frac{\sum_{i=0}^{100} i \vec{b}_{i+1}}{\sum_{i=0}^{100} \vec{b}_{i+1}}$$

maps them to the MAC. With these functions, we aim to minimise

$$F : (\mathbb{R}^+)^{101} \rightarrow \mathbb{R}^+ : \vec{b} \mapsto \frac{|f_1(\vec{b}) - S_{mf}^b(y, r)|}{100} + |f_2(\vec{b}) - S_{if}^b(y, r)|. \quad (37)$$

The scaling factor 1/100 balances that the two objectives are on different scales (MAC 30, births > 10000). With 101 degrees of freedom, this optimisation problem would be heavily underdetermined. Consequently, we reduce the degrees of freedom by

$$G : \mathbb{R}^3 \rightarrow (\mathbb{R}^+)^{101} : \vec{\theta} \mapsto \vec{b} = \left(\vec{\theta}_1 \exp \left(-\frac{((i-1) - \vec{\theta}_2)^2}{\vec{\theta}_3^2} \right) \right)_{i=0}^{100}. \quad (38)$$

Using the Gaussian bell curve fit we achieve the reduced minimisation problem $F \circ G$ which has three-dimensional input. As seen in Figure 3, the Gaussian bell curve G is a proper model for the age-distribution of the birth rates later than 1990, after which the distribution stopped being skewed to the left.

5. After defining the minimisation problem $F \circ G$ we solve it numerically for every federalstate and year $y \in [y_0, 2100]$ using a Broyden-Fletcher-Goldfarb-Shanno Quasi Newton method. For the optimised $\vec{\theta}_{opt}$, we compute $\vec{b}_{opt} = G(\vec{\theta}_{opt})$ and use it to get the entries for the age-dependent birth forecast X_2 :

$$X_2(y, r, f, i) = \vec{b}_{opt,i} \frac{\hat{P}(y, r, f, i) + \hat{P}(y+1, r, f, i)}{2}.$$

6. We combine X_1 and X_2 into $\hat{B}_m$ which has now the following resolution:

$$\hat{B}_m(y, r, s, a) : y \in [1962, 2100], r \in \{\text{federalstates}\}, s \in \{f\}, a \in [0, 100^+].$$

7. To be more flexible we computed the data for other regional-levels as well. Hereby the data was not only aggregated to coarser levels but also disaggregated to a finer one. Hereby the fertility rates were disaggregated, assuming that the rates are equivalent for each subregion. Recomputing the births using the fine-grained population data gives an estimate for high-resolution birth data.

Parameter Value 6.2 (Births per Mother $\hat{B}_m$).

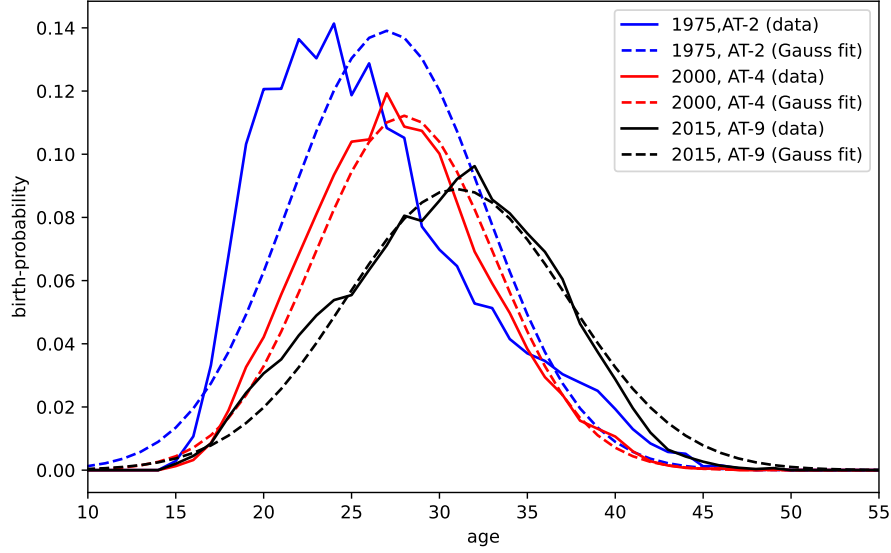


Figure 3: Fit of the Gaussian bell curve through the relative births per female inhabitant ($S_i^b/1000$)

<i>Contents</i>	
<i>Total number of newborn children of female persons with age a living in region r in the course of year y.</i>	
<i>Resolution</i>	
<i>Time-Frame</i>	[1962, 2100]
<i>Regional-Level</i>	<i>federalstates</i>
<i>Age-Resolution</i>	[0, 100 ⁺]
<i>Other Regional-Levels</i>	<i>districts_districts / country</i>

In comparison with the ground-truth provided by Statistics-Austria, we would score the validity of the parameter values (0 invalid, 10 fully valid) as follows:

years	regions	sex	age	score	description
[1962, $y_0 - 1$]	all	all	all	10	raw data
[y_0 , 2100]	all	all	all	6.5	extrapolation method with a Gaussian bell curve to conserve certain sums (total births, average fertility age ...)

6.2.2 Birth Probability of a Male Child $\hat{\alpha}_m$

Next, we will compute the probability that a newborn person in Austria is male or female. This is done using Source 5.9 for Austria as a whole. Table 1 “Lebendgeborene” includes the total number of male and female newborn since 1961 which can be used to compute the ratio. In the regarded time-frame, the ratio of male newborn varies between 0.5082 in 1982 and 0.5170 in 2012, yet, in general, no trend can be observed. The average lies at 0.513234 which we will furthermore use as a static parameter value.

Parameter Value 6.3 (Probability of a male newborn $\hat{\alpha}_m$).
<i>Contents</i>
<i>Probability that a newborn child is male:</i>
$P(s = m s \in \{f, m\}) = \hat{\alpha}_m = 0.513234$

6.2.3 Births $\hat{B}$

Finally, combining probability $\hat{\alpha}_m$ with the total births per mother's age $\hat{B}_m(y, r, f, a)$ via

$$\hat{B}(y, r, s, 0) = \begin{cases} \hat{\alpha}_m \sum_{a=0}^{100} \hat{B}_m(y, r, f, a), & s = m, \\ (1 - \hat{\alpha}_m) \sum_{a=0}^{100} \hat{B}_m(y, r, f, a), & s = f, \end{cases} \quad (39)$$

we get a proper estimate for the total number of newborns per sex of the newborn $\hat{B}$.

Parameter Value 6.4 (Births $\hat{B}$).	
<i>Contents</i>	
<i>Total number of newborn children with sex s born in region r in the course of year y. Regional-levels can be <code>districts</code>, <code>districts</code>, <code>country</code> or <code>federalstates</code></i>	
<i>Resolution</i>	
<i>Time-Frame</i>	[1962, 2100]
<i>Regional-Level</i>	<code>federalstates</code>
<i>Age-Resolution</i>	{0}
<i>Other Regional-Levels</i>	<code>districts</code> , <code>districts</code> / <code>country</code>

6.3 Deaths $\hat{D}$

In this section we compute a harmonized census for deaths. Like the processing steps for the births, additional assumptions need to be made due to the forecast information.

Again, we will use the Proportional Disaggregation method, Algorithm 4.1 because it perfectly conserves the distribution.

We will use the following sources for computing deaths:

name of source	Demographische Zeitreihen -indikatoren	Bevölkerungs- bewegung 1961 bis 2100 nach Bundesland, Bewegungsarten und Szenarien	Demographische Indikatoren 1961 bis 2100	Population $\hat{P}$
source	S. 5.9 (Table 3)	S. 5.7 (deaths)	S. 5.10 (life expectancies)	Parameter Value 6.1
variable	$S_i^d(y, r, s, a)$	$S_{mf}^d(y, r)$	$S_{if}^d(y, s, a)$	$\hat{P}(y, r, s, a)$
$y \in$	$[1961, y_0 - 1]$	$[1961, 2100]$	$[1961, 2100]$	$[1962, 2101]$
$r \in$	federalstates	federalstates	country	federalstates
$s \in$	$\{m, f\}$	none	$\{m, f\}$	$\{m, f\}$
$a \in$	$[0, a_{max}^+]$, a_{max} varies	none	$\{0, 65\}$	$[0, 100^+]$

Due to the lack of age-information for the forecast, calculation of deaths works quite similar to the one for births: we will specify a minimisation problem.

1. We will start with the death table S_i^d , which contains the absolute number of deaths per federalstate.
2. In the next step we will take care about deficiencies in this data set due to the varying a_{max} class. As with the births, we will drop year 1961 due to missing population data.
3. Since $a_{max} = 95$ for years before 1971, we use the $[95, 96, 97, 98, 99^+]$ data for 1971-1973 to disaggregate the 95^+ cohort with the Proportional Disaggregation method, Algorithm 4.1. Let $X_1(r, s, a) = \sum_{y=1971}^{1973} S_i^d(y, r, s, a)$ and compute

$$\begin{aligned}
S_i^d : ([1962, 1970], \{\text{federalstates}\}, \{m, f\}, 95^+) \\
\rightarrow ([1962, 1970], \{\text{federalstates}\}, \{m, f\}, \{95, 96, 97, 98, 99^+\}) \\
\text{via } X_1 \text{ key } (\{\text{federalstates}\}, \{m, f\})
\end{aligned}$$

4. Since $a_{max} = 99$ for years before 2017, we use the $[99, 100^+]$ data for 2017-2019 to disaggregate the 99^+ cohort. Let $X_2(r, s, a) = \sum_{y=2017}^{2019} S_i^d(y, r, s, a)$ and compute

$$\begin{aligned}
S_i^d : ([1962, 2016], \{\text{federalstates}\}, \{m, f\}, 99^+) \\
\rightarrow ([1962, 2016], \{\text{federalstates}\}, \{m, f\}, \{99, 100^+\}) \\
\text{via } X_2 \text{ key } (\{\text{federalstates}\}, \{m, f\})
\end{aligned}$$

At this stage we have the total number of deaths by age, region, and sex via

$$X_2(y, r, s, a) : y \in [1962, y_0], r \in \{\text{federalstates}\}, s \in \{m, f\}, a \in [0, 100^+]$$

5. In the next steps, we will compute death probabilities using the classic Farr Formula from Theorem 3.1. Our goal is to emulate Statistic Austria's death tables to create a forecast which is compatible with the provided indicators

$$X_3(y, r, s, a) = \frac{X_2(y, r, s, a)}{\frac{1}{2} \left(\hat{P}(y, r, s, a) + \hat{P}(y+1, r, s, a) \right) + \alpha(a) X_2(y, r, s, a)}$$

We identified that we can reproduce the death tables of Statistic Austria using $\alpha(0) = 0.923$ and $\alpha(a) = 0.5$ else.

6. Analogous to the birth computation (see Section 6.2) we furthermore define a minimisation problem. First we need to specify the reference:

For any given year $y \in [y_0, 2100]$ five values are available: The total number of deaths $d = \sum_r S_{mf}^d(y, r)$, the life-expectancies of 0-year old males $c_{0,m} = S_{10}(y, m, 0)$ and females $c_{0,f} = S_{10}(y, f, 0)$, and the life expectancies of 65-year old males $c_{65,m} = S_{10}(y, m, 65)$ and females $c_{65,f} = S_{10}(y, f, 65)$. These five values will provide the target of the minimisation problem.

The two death-probability vectors for males and females $\vec{q}_m$ and $\vec{q}_f$ pose the degrees of freedom. Furthermore we define the five functions $f_1(\vec{q}_m, \vec{q}_f)$, $f_2(\vec{q}_m)$, $f_3(\vec{q}_f)$, $f_4(\vec{q}_m)$ and $f_5(\vec{q}_f)$ which map the death probabilities onto the five desired outcomes: total deaths, and life-expectancy of 0/65-year old males and females.

We furthermore define the target function for the optimisation:

$$F : (\mathbb{R}^+)^{2 \cdot 101} \rightarrow \mathbb{R}^+ :$$

$$(\vec{q}_m, \vec{q}_f) \mapsto \left\| \frac{f_1(\vec{q}_m, \vec{q}_f) - d}{2000} \right\|_2 + \|f_2(\vec{q}_m) - c_{0,m}\|_2 + \|f_3(\vec{q}_f) - c_{0,f}\|_2 + \|f_4(\vec{q}_m) - c_{65,m}\|_2 + \|f_5(\vec{q}_f) - c_{65,f}\|_2$$

The weight of $1/2000$ was determined experimentally to give useful results, since f_1 is on a much higher scale than the other four outcomes.

7. Before defining the five functions $f_1 - f_5$, we reduce the degrees of freedom of the problem (currently $2 \cdot 101 = 202$). Our approach to solve this problem is a parametrised reference distribution. We define two reference distributions $q_{ref,m}$, $q_{ref,f}$ from the last three available data years via

$$q_{ref,m}(a) = \frac{1}{3}(X_3(y_0 - 2, m, a) + X_3(y_0 - 1, m, a) + X_3(y_0, m, a)).$$

$$q_{ref,f}(a) = \frac{1}{3}(X_3(y_0 - 2, f, a) + X_3(y_0 - 1, f, a) + X_3(y_0, f, a)).$$

Note that for the most recent update, the years 2020, 2021 and 2022 were omitted in this computation due to slightly different mortality by the COVID-19 crisis.

Furthermore, we parametrise the reference distributions with six parameters, three for each sex. We define

$$G : \mathbb{R}^6 \rightarrow (\mathbb{R}^+)^{2 \cdot 101} : (\vec{\theta}) \mapsto \left(\begin{aligned} &(\theta_1 \phi_1(a) + \theta_2 \phi_2(a) + \theta_3 \phi_3(a)) \cdot q_{ref,m}(a), & a \in [0, 100] \\ &(\theta_4 \phi_1(a) + \theta_5 \phi_2(a) + \theta_6 \phi_3(a)) \cdot q_{ref,f}(a), & a \in [0, 100] \end{aligned} \right). \quad (40)$$

The three actuation functions ϕ_1, ϕ_2, ϕ_3 are chosen to smoothly raise and lower death probabilities in certain age groups. We chose

$$\phi_1(a) = 1 - \frac{1}{1 + \exp(-0.5 \cdot (a - 16))}, \quad (41)$$

$$\phi_2(a) = \frac{1}{1 + \exp(-0.5 \cdot (a - 16))} - \frac{1}{1 + \exp(-0.5 \cdot (a - 65))}, \quad (42)$$

$$\phi_3(a) = \frac{1}{1 + \exp(-0.5 \cdot (a - 65))}. \quad (43)$$

Figure 4 displays the three functions. However, with $F \circ G$ the optimisation problem becomes 6 dimensional. Figure 5 shows, how well the strategy fits historic data, analogous to Figure 3 for the birth parameter processing.

8. We furthermore define f_1 to f_5 . The first one maps the probabilities onto the total number of deaths per federalstate. We compute this using the inverse of the Farr formula and a sum over all ages and sex:

$$f_1(\vec{q}_m, \vec{q}_f) = \sum_{s \in \{m, f\}} \sum_{i=1}^{101} \frac{\hat{P}(y, r, s, i-1) + \hat{P}(y+1, r, s, i-1)}{2} \frac{(\vec{q}_s)_i}{1 - \alpha(i-1)(\vec{q}_s)_i}$$

Moreover, f_2 to f_5 use the life expectancy formula from Definition 3.7:

$$f_2(\vec{q}_m) = LE(\vec{q}_m, y, r, m, 0), \quad f_3(\vec{q}_f) = LE(\vec{q}_f, y, r, f, 0),$$

$$f_4(\vec{q}_m) = LE(\vec{q}_m, y, r, m, 65), \quad f_5(\vec{q}_f) = LE(\vec{q}_f, y, r, f, 65).$$

9. After defining the minimisation problem $F \circ G$ we solve it numerically for every federalstate and year $y \in [y_0, 2100]$ using a Broyden-Fletcher-Goldfarb-Shanno (BFGS) Quasi Newton method. The method requires feasible bounds and initial values to converge properly. We furthermore call the optimised vector $\vec{\theta}_{opt}$.

10. Finally, we compute the target vectors $\vec{q}_m, \vec{q}_f$ for the forecast by $G(\vec{\theta}_{opt})$ and apply the Farr Formula inversely

$$\frac{\hat{P}(y, r, s, i-1) + \hat{P}(y+1, r, s, i-1)}{2} \frac{(\vec{q}_s)_i}{1 - \alpha(i-1)(\vec{q}_s)_i}$$

to get the total number of deaths per age. We merge them with X_3 to get a complete data-set $\hat{D}$:

$$\hat{D}(y, r, s, a) : y \in [1962, 2100], r \in \{\text{federalstates}\}, s \in \{m, f\}, a \in [0, 100^+]$$

11. To be more flexible we also computed the death data for other regional-levels as well. For those with finer regional resolution we applied the same strategy as for the birth, i.e. assuming that death rates are equivalent in subregions.

Parameter Value 6.5 (Deaths $\hat{D}$).	
<i>Contents</i>	
<i>Total number of died persons living in region r with sex s and a-th birthday in year y.</i>	
<i>Resolution</i>	
<i>Time-Frame</i>	[1962, 2100]
<i>Regional-Level</i>	<i>federalstates</i>
<i>Age-Resolution</i>	[0, 100 ⁺]
<i>Other Regional-Levels</i>	<i>districts _ districts / country</i>

In comparison with the ground-truth provided by Statistics-Austria, we would score the validity of the parameter values (0 invalid, 10 fully valid) as follows:

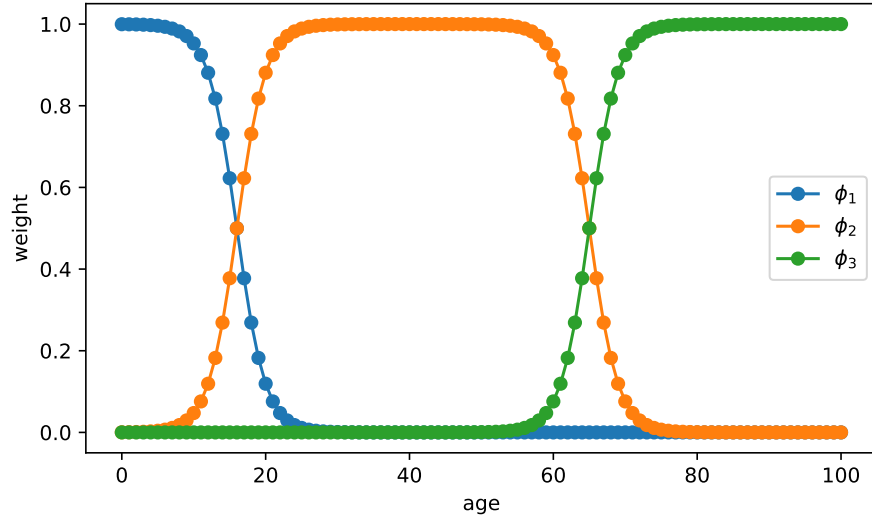


Figure 4: Activation functions f_1, f_2 and f_3 used to parametrise the distribution.

years	regions	sex	age	score	description
[1962, $y_0 - 1$]	all	all	[0, 94]	10	raw data
[1971, $y_0 - 1$]	all	all	[0, 98]	10	raw data
[2017, $y_0 - 1$]	all	all	[0, 100 ⁺]	10	raw data
[1961, 1970]	all	all	[95, 100 ⁺]	9	distributed deaths from 95+ according to distribution for 95-99 as of 1971-1973
[1971, 2016]	all	all	[99, 100 ⁺]	9	distributed deaths from 95+ according to distribution 99-100+ as of 2017-2019
[y_0 , 2100]	all	all	[99, 100 ⁺]	6	extrapolation method with the probability distribution from 2017-2019 to conserve certain sums (total deaths, life expectancies,...)

6.4 Emigrants $\hat{E}$

In this section we compute a harmonized census for emigrants.

Again, we will use the Proportional Disaggregation method, Algorithm 4.1 because it perfectly conserves the distribution.

We will use the following sources:

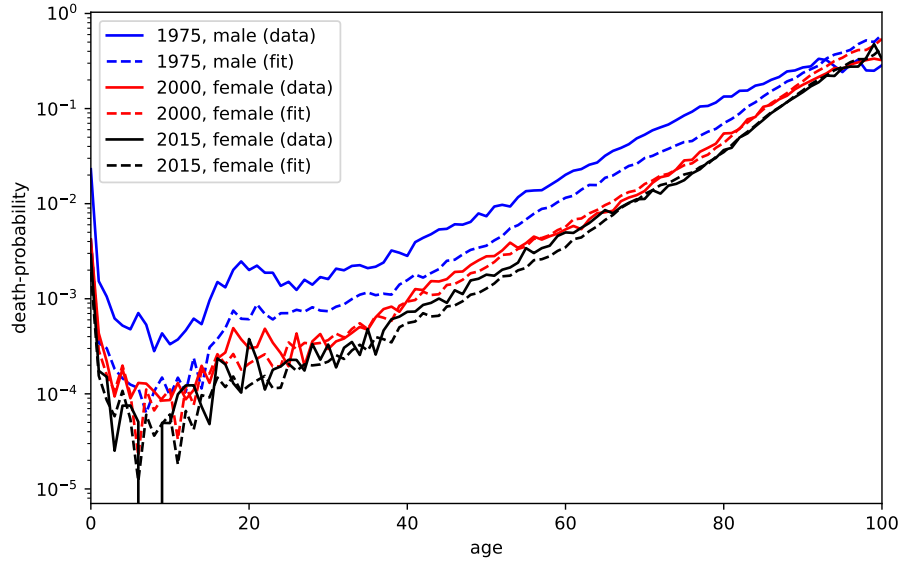


Figure 5: Fit of the parametrised curve through the death probabilities given in the data. The fit is better the more recent the data which justifies its use for the forecast.

name of source	Wanderungen mit dem Ausland von 2002 bis 2014 nach Altersgruppen, Gemeinde und Staatsangehörigkeit	Wanderungen mit dem Ausland ab 2015 nach Altersgruppen, Gemeinde und Staatsangehörigkeit	Wanderungen mit dem Ausland ab 2002 nach Alter, Geschlecht und Staatsangehörigkeit
source	S. 5.5 (emigration)	S. 5.6 (emigration)	S. 5.4 (emigration)
variable	$S_{m1}^e(y, r, s, a)$	$S_{m2}^e(y, r, s, a)$	$S_{mc}^e(y, s, a)$
$y \in$	[2002,2014]	[2015, $y_0 - 1$]	[2002, $y_0 - 1$]
$r \in$	municipalities_districts	municipalities_districts	country
$s \in$	none	none	$\{m, f\}$
$a \in$	$\{0, 5, \dots, 100^+\}$	$\{0, 5, \dots, 100^+\}$	$[0, 100^+]$
name of source	Bevölkerungs- bewegung 1961 bis 2100 nach Bundesland, Bewegungsarten und Szenarien	Tabellensammlung Außenwanderung	Population $\hat{P}$
source	S. 5.7 (emigrants)	S. 5.8 (emigrants)	Parameter Value 6.1
variable	$S_{mf}^e(y, r)$	$S_{mb}^e(y, r)$	$\hat{P}(y, r, s, a)$
$y \in$	[2002,2100]	[1996, y_0]	[1962,2101]
$r \in$	federalstates	federalstates	districts_districts
$s \in$	none	none	$\{m, f\}$
$a \in$	none	none	$[0, 100^+]$

1. In the first step we merge the two datasets S_{m1}^e and S_{m2}^e from the two sources with external migration data (Sources 5.5 and Source 5.6). They contain total number of emigrants per year, municipality_district resolution and five-year age-classes and can be merged seamlessly into a data set X_1 :

$$X_1(y, r, a) : y \in [2002, y_0 - 1], r \in \{\text{municipalities_districts}\}, a \in \{0, 5, \dots, 100^+\}.$$

2. To get a sex-resolution and a finer age-resolution, we use S_{mc}^e from Source 5.4 which contains single-age resolved migration data on country level. Since there is a straightforward mapping between the 5-year and single-year age-classes, we will use S_{mc}^e to fine-tune data from X_1 with the Proportional Disaggregation strategy:

$$\begin{aligned} X_1 : ([2002, y_0 - 1], \{\text{municipalities_districts}\}, \{0, 5, \dots, 100^+\}) \\ \rightarrow ([2002, y_0 - 1], \{\text{municipalities_districts}\}, \{m, f\}, [0, 100^+]) \\ \text{via } S_{mc}^e \text{ key } (\{0, 5, \dots, 100^+\}). \end{aligned}$$

The resulting X_2 has a sex resolution and single-age classes $[0, 100^+]$.

3. For the parameter values beyond $[2002, y_0 - 1]$ we use a similar yet much simpler strategy as for births and deaths. Since the only quantity available for computing the forecast is a scalar total number of emigrants per federalstate from S_{mf}^e , we can only assume that the age- and regional distribution of emigrants remains the same. We define

$$X_3(r, s, a) = \sum_{y=y_0-3}^{y_0-1} X_1(y, r, s, a)$$

and disaggregate S_{mf}^e :

$$\begin{aligned} S_{mf}^e : ([y_0, 2100], \{\text{federalstates}\}, \text{none}, \text{none}) \\ \rightarrow ([y_0, 2100], \{\text{municipalities_districts}\}, \{m, f\}, [0, 100^+]) \\ \text{via } X_3 \text{ key } (\{\text{federalstates}\}). \end{aligned}$$

We merge the result with X_2 to get

$$X_4(y, r, s, a) : y \in [2002, 2100], r \in \{\text{municipalities_districts}\}, s \in \{m, f\}, a \in [0, 100^+].$$

4. Since the historic data S_{mb}^e has precisely the same resolution as the forecast data, we use the same strategy to extend the data into the past until 1996 using the distribution for 2002, 2003, 2004 as reference for disaggregation.
5. Having the data with the given regional resolution (municipalities_districts) is not only unnecessarily fine for the simulation model, but also much too noisy to compute stable probabilities. As a result, we decided to aggregate the data to the districts_districts level and use this as the main output of the algorithm. Thus

$$\hat{E}(y, r, s, a) : y \in [1996, 2100], r \in \{\text{districts_districts}\}, s \in \{m, f\}, a \in [0, 100^+].$$

6. Finally, to be more flexible, we decided to keep the high-resolution data and aggregate it to various coarser regional-levels as well.

Parameter Value 6.6 (Emigrants $\hat{E}$).	
<i>Contents</i>	
<i>Total number of emigrated persons living in region r with sex s and a-th birthday in year y.</i>	
<i>Resolution</i>	
<i>Time-Frame</i>	[1996, 2100]
<i>Regional-Level</i>	<i>districts</i> _ <i>districts</i>
<i>Age-Resolution</i>	[0, 100 ⁺]
<i>Other Regional-Levels</i>	<i>country</i> / <i>federalstates</i> / <i>municipalities</i> _ <i>districts</i>

In comparison with the ground-truth provided by Statistics-Austria, we would score the validity of the parameter values (0 invalid, 10 fully valid) as follows:

years	regions	sex	age	score	description
[2002, y_0]	all	all	all	8.5	Statistical linkage between single-age data and age-class data.
[$y_0 + 1$, 2100]	all	all	all	4	Used district, age and sex distribution of the last three available years to extrapolate until 2100 given only the total expected number of emigrants per federalstate and sex
[1996, 2001]	all	all	all	4	Used district, age and sex distribution of the first three available years to extrapolate before 2002 given only the total number of emigrants per federalstate and sex

6.5 Immigrants $\hat{I}$

In this section we compute the parameter values for the immigration processes. That means, we need to identify the total number of immigrants per year, so that the interface-agent can sample them into the model. Since the model needs to sample new coordinates for the involved agents, we aim to get the same resolution as for $\hat{P}$.

Note that the model requires the numbers to be integers. Therefore, we will finally apply the Huntington-Hill Disaggregation method, Algorithm 4.2, to get a well structured parameter file.

Essentially we will use the same sources for immigration as for emigration, however we will make use of the information we already have from the other harmonised census data $\hat{B}$, $\hat{E}$, $\hat{D}$:

name of source	Wanderungen mit dem Ausland von 2002 bis 2014 nach Altersgruppen, Gemeinde und Staatsangehörigkeit	Wanderungen mit dem Ausland ab 2015 nach Altersgruppen, Gemeinde und Staatsangehörigkeit	Wanderungen mit dem Ausland ab 2002 nach Alter, Geschlecht und Staatsangehörigkeit
source	S. 5.5 (immigration)	S. 5.6 (immigration)	S. 5.4 (immigration)
variable	$S_{m1}^i(y, r, s, a)$	$S_{m2}^i(y, r, s, a)$	$S_{mc}^i(y, s, a)$
$y \in$	[2002,2014]	[2015, $y_0 - 1$]	[2002, $y_0 - 1$]
$r \in$	municipalities_districts	municipalities_districts	country
$s \in$	none	none	$\{m, f\}$
$a \in$	$\{0, 5, \dots, 100^+\}$	$\{0, 5, \dots, 100^+\}$	$[0, 100^+]$
name of source	Bevölkerungs-bewegung 1961 bis 2100 nach Bundesland, Bewegungsarten und Szenarien	Tabellensammlung Außenwanderung	Population $\hat{P}$
source	S. 5.7 (immigrants)	S. 5.8 (immigrants)	Parameter Value 6.1
variable	$S_{mf}^i(y, r)$	$S_{mb}^i(y, r)$	$\hat{P}(y, r, s, a)$
$y \in$	[2002,2100]	[1996, y_0]	[1962,2101]
$r \in$	federalstates	federalstates	municipalities_registrationdistricts
$s \in$	$\{m, f\}$	$\{m, f\}$	$\{m, f\}$
$a \in$	none	none	$[0, 100^+]$
name of source	Births $\hat{B}$	Deaths $\hat{D}$	Emigrants $\hat{E}$
source	Parameter Value 6.4	Parameter Value 6.5	Parameter Value 6.6
variable	$\hat{B}(y, s, a)$	$\hat{D}(y, s, a)$	$\hat{E}(y, s, a)$
$y \in$	[1962,2100]	[1962,2100]	[1996,2100]
$r \in$	country	country	country
$s \in$	$\{m, f\}$	$\{m, f\}$	$\{m, f\}$
$a \in$	$\{0\}$	$[0, 100^+]$	$[0, 100^+]$

1. The first steps of immigration processing will be analogous to the one from calculating the emigration parameters: We merge S_{m1}^i and S_{m2}^i into one dataset, and use S_{mc}^i to get a sex resolution and to refine the age-classes into single-age groups. Let

$$X_1(y, r, s, a) : y \in [2002, y_0 - 1], r \in \{\text{municipalities_districts}\}, s \in \{m, f\}, a \in [0, 100^+],$$

stand for the resulting file containing absolute number of immigrants per region r , sex s , age a and year y .

2. Also for the forecast we use a similar strategy as for the emigration. However, since S_{mf}^i only contains total number of immigrants without sex-resolution, and immigrants (in particular refugees) are heavily biased here, we use a different reference though: We use balance equation from Corollary 3.2 for $y \in [y_0, 2100]$

and get

$$X_2(y, s) = \left[\sum_{a=0}^{101} \hat{P}(y+1, s, a) - \hat{P}(y, s, a) - \hat{B}(y, s, a) + \hat{E}(y, s, a) + \hat{D}(y, s, a) \right].$$

Since the immigration data must contain integers only, $[\cdot]$ refers to rounding to the nearest integer. With X_2 we find the total number of immigrants per sex and year

$$X_2(y, s) : y \in [y_0, 2100], s \in \{m, f\},$$

so that they perfectly (apart from rounding) align with the other parameter-files in the model processing. It is clear that any of the four (immigrants, emigrants, births, deaths) can be computed knowing the other three. We chose immigrants in our processing steps in favour of the others since it is the most fluctuating quantity over the years.

3. For the first disaggregation step we use the federal-state distribution from S_{mf}^i using the Huntington-Hill Disaggregation strategy:

$$X_2 : ([y_0, 2100], \text{none}, \{m, f\}, \text{none}) \rightarrow ([y_0, 2100], \{\text{federalstates}\}, \{m, f\}, \text{none}) \\ \text{via } S_{mf}^i \text{ key } ([y_0, 2100]).$$

4. From this point we continue analogous to the emigration parameter calculation again. We use

$$X_3(r, s, a) = \sum_{y=y_0-3}^{y_0-1} X_2(y, r, s, a)$$

as distribution and disaggregate X_2 :

$$X_2 : ([y_0, 2100], \{\text{federalstates}\}, \{m, f\}, \text{none}) \\ \rightarrow ([y_0, 2100], \{\text{municipalities_districts}\}, \{m, f\}, [0, 100^+]) \\ \text{via } X_3 \text{ key } (\{\text{federalstates}\}, \{m, f\}).$$

We summarise all data into

$$X_4(y, r, s, a) : y \in [2002, 2100], r \in \{\text{municipalities_districts}\}, s \in \{m, f\}, a \in [0, 100^+].$$

5. We proceed analogously with the historic data from 1996 and get

$$X_5(y, r, s, a) : y \in [1996, 2100], r \in \{\text{municipalities_districts}\}, s \in \{m, f\}, a \in [0, 100^+].$$

6. To match the spatial resolution of the population data $\hat{P}$, we disaggregate X_5 further into Viennese registration-districts. With the assumption that immigrants distribute analogously to the population, we disaggregate with the Huntington-Hill Disaggregation method:

$$X_5 : ([1996, 2100], \{\text{municipalities_districts}\}, \{m, f\}, [0, 100^+]) \\ \rightarrow ([1996, 2100], \{\text{municipalities_registrationdistricts}\}, \{m, f\}, [0, 100^+]) \\ \text{via } \hat{P} \text{ key } ([1996, 2100], \{\text{municipalities_districts}\}, \{m, f\}, [0, 100^+]).$$

The resulting data-set $\hat{I}$ can be regarded as the result of this parameter calculation process.

7. To be more flexible, we also aggregated the data to coarser regional-levels.

Parameter Value 6.7 (Immigrants $\hat{I}$).	
<i>Contents</i>	
<i>Total number of immigrated persons into region r with sex s and a-th birthday in year y.</i>	
<i>Resolution</i>	
<i>Time-Frame</i>	[1996, 2100]
<i>Regional-Level</i>	<i>municipalities_registrationdistricts</i>
<i>Age-Resolution</i>	[0, 100 ⁺]
<i>Other Regional-Levels</i>	<i>country / federalstates / districts_districts / municipalities_districts</i>

In comparison with the ground-truth provided by Statistics-Austria, we would score the validity of the parameter values (0 invalid, 10 fully valid) as follows:

years	regions	sex	age	score	description
[2002, $y_0 - 1$]	real municipalities	all	all	9.5	Statistical linkage between single-age data and age-class data.
[2002, $y_0 - 1$]	Viennese registration districts	all	all	8.5	Statistical linkage with population data from Vienna.
[y_0 , 2100]	all	all	all	4	Total numbers computed by balance equation, age/sex/municipality distribution from the last three available years
[1996, 2001]	all	all	all	4	Total numbers computed by balance equation, age/sex/municipality distribution from the first three available years

6.6 Birth-, Death-, and Emigration probabilities $\hat{B}^p, \hat{D}^p, \hat{E}^p$

In the final step, we compute probabilities from the computed harmonised / augmented census information developed in the previous sections:

name of source	Population $\hat{P}$	Deaths $\hat{D}$	Emigrants $\hat{E}$	Births per Mother $\hat{B}_m$
source	Parameter Value 6.1	Parameter Value 6.5	Parameter Value 6.6	Parameter Value 6.2
variable	$\hat{P}(y, r, s, a)$	$\hat{D}(y, r, s, a)$	$\hat{E}(y, r, s, a)$	$\hat{B}_m(y, r, s, a)$
$y \in$	[1962, 2101]	[1996, 2100]	[1996, 2100]	[1962, 2100]
$r \in$	country / federalstates / districts_districts			
$s \in$	$\{m, f\}$	$\{m, f\}$	$\{m, f\}$	$\{f\}$
$a \in$	[0, 100]	[0, 100]	[0, 100]	[0, 100]

With

$$Q(y, r, s, a) = \hat{D}(y, r, s, a) + \hat{E}(y, r, s, a)$$

as the total number of individuals leaving the cohort, we apply the enhanced Farr Formula, Corollary 3.1, for computation of probabilities which are compatible with definition 3.4. Note that we can use only the overlapping time period [1996, 2100] for this computation. This results in the following data sets:

Parameter Value 6.8 (Birth-Probabilities $\hat{B}^p$).	
<i>Contents</i>	
<i>Probability that a female agent living in region r who has its a-th birthday in year y gives birth to a child until its next birthday.</i>	
<i>Resolution</i>	
<i>Time-Frame</i>	[1996, 2100]
<i>Regional-Level</i>	<i>federalstates</i>
<i>Age-Resolution</i>	[0, 100 ⁺]
<i>Other Regional-Levels</i>	<i>country / districts_ districts</i>

Parameter Value 6.9 (Death-Probabilities $\hat{D}^p$).	
<i>Contents</i>	
<i>Probability that an agent with sex s living in region r who has its a-th birthday in year y dies until its next birthday.</i>	
<i>Resolution</i>	
<i>Time-Frame</i>	[1996, 2100]
<i>Regional-Level</i>	<i>federalstates</i>
<i>Age-Resolution</i>	[0, 100 ⁺]
<i>Other Regional-Levels</i>	<i>country / districts_ districts</i>

Parameter Value 6.10 (Emigration-Probabilities $\hat{E}^p$).	
<i>Contents</i>	
<i>Probability that an agent with sex s living in region r who has its a-th birthday in year y emigrates until its next birthday.</i>	
<i>Resolution</i>	
<i>Time-Frame</i>	[1996, 2100]
<i>Regional-Level</i>	<i>districts_ districts</i>
<i>Age-Resolution</i>	[0, 100 ⁺]
<i>Other Regional-Levels</i>	<i>country / federalstates</i>

Note that we chose federalstates as the default regional-level for death and birth probabilities and districts_ districts as the one for emigration. This was chosen as a compromise between having regional features well represented, which motivates a fine regional-level, and outlier-robust probability computation, which motivates larger population counts and therefore a coarse regional-level.

6.7 Region-Families ($A_j^{r,x}$)

For all regional-levels r specified in Section 3.1 we specify the regions for the region families $A_j^{r,x}$ in form of so called multi-polygons with internal rings (MPIR). That means, every region $A_j^{r,x}$ is defined as a tuple

$$A_j^{r,x} = ((P_1, \dots, P_{J_j}), (PI_1, \dots, PI_{K_j})) : \\ P_i = ((lat, long)_1, (lat, long)_2, \dots, (lat, long)_{L_{ji}}), \\ PI_i = ((lat, long)_1, (lat, long)_2, \dots, (lat, long)_{M_{ji}})$$

We can make use of this structure in various ways. First of all, we can easily find in which region a given point lies in using an appropriate point-in-polygon algorithm (see [13] for examples). A point lies in a region $A_j^{r,x}$ if it lies in one of the polygons P_i but not in any of the interior rings PI . Moreover, we may triangulate the regions using a Constrained Delaunay triangulation [8]. This allows for efficient sampling of uniformly distributed random coordinates inside each MPIR (see [6]).

For r_{min} , which is needed to fine-tune the coordinate sampling process (we refer to the documentation of GEPOC ABM Geography in [3] for details), we use a raster-image format. I.e. the regions are defined as rectangles:

$$A_{i,j}^{r_{min}} = \{(x, y) : i\Delta lat \leq x - lat_0 \leq (i+1)\Delta lat, j\Delta long \leq y - long_0 \leq (j+1)\Delta long\}.$$

This makes a lot of computations easier and faster, e.g. finding if a given coordinate lies inside a certain region or not.

6.7.1 Multi-Polygon Processing

To get compatible model input files, we require high quality MPIR files for whole Austria which are consistent across different regional-levels. The only way to guarantee consistency is by aggregating data from finer to coarser regional-levels. Suppose we have regions A_j for the municipalities_registrationdistricts level, we can aggregate:

1. municipalities_registrationdistricts $\rightarrow$ municipalities_districts, by merging all those regions in Vienna which share the same first three digits in their id (leading to 23 districts in Vienna),
2. municipalities_districts $\rightarrow$ municipalities, by merging all regions in Vienna into one region,
3. municipalities_districts $\rightarrow$ districts_districts, by merging all regions which share the same first three digits in their id,
4. districts_districts $\rightarrow$ districts, by merging all regions in Vienna,
5. districts $\rightarrow$ federalstates, by merging all regions which share the same first number in their id,
6. federalstates $\rightarrow$ country, by merging all regions.

It remains to get proper data for the municipalities_registrationdistricts level. For this purpose, we use Source 5.11 and 5.12. The former GEOJSON file contains the MPIR borders of all municipalities in Austria, the latter for all registrationdistricts in Vienna for area-status 2017 and 2019 respectively. While the data seems to be well suited for GEPOC parametrisation, there were a couple of preprocessing steps involved before we could use it. Since many steps were done manually, it is not possible to describe them in a reproducible fashion:

1. First of all, both data files were internally inconsistent. That means that the included geometries had (a) (self-)intersections and (b) gaps in between bordering regions. Problem (a) was solved by cleaning the geometries automatically, intersecting all geometries with each other, and cutting out duplicate areas, problem (b) was solved manually afterwards. Using QGIS and MapShaper, left-out areas within Austria were identified and manually assigned to suitable regions.
2. In a second step, both datasets have been joined to get a map for the municipalities in Austria including registration districts in Vienna. For the Viennese borders, the same problems (a) and (b) could be identified which were solved analogously.
3. Since the registration districts have not changed for a while, the resulting data has area-status 2017. It remains to update the area-status to a given year y_0 which involves various merges and splits of municipalities. Until 2025, for example, these are:

2017→2018

- 40803+40819 → 40835 (Peuerbach)
- 41625+41340 → 41628 (Vorderweißenbach)

2018→2019

- 41310+41302→41345 (Helfenberg)
- 41335+41301→41346 (St. Stefan-Afiesl)

2019→2020

- 61056+61058+62347 →61060+61061 (Sankt Veit in der Südsteiermark, Straß in Steiermark)

2021→2022

- 70327+70330+70341→70370 (Matrei am Brenner)

2024→2025

- 62252+62267→62280 (Fürstenfeld)

With exception of the 3 to 2 merge in 2020, which was manually executed using the geo-software QGIS, all merges could be done automatically using standard shape tools.

Parameter Value 6.11 (Region-Families A^{r_x}).	
<i>Contents</i>	
<i>Borders of regions with different regional-levels within Austria valid for y_0-01-01.</i>	
<i>Resolution</i>	
<i>Regional-Level</i>	<i>all levels specified in Section 3.1</i>

6.7.2 Raster Image for $A^{r_{min}}$

For $A^{r_{min}}$ we use the Global Human Settlement layer of Europe, Source 5.13. To make usage more efficient we use a tight cutout around the borders of Austria.

Parameter Value 6.12 ($A^{r_{min}}$).	
<i>Contents</i>	
<i>Estimated number of inhabitants per grid-cell.</i>	
<i>Resolution</i>	
<i>Regional-Level</i>	<i>100×100m raster</i>
<i>Age Resolution</i>	<i>no age resolution</i>

6.8 Internal Migrants $\hat{O}\hat{D}, \hat{I}\hat{E}, \hat{I}, \hat{M}$

In the following we compute parameters for the internal migration models in GEPOC, in specific the interregional, biregional and fully regional model of GEPOC ABM IM.

6.8.1 Interregional Flows $\hat{O}\hat{D}$

The first parameter values which will be computed for parametrisation of internal migrations will be the ones for the interregional model (see GEPOC IM, [3]), i.e. the model which computes the destination region (solely) based on the origin region and sex.

For one-sided disaggregation we will apply the Proportional Disaggregation method, Algorithm 4.1, since this data will only be used as probabilities in the model. For two-sided disaggregation we will use the 2D Iterative Proportional Fitting (IPF 2D) method, Algorithm 4.3.

We will use the following sources:

name of source	Wanderungen innerhalb Österreichs	Bevölkerungsbewegung 1961 bis 2100 nach Bundesland, Bewegungsarten und Szenarien S. 5.7	
		Tabellensammlung Binnenwanderungen 2024 S. 5.16	
source	S. 5.14	(i. emigration and immigration)	(i. emigration and immigration)
variable	$S_{od}^e(y, r, s, r_2)$	$S_{mf}^{ie}(y, r), S_{mf}^{ii}(y, r)$	$S_{mb}^{ie}(y, r), S_{mb}^{ii}(y, r)$
$y \in$	$[2002, y_0 - 1]$	$[2002, 2100]$	$[1996, 2024]$
$r, r_2 \in$	municipalities_dist.	federalstates	federalstates
$s \in$	$\{m, f\}$	none	none
$a \in$	none	none	none

The main problem in processing of origin destination flows/probabilities from the available information is the availability of only two one-dimensional data-sets, internal emigrants S_{mf}^{ie} (origin) and internal immigrants S_{mf}^{ii} (destination), for the forecast. Not only do the data-sets lack the 2-nd dimension (origin+destination), but they are only on federalstates level and do not distinguish between male and female. To make things even worse, they do not incorporate emigrants which remained in the same federalstate.

1. In order to extend the data beyond $y_0 - 1$, we utilise the two forecasts for $[y_0, 2100]$. We attempt to establish an origin destination mapping $X_1(y, r_1, r_2)$ between the different federalstates, so that the sum over all destinations (not equal to the origin) matches $S_{mf}^{ie}(y, r)$ and the sum over all origins (not equal to the destination) matches $S_{mf}^{ii}(y, r)$, that means, we aim to find X_1 , so that

$$\forall r_1 : \sum_{r_2, r_2 \neq r} X_1(y, r_1, r_2) = S_{mf}^{ie}(y, r_1), \quad \forall r_2 : \sum_{r_1, r_1 \neq r_2} X_1(y, r_1, r_2) = S_{mf}^{ii}(y, r_2). \quad (44)$$

Thus, we utilise IPF 2D using $\vec{a}_i := S_{mf}^{ie}(y, r_i)$ and $\vec{b}_j := S_{mf}^{ii}(y, r_j)$ as marginals.

2. The algorithm requires a feasible initial estimate, since the problem with $2 \cdot n$ equations ($n = 9$ is equal to the number of federalstates) for finding $n(n - 1)$ variables is heavily underdetermined. Therefore, we use the most recent original OD-data from S_{od} . We aggregate sex, regions from municipalities_districts to federalstates, and average the flows over the last three available years of the original data

$$X_2(r_1, r_2) = \frac{1}{3} \sum_{y=y_0-3}^{y_0-1} \sum_{d_1 \in r_1} \sum_{s \in \{m, f\}} \sum_{d_2 \in r_2} S_{od}(y, d_1, s, d_2).$$

Since the forecast data does not consider migrations within the same region, we define a new dataset X_3 where we set the corresponding diagonal entries to zero:

$$\forall r : X_3(r_1, r_2) = \begin{cases} X_2(r_1, r_2), & r_1 \neq r_2, \\ 0, & \text{else.} \end{cases}$$

Note how the IDF 2D method only up and downscales values from the initial condition. Therefore any zeros in the initial matrix will remain as they are. Thus, X_3 , is a feasible starting value for the IPF 2D, since the row- and column-sums do not include the diagonal elements.

3. We regard the result of the IPF 2D method (applied with convergence threshold of 10^{-10}) as X_4 :

$$X_4(y, r_1, r_2) : y \in [y_0 - 1, 2100], r_1 \neq r_2 \in \{\text{federalstates}\}.$$

4. The diagonal entries (currently 0, by design) are still to be defined. We decided to assume that the migration within the federalstate grows/declines directly proportional to the migration into other federalstates. Therefore we simply upscale the values from the original distribution X_2 : For every year $y \in [y_0, 2100]$ and federalstate r we define

$$X_4(y, r, r) := X_2(r, r) \frac{\sum_{r_2 \neq r} X_4(y, r, r_2)}{\sum_{r_2 \neq r} X_2(r, r_2)}.$$

5. In the next step, we disaggregate X_4 to match the resolution of S_{od} using the Proportional Disaggregation method, Algorithm 4.1. As reference distribution we aggregate the last three available years:

$$X_5(r_1, s, r_2) = \sum_{y=y_0-3}^{y_0-1} S_{od}(y, d_1, s, d_2).$$

and disaggregate

$$\begin{aligned} X_4 : ([y_0, 2100], \{\text{federalstates}\}, \{\text{federalstates}\}) \\ \rightarrow ([y_0, 2100], \{\text{municipalities_districts}\}, \{m, f\}, \{\text{municipalities_districts}\}) \\ \text{via } X_5 \text{ key } (\{\text{federalstates}\}, \{\text{federalstates}\}). \end{aligned}$$

We merge the result with S_{od} and get

$$X_6(y, r_1, s, r_2) : y \in [2002, 2100], r_1, r_2 \in \{\text{municipalities_districts}\}, s \in \{m, f\}.$$

6. In a final step, we proceed analogously for the historic internal migration data S_{mb}^{ii} and S_{mf}^{ie} since it has the same structure as the forecast data. However, we use years 2002 – 2004 to compute the distribution / initial condition for the IPF 2D. The results of all procedures are joined along the time-axis to get one final dataset

$$\hat{OD}(y, r_1, s, r_2) : y \in [1996, 2100], r_1, r_2 \in \{\text{municipalities_registrationdistricts}\}, s \in \{m, f\}.$$

7. While this file will be regarded as the primary output of the processing, aggregation to coarser regional-levels will be done to become more flexible.

Parameter Value 6.13 (Interregional Flows $\hat{OD}$).	
<i>Contents</i>	
<i>Total number of persons migrating from region r_1 into region r_2 with sex s in year y.</i>	
<i>Resolution</i>	
<i>Time-Frame</i>	[1996, 2100]
<i>Regional-Level</i>	municipalities_ districts
<i>Age-Resolution</i>	none
<i>Other Regional-Levels</i>	federalstates / districts_ districts

In comparison with the ground-truth provided by Statistics-Austria, we would score the validity of the parameter values (0 invalid, 10 fully valid) as follows:

years	regions	sex	age	score	description
[2002, $y_0 - 1$]	all	all	all	10	raw data.
[y_0 , 2100]	all	all	all	3	Forecast based on immigration and emigration numbers on federalstate level and the IPF 2D method.
[1996, 2001]	all	all	all	3	Historic flows are based on immigration and emigration numbers on federalstate level and the IPF 2D method.

6.8.2 Internal Emigrants $\hat{IE}$

In the next step we compute the harmonised census for the internal emigrants $\hat{IE}$ which will be required to compute the internal emigration probability for all 3 IM models. That means, we focus on internal emigrants and their age and sex distribution without analysing their destination region. The Proportional Disaggregation method, Algorithm 4.1, is used for disaggregation of age-classes.

We will use the following sources:

name of source	Binnenwanderungen innerhalb Österreichs ab 2002	Interregional Flows $\hat{OD}$
source	S. 5.15 (i. emigrants)	Parameter Value 6.13
variable	$S_m^{ie}(y, r, s, a)$	$\hat{OD}(y, r, s, r_2)$
$y \in$	$[2002, y_0 - 1]$	$[1996, 2100]$
$r, r_2 \in$	districts_districts	districts_districts
$s \in$	$\{m, f\}$	$\{m, f\}$
$a \in$	$[0, 95^+]$	none

The core of the age-dependent internal-emigrant is source S_m^{ie} which was collected from Statistics Austria's STATCube. To establish the forecast, we utilise the $\hat{OD}$ parameter values which were derived in the previous section. This way, we remain consistent w.r. to total numbers.

1. To compute the forecast, we use the age distribution of S_m^{ie} from $y_0 - 3$ to $y_0 - 1$:

$$X_1(r, s, a) = \sum_{y=y_0-3}^{y_0-1} S_m^{ii}(y, r, s, a)$$

and the aggregated emigrants from $\hat{OD}$ (restricted to $[y_0, 2100]$)

$$X_2(y, r, s) = \sum_{r_1} \hat{OD}(y, r, s, r_1),$$

and de-aggregate with the Proportional Method:

$$\begin{aligned} X_2 : ([y_0, 2100], \{\text{districts_districts}\}, \{m, f\}) \\ \rightarrow ([y_0, 2100], \{\text{districts_districts}\}, \{m, f\}, [0, 95^+]) \\ \text{via } X_1 \text{ key } (\{\text{districts_districts}\}, \{m, f\}). \end{aligned}$$

2. We proceed identically with the historic OD data between 1996 and 2001 using the distribution of 2002 to 2004. The results are merged with S_m^{ie} into

$$\hat{IE}(y, r, s, a) : y \in [1996, 2100], r \in \{\text{districts_districts}\}, a \in [0, 95^+].$$

3. Aggregation to federalstate level is considered as secondary output of this procedure.

Parameter Value 6.14 (Internal Emigrants $\hat{IE}$).	
<i>Contents</i>	
<i>Total number of persons to internally migrate from region r with sex s and age a (at the date of migration) in year y.</i>	
<i>Resolution</i>	
<i>Time-Frame</i>	[1996, 2100]
<i>Regional-Level</i>	<i>districts_districts</i>
<i>Age-Resolution</i>	[0, 95 ⁺]
<i>Other Regional-Levels</i>	<i>federalstates</i>

In comparison with the ground-truth provided by Statistics-Austria, we would score the validity of the parameter values (0 invalid, 10 fully valid) as follows:

years	regions	sex	age	score	description
[2002, y_0]	all	all	all	10	raw data.
[y_0 , 2100]	all	all	all	4	Age distribution from last three available years, <i>districts_districts</i> distribution from <i>federalstates</i> .
[1996, 2001]	all	all	all	4	Age distribution from first three available years, <i>districts_districts</i> distribution from <i>federalstates</i> .

6.8.3 Internal Immigrants $\hat{II}$

To compute parameter $\hat{II}$ for the biregional model, we focus on internal immigrants and their age and sex distribution without analysing their origin region. The Proportional Disaggregation method, Algorithm 4.1, is used for disaggregation of age-classes.

We will use the following sources:

name of source	Binnenwanderungen innerhalb Österreichs ab 2002	Interregional Flows $\hat{OD}$	Internal Emigrants $\hat{IE}$
source	S. 5.15 (i. immigrants)	Parameter Value 6.13	Parameter Value 6.14
variable	$S_m^{ii}(y, r, s, a)$	$\hat{OD}(y, r, s, r_2)$	$\hat{IE}(y, r, s, a)$
$y \in$	[2002, $y_0 - 1$]	[1996, 2100]	[1996, 2100]
$r, r_2 \in$	<i>districts_districts</i>	<i>districts_districts</i>	<i>districts_districts</i>
$s \in$	$\{m, f\}$	$\{m, f\}$	$\{m, f\}$
$a \in$	[0, 95 ⁺]	none	[0, 95 ⁺]

With one exception, the computation for internal immigrants is identical to that for internal emigrants. However, a final adjustment needs to be made to guarantee that the age distributions of the internal emigrants and immigrants align.

1. We proceed analogous to the internal emigrant harmonisation, while the OD data ($\hat{OD}$ parameter file) is aggregated w.r. to the destination dimension. We denote the result of this procedure as

$$X_1(y, r, s, a) : y \in [1996, 2100], r \in \{\text{districts_districts}\}, a \in [0, 95^+].$$

2. Unfortunately, this way, we receive a mismatch with age distribution of the computed internal emigrants $IE(y, r, s, a)$ since

$$\forall y \in [1996, 2001] \cup [y_0, 2100], s \in \{m, f\} : \sum_r IE(y, r, s, a) = \sum_r II(y, r, s, a)$$

is not guaranteed. To solve this problem, we once again apply the IPF 2D method: for any year $y \in [1996, 2001] \cup [y_0, 2100]$ and sex s , we use

$$\sum_r IE(y, r, s, a), \text{ and } \sum_a II(y, r, s, a)$$

as marginals, and $II(y, r, s, a)$ as initial matrix of the algorithm. Note that the algorithm is now applied for the region and age dimension, instead of the two region dimensions, as for the OD computation. Therefore, the result is guaranteed to match the age distribution of the aggregated internal emigration data and to maintain the overall distribution of destination regions.

3. The results of the IPF 2D (tolerance 10^{-10}) are merged with S_m^{ii} into

$$\hat{II}(y, r, s, a) : y \in [1996, 2100], r \in \{\text{districts_districts}\}, a \in [0, 95^+].$$

4. Aggregation to federalstates level is considered as secondary output of this procedure.

Parameter Value 6.15 (Internal Immigrants $\hat{II}$).	
<i>Contents</i>	
<i>Total number of persons to migrate into region r_2 with sex s and age a (at the date of migration) in year y.</i>	
<i>Resolution</i>	
<i>Time-Frame</i>	[1996, 2100]
<i>Regional-Level</i>	<i>districts_districts</i>
<i>Age-Resolution</i>	[0, 95 ⁺]
<i>Other Regional-Levels</i>	<i>federalstates</i>

In comparison with the ground-truth provided by Statistics-Austria, we would score the validity of the parameter values (0 invalid, 10 fully valid) as follows:

years	regions	sex	age	score	description
[2002, y_0]	all	all	all	10	raw data.
[y_0 , 2100]	all	all	all	4	Age distribution from last three available years, <i>districts_districts</i> distribution from federalstates.
[1996, 2001]	all	all	all	4	Age distribution from first three available years, <i>districts_districts</i> distribution from federalstates.

6.8.4 Internal Migrants by Age, Origin and Destination $\hat{M}$

In this section we will derive migration flows $\hat{M}$ which describe the number of persons with sex s and age a emigrating from a certain origin r_1 into a certain destination r_2 . Note that all parameter values computed for internal migration so far have in common that they do not contain full information for parametrisation of the migration process. While the origin-destination parameters $\hat{OD}$ lack the age resolution, internal emi- and immigrants $\hat{IE}$ and $\hat{II}$ lack the corresponding destination/origin mapping. Core of this computation is the 3-dimensional Iterative Proportional Fitting, Algorithm 4.4.

We will use no additional sources but only derived parameter values:

name of source	Interregional Flows $\hat{OD}$	Internal Emigrants $\hat{IE}$	Internal Immigrants $\hat{II}$
source	Parameter Value 6.13	Parameter Value 6.14	Parameter Value 6.15
variable	$\hat{OD}(y, r_1, s, r_2)$	$\hat{IE}(y, r_1, s, a)$	$\hat{II}(y, r_2, s, a)$
$y \in$	[1996,2100]	[1996,2100]	[1996,2100]
$r_1, r_2 \in$	districts_districts/ federalstates	districts_districts/ federalstates	districts_districts/ federalstates
$s \in$	$\{m, f\}$	$\{m, f\}$	$\{m, f\}$
$a \in$	none	$[0, 95^+]$	$[0, 95^+]$

Since no additional data from Statistics Austria is used, the only knowledge we have for computing $\hat{M}$ is the causal relation between the data: For all years y and sex s , we know that

$$\forall r_1, a : \sum_{r_2} \hat{M}(y, r_1, s, a, r_2) = \hat{IE}(y, r_1, s, a), \quad (45)$$

$$\forall r_2, a : \sum_{r_1} \hat{M}(y, r_1, s, a, r_2) = \hat{II}(y, r_2, s, a), \quad (46)$$

$$\forall r_1, r_2 : \sum_a \hat{M}(y, r_1, s, a, r_2) = \hat{OD}(y, r_1, s, r_2). \quad (47)$$

Let n denote the total number of regions (in our case around 100 districts) and m the total number of age-classes (in our case 95), then the problem has $2mn + n^2$ equations and n^2m unknowns.

This provides the perfect setup for applying IPF 3D. We use the three datasets at districts_districts level as marginals, the simple tensor $M_0(y, r_1, s, a, r_2) \equiv 1$ as initial condition, and tolerance 10^{-4} is used as stopping criterion. This choice for the initial condition turned out to be sufficiently good to result in convergence in less than 200 iterations.

The result

$$\hat{M}(y, r_1, s, a, r_2) : y \in [1996, 2100], r_1, r_2 \in \{\text{districts_districts}\}, s \in \{m, f\}, a \in [0, 95^+]$$

is regarded as the primary output of this procedure. Aggregation to the federalstates level is provided as a secondary outcome.

Parameter Value 6.16 (Internal Migrants $\hat{M}$).

Contents	
<i>Total number of persons migrating from region r_1 into region r_2 with sex s and age a (at the date of migration) within year y.</i>	
Resolution	
Time-Frame	[1996, 2100]
Regional-Level	<i>districts_districts</i>
Age-Resolution	[0, 95+]
Other Regional-Levels	<i>federalstates</i>

In comparison with the ground-truth provided by Statistics-Austria, we would score the validity of the parameter values (0 invalid, 10 fully valid) as follows:

years	regions	sex	age	score	description
[2002, $y_0 - 1$]	all	all	all	8	Matched with IPF 3D.
[y_0 , 2100]	all	all	all	2.5	Forecast based on immigration and emigration numbers on federalstate level and IPF matching.
[1996, 2001]	all	all	all	2.5	Historic information based on immigration and emigration numbers on federalstate level and IPF matching.

6.9 Internal Emigration Probability $\hat{I}E^p$

Finally, after harmonising the internal migration census, the internal emigration probability which is required for all three internal migration models of GEPOC ABM IM is computed analogously to all other probabilities using the enhanced Farr formula, Corollary 3.1.

name of source	Population $\hat{P}$	Deaths $\hat{D}$	Emigrants $\hat{E}$	Internal Emigrants $\hat{I}E$
source	Parameter Value 6.1	Parameter Value 6.5	Parameter Value 6.6	Parameter Value 6.14
variable	$\hat{P}(y, r, s, a)$	$\hat{D}(y, r, s, a)$	$\hat{E}(y, r, s, a)$	$\hat{I}E(y, r, s, a)$
$y \in$	[1962, 2101]	[1962, 2100]	[1996, 2100]	[1996, 2100]
$r \in$		federalstates / districts_districts		
$s \in$	$\{m, f\}$	$\{m, f\}$	$\{m, f\}$	$\{m, f\}$
$a \in$	[0, 100]	[0, 100]	[0, 100]	[0, 95+]

After aggregating all data to the same age resolution [0, 95+], again,

$$Q(y, r, s, a) = \hat{D}(y, r, s, a) + \hat{E}(y, r, s, a)$$

is considered the total number of individuals leaving the cohort, then we apply the enhanced Farr Formula, Corollary 3.1, for computation of an internal emigration probability which is compatible with Definition 3.4. We want to emphasise that an internally migrated agent remains a component of the model and is still capable of having other events in the course of the remaining life-year. Therefore, internal migrants must not be considered as cohort-leavers.

This results in the following data-set:

Parameter Value 6.17 (Internal Emigration Probabilities $\hat{I}E^p$).	
<i>Contents</i>	
<i>Probability that an agent living in region r who has its a-th birthday in year y emigrates internally until its next birthday.</i>	
<i>Resolution</i>	
<i>Time-Frame</i>	[1996, 2100]
<i>Regional-Level</i>	<i>districts_districts</i>
<i>Age-Resolution</i>	[0, 95 ⁺]
<i>Other Regional-Levels</i>	<i>federalstates</i>

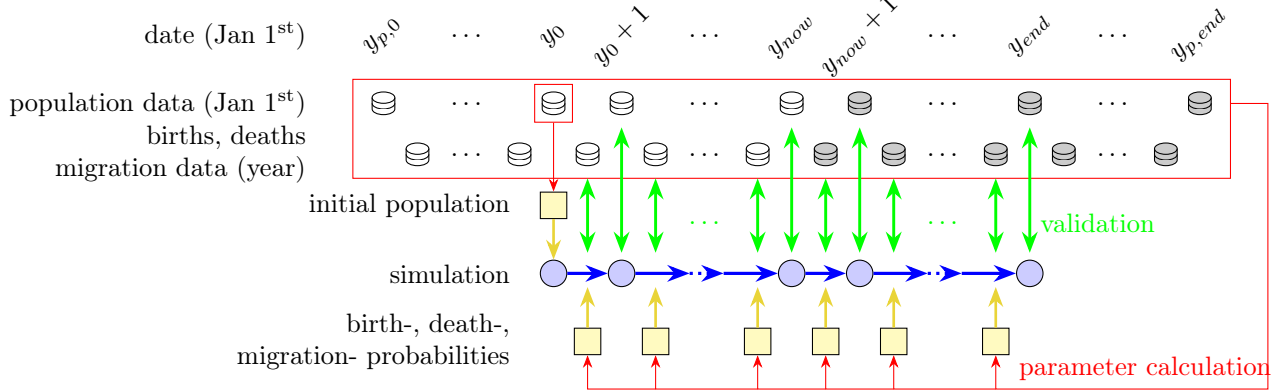


Figure 6: General parametrisation and validation scheme of GEPOC ABM. Population, death, birth, and migration counts in form of census counts (white) and forecasts (grey) pose the source for the parameter values and the validation reference at the same time.

7 Validation

In the following we will quantitatively compare the fully parametrised model GEPOC ABM (see, [3]) with the given information about the population of Austria as reference (see Figure 6).

Reference. We will treat the sources described in Section 5 as calibration reference. That means that we consider the available data about the Austrian population as ground truth and do not question the accuracy of the data compared to the actual population. This is particularly relevant for the validation of the future population dynamics, where the main forecasting scenario of Statistik Austria (“Prognose Hauptszenario”) will be considered as ground truth, even though there is not yet any actual population to compare to. The model will be regarded as *quantitatively valid* if the differences between the synthetic census, generated by the simulation, and the reference data is sufficiently small.

Error Metric. For each simulation scenario (see below), we run Monte Carlo runs and average the results. The number of runs was carefully chosen by contrasting the sample standard deviation with the sample mean (using the Gaussian approach described in [7]). From the simulations, we will carry out all information about demographic quantities introduced in Section 3.2 on a yearly basis and using the highest possible regional, age, and sex resolution. Data for any lower resolution required for comparison with data is aggregated in post-processing.

The quantitative validation will be performed using the following error function for a time series Y compared to a reference X :

$$e_{max}(Y, X) := \max_{t \in [t_1, t_2]} \left(\frac{Y(t) - X(t)}{\max(1, X(t))} \right), \quad e_{min}(Y, X) := \min_{t \in [t_1, t_2]} \left(\frac{Y(t) - X(t)}{\max(1, X(t))} \right). \quad (48)$$

We hereby quantify the highest positive and negative relative distance from the outcome Y to the reference X on the time-frame $[t_1, t_2]$. The max expression is used to avoid divisions by zero. We argue that this quantity is comparably strict in contrast to usual scores, since it harshly punishes short term fluctuations, which would be averaged when using e.g. classical R^2 scores. Moreover, since signs are not cancelled out, it is also well suitable for analysing over- and underestimation.

Simulation Scenarios. In total, seven simulation scenarios (SC1-SC7) are defined to analyse the validity of the model.

In the first part, we will investigate and compare the outcomes of one scenario for GEPOC ABM Geography (**SC1**) and one for GEPOC ABM IM (**SC2**) using the full-regional internal migration model. SC2 will be considered as the main-scenario and a high-accuracy fit is to be expected. For SC1, we want to analyse how well the model can fit the data without regarding internal migration. For both scenarios we will use January 2000 as start date and investigate the population change until 2050 in yearly steps. This way, the time-span covers three different intervals w.r. to the parametrization: for 1996 to 1999 (2001, for migration respectively), data for demographic change is only available on low-resolution. Then, from 2000 (2002) until 2024, the model parameters are estimated based on high-resolution data. Then again, from 2025, the parameters are based on low-resolution forecast data. For details of which parameter values and which regional-levels were used for which parameter in which scenario, we refer to Table 4.

In the second part, two additional scenarios for GEPOC ABM IM are evaluated, one with the biregional (**SC3**) and one with the interregional (**SC4**) internal migration model. They are compared with the full regional model (**SC2**) to analyse the impact of the internal migration modelling approach. Finally, in the third part, we will investigate three additional scenarios where other fundamental model-parameters from SC2 are varied. Scenario **SC5** will extend the simulations beyond 2050 until 2100, **SC6** uses monthly instead of yearly macro-steps and **SC7** will use a downscaled population in which every agent represents 10 real inhabitants of Austria. We refer to the corresponding sections for details about parametrisation and start with the evaluation of SC1 and SC2.

7.1 Population

Since it is the most important variable for GEPOC, we start by comparing the simulation results from SC1 and SC2 with the data for the total population.

7.1.1 Comparison with Source 5.1 (Bevölkerungsstand)

First of all, we compare the values with the highly resolved data given by Source 5.1 for the period between 2002 and 2025.

Total and sex. Figure 7 shows the a comparison between SC1, SC2 and the reference from Source 5.1 for the total male and female population. After 25 years of simulation, the overall population per sex remained within 0.4% deviation of the actual population and the two cohorts are less than 20000 persons off. The IM model performs even better, where the error remained within 0.15%. This is not surprising, since the model should, in theory, conserve the regional age structure better. Yet the figure also indicates a trend to slightly overestimate the total population for both models.

Age-classes. Difference curves for twenty-year age-classes are displayed in Figure 8. The differences vary between -13000 and 40000 individuals, which may sound large at the first glance, but effectively correspond to relative differences within $\pm 1.6\%$. In contrast to Figure 8, Figure 9 displays the absolute quantities where, the curves from SC1 and SC2 are almost indistinguishable from the reference data. This type of plot is only being used this single time to visualise that an offset of 40000 persons might only seem large on the first glance. Anyhow, in particular for the age class 20-39, the internal migration model gives more accurate results.

Federal-states. Figure 10 shows the analogous plot to Figure 8 for the nine federal states of Austria. As expected, scenario SC1 without internal migration results in larger deviations, up to 7%. The two federal

Table 4: Parametrisation setup for simulation scenarios SC1 and SC2.

		SC1	SC2
GEPOC Model		Geography	IM
sim start-time	t_0	2000-01-01	2000-01-01
time-step lengths	Δt	years	years
sim end-time	t_e	2050-01-01	2050-01-01
scale	σ	1.0	1.0
Monte-Carlo runs		9	9
probability for male person-agent at birth	$\hat{\alpha}_m$	P.V. 6.3	P.V. 6.3
maximum age	$\hat{a}_{max}$	100	100
regional-level for initial population	r_0	municipalities_registrationdistricts	municipalities_registrationdistricts
data for initial population	$\hat{P}$	P.V. 6.1	P.V. 6.1
regional-level for initial population refinement	r_{min}	100x100[m] raster	100x100[m] raster
regional-level for births	r_b	federalstates	federalstates
birth probabilities	$\hat{B}^p$	P.V. 6.8	P.V. 6.8
regional-level for deaths	r_d	federalstates	federalstates
death probabilities	$\hat{D}^p$	P.V. 6.9	P.V. 6.8
regional-level for emigrants	r_e	districts_districts	districts_districts
emigration probabilities	$\hat{E}^p$	P.V. 6.10	P.V. 6.10
regional-level for immigrants	r_i	municipalities_registrationdistricts	municipalities_registrationdistricts
immigrants	$\hat{I}$	P.V. 6.7	P.V. 6.7
internal migration model			full regional
regional-level for internal migration	r_{ie}	-	districts_districts
internal emigration probabilities	$\hat{IE}^p$	-	P.V. 6.14
internal migrants	$\hat{M}^p$	-	P.V. 6.16
region families for all regional-levels except r_{min}	$A_j^{r_0}$	P.V. 6.11	P.V. 6.11
region families for population refinement (level r_{min})	$A_j^{r_{min}}$	P.V. 6.12	P.V. 6.12

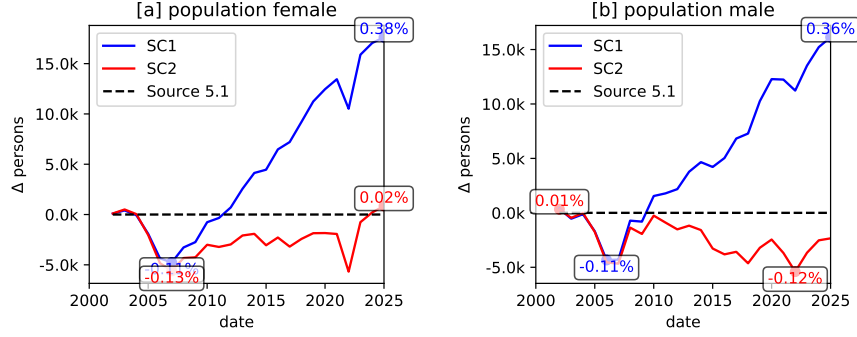


Figure 7: Differences between the male and female population data (Source 5.1) and the two simulation scenarios SC1 and SC2.

states with the highest deviations, AT-1 and AT-2, are well known for being subject to high internal migration-dynamics. Scenario SC2 with internal migration showed its highest deviation in AT-6 and AT-9 with roughly -0.5% .

Finer aggregation levels. The maximum deviations for the mentioned and also more detailed aggregation levels are summarised in Table 5. Regionally, comparably large offsets can be found for elderly people in both models, however, SC2 performs, on average, much better.

7.1.2 Comparison with Source 5.3 (Bevölkerung zum Jahresanfang 1952 bis 2101)

For analysing the quality beyond the period, for which actual data is available, we use Source 5.3 between 2026 to 2050. Note that the age and regional resolution given by the source is sufficient to perform the same analysis as in the prior section. Figures 11, 12, and 13 provide the analogous plots to the ones shown in the previous section, but for the time-frame between 2026-01-01 and 2050-01-01.

Total and sex. Figure 11 shows that the trend towards overestimation indicated by Figure 7 is not prolonged, but turns into a clear underestimation with -1% at the end of the period. Here, both models perform roughly equally well. In general, the trend towards underestimation can be explained by inaccurate disaggregation of the low-resolution forecast data. At least on this high aggregation level, the problem seems to be more severe than the missing internal migration and causes both model scenarios to become inaccurate eventually.

Federal-states, age-classes and finer. On the regional-level, the results from SC1 start to differ more heavily from the ground truth, whereas SC2 captures the regional trends much better and keeps them within the $\pm 2\%$ region. The differences for the age-classes remain in a similar magnitude as for 2002-2025 for both models. On a less aggregated scope (see Table 6), we find the largest differences for the SC1 model for the [20,39] cohort in AT-2, the elderly in AT-9 (Vienna) and the [0,19] cohort in AT-1. The latter two can be explained by the high internal migration dynamics between AT-1, AT-3 and the city of Vienna AT-9.

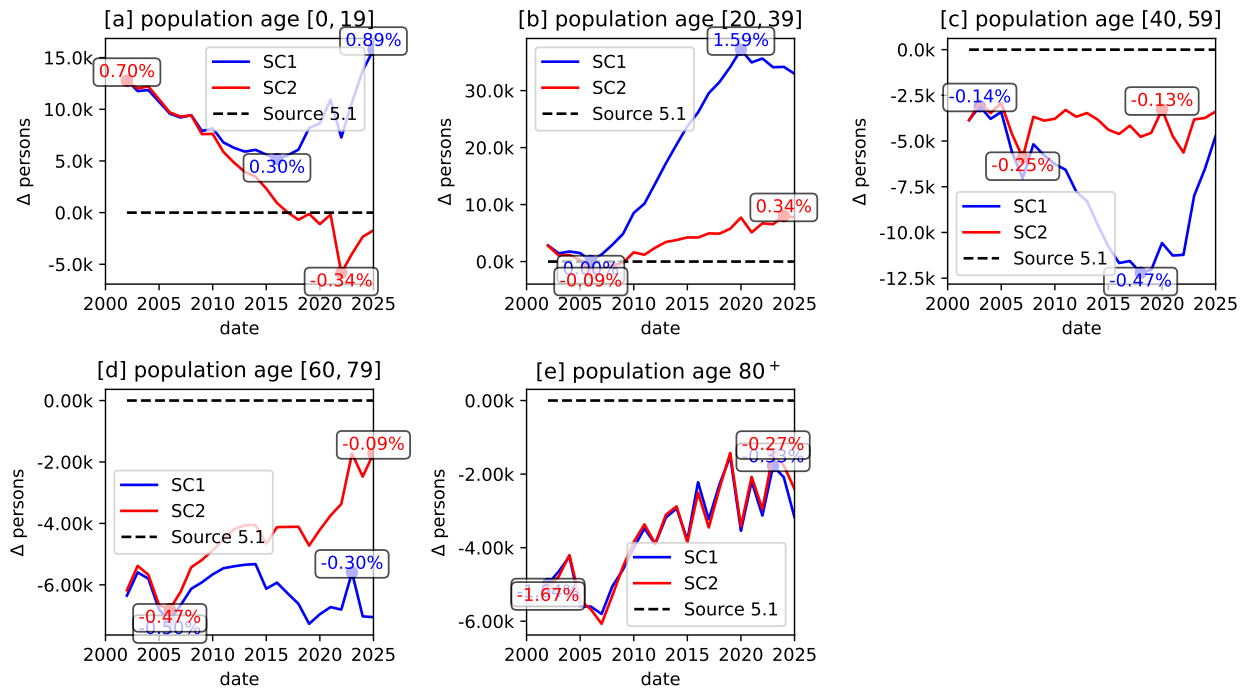


Figure 8: Differences between the population data (Source 5.1) and the two simulation scenarios SC1 and SC2 after aggregation to Austria and 20-year age classes.

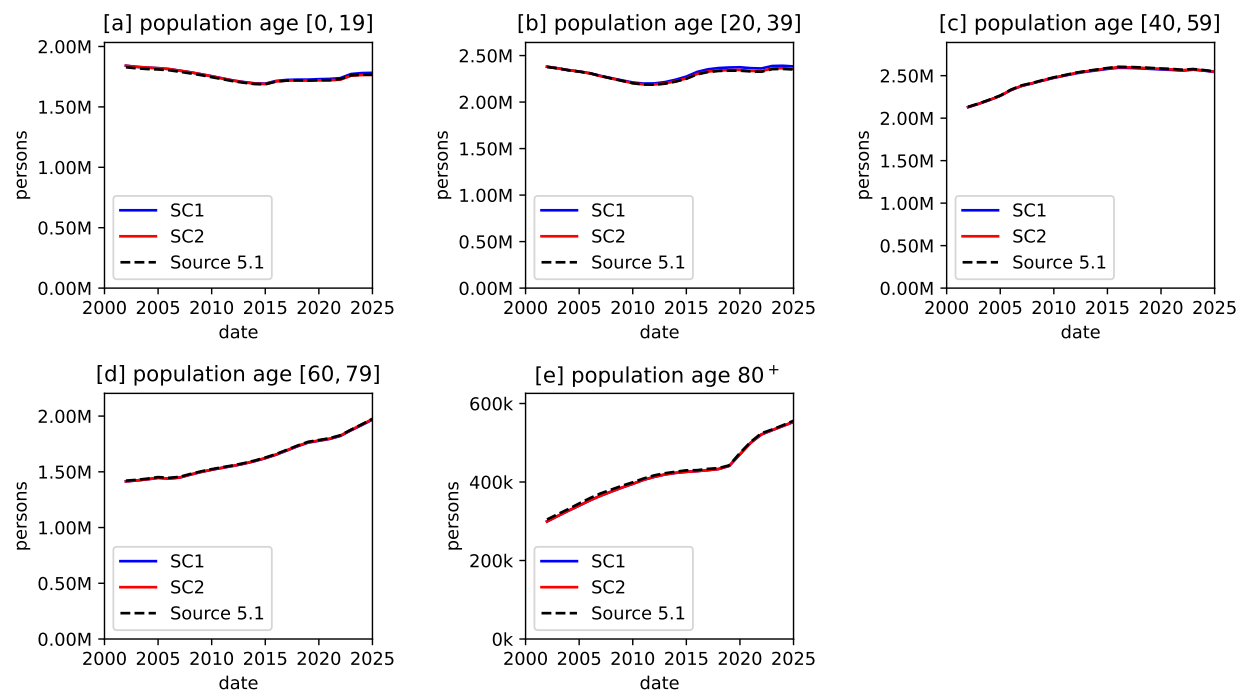


Figure 9: Comparison between the population data (Source 5.1) and the two simulation scenarios SC1 and SC2 after aggregation to Austria and 20-year age classes.

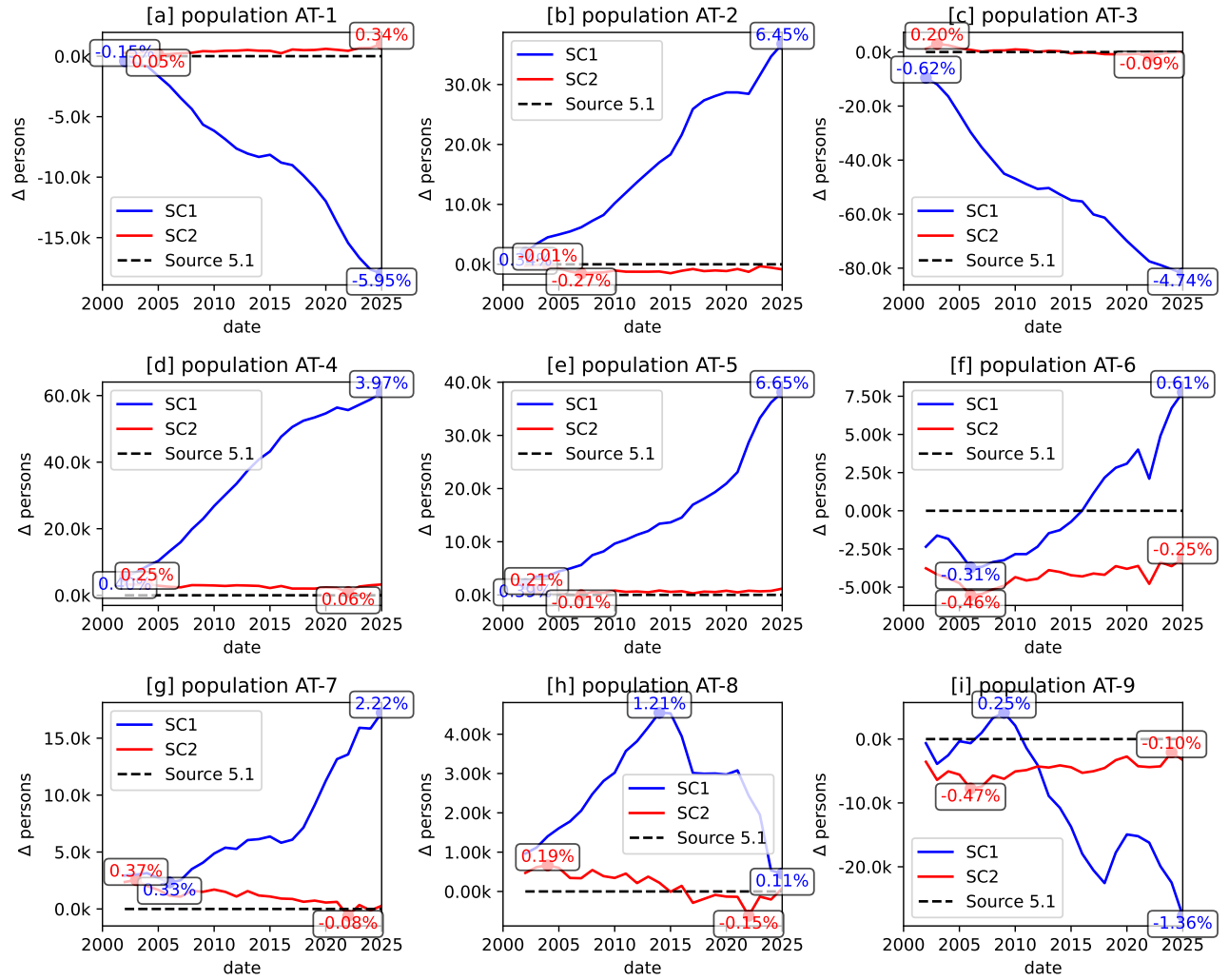


Figure 10: Differences between the population data (Source 5.1) and the two simulation scenarios SC1 and SC2 for all federalstates of Austria.

Table 5: Relative maximum differences between the total population data (Source 5.1 and the simulations SC1 and SC2 between 2002 and 2025.

region	sex	age	SC1 (e_{min}, e_{max})		SC2 (e_{min}, e_{max})	
-	-	-	-0.11%	0.37%	-0.12%	0.01%
-	female	-	-0.11%	0.38%	-0.13%	0.02%
-	male	-	-0.11%	0.36%	-0.12%	0.01%
-	-	[0, 19]	0.30%	0.89%	-0.34%	0.70%
-	-	[20, 39]	0.00%	1.59%	-0.09%	0.34%
-	-	[40, 59]	-0.47%	-0.14%	-0.25%	-0.13%
-	-	[60, 79]	-0.50%	-0.30%	-0.47%	-0.09%
-	-	80 ⁺	-1.64%	-0.33%	-1.67%	-0.27%
AT-1	-	-	-5.95%	-0.15%	0.05%	0.34%
AT-2	-	-	0.34%	6.45%	-0.27%	-0.01%
AT-3	-	-	-4.74%	-0.62%	-0.09%	0.20%
AT-4	-	-	0.40%	3.97%	0.06%	0.25%
AT-5	-	-	0.39%	6.65%	-0.01%	0.21%
AT-6	-	-	-0.31%	0.61%	-0.46%	-0.25%
AT-7	-	-	0.33%	2.22%	-0.08%	0.37%
AT-8	-	-	0.11%	1.21%	-0.15%	0.19%
AT-9	-	-	-1.36%	0.25%	-0.47%	-0.10%
AT-1	-	[0, 19]	-6.57%	0.19%	-0.34%	1.19%
AT-2	-	[0, 19]	0.77%	13.23%	-0.21%	0.78%
AT-3	-	[0, 19]	-6.86%	-0.18%	-0.49%	0.83%
AT-4	-	[0, 19]	0.76%	6.33%	-0.00%	0.83%
AT-5	-	[0, 19]	1.33%	8.73%	0.14%	0.95%
AT-6	-	[0, 19]	0.35%	2.00%	-0.39%	0.71%
AT-7	-	[0, 19]	0.67%	2.18%	-0.64%	1.14%
AT-8	-	[0, 19]	-1.15%	0.63%	-0.71%	0.60%
AT-9	-	[0, 19]	-2.73%	3.92%	-0.73%	0.36%
AT-1	-	[20, 39]	1.01%	9.21%	-0.16%	0.97%
AT-2	-	[20, 39]	1.19%	15.99%	-0.40%	0.25%
AT-3	-	[20, 39]	-0.74%	5.00%	0.12%	0.49%
AT-4	-	[20, 39]	1.20%	8.82%	0.14%	0.63%
AT-5	-	[20, 39]	0.77%	11.85%	-0.09%	0.43%
AT-6	-	[20, 39]	-0.25%	2.38%	-0.80%	0.35%
AT-7	-	[20, 39]	-0.08%	2.90%	-0.03%	0.32%
AT-8	-	[20, 39]	0.06%	4.12%	0.11%	0.38%
AT-9	-	[20, 39]	-11.26%	-0.92%	-0.36%	0.71%
AT-1	-	[40, 59]	-8.15%	-0.74%	-0.01%	0.30%
AT-2	-	[40, 59]	-0.36%	4.50%	-0.24%	-0.01%
AT-3	-	[40, 59]	-10.05%	-0.74%	-0.07%	0.09%
AT-4	-	[40, 59]	0.11%	3.06%	0.13%	0.25%
AT-5	-	[40, 59]	0.02%	4.79%	-0.20%	0.23%
AT-6	-	[40, 59]	-0.65%	0.66%	-0.81%	-0.43%
AT-7	-	[40, 59]	0.35%	2.18%	-0.03%	0.30%
AT-8	-	[40, 59]	-0.43%	0.18%	-0.07%	0.16%
AT-9	-	[40, 59]	0.02%	4.84%	-0.84%	-0.45%
AT-1	-	[60, 79]	-12.19%	-0.86%	-0.45%	0.22%
AT-2	-	[60, 79]	-2.27%	-0.40%	-0.55%	-0.23%
AT-3	-	[60, 79]	-6.11%	-0.81%	-0.42%	0.02%
AT-4	-	[60, 79]	-0.61%	0.35%	-0.60%	0.18%
AT-5	-	[60, 79]	-0.57%	2.73%	-0.63%	0.13%
AT-6	-	[60, 79]	-1.78%	-0.41%	-0.60%	-0.36%
AT-7	-	[60, 79]	-0.05%	1.91%	-0.14%	0.27%
AT-8	-	[60, 79]	-0.23%	1.50%	-0.27%	0.29%
AT-9	-	[60, 79]	-0.07%	7.50%	-0.72%	-0.38%
AT-1	-	80 ⁺	-8.35%	-2.38%	-1.80%	0.73%
AT-2	-	80 ⁺	-3.60%	-1.01%	-3.06%	-0.21%
AT-3	-	80 ⁺	-5.73%	-2.67%	-1.84%	-0.21%
AT-4	-	80 ⁺	-2.43%	-0.48%	-2.45%	-0.46%
AT-5	-	80 ⁺	-2.33%	0.74%	-2.23%	-0.05%
AT-6	-	80 ⁺	-2.29%	-0.75%	-2.70%	-0.42%
AT-7	-	80 ⁺	-1.65%	0.83%	-1.40%	0.54%
AT-8	-	80 ⁺	-2.66%	0.52%	-2.58%	-0.07%
AT-9	-	80 ⁺	-0.01%	6.99%	-1.28%	-0.22%

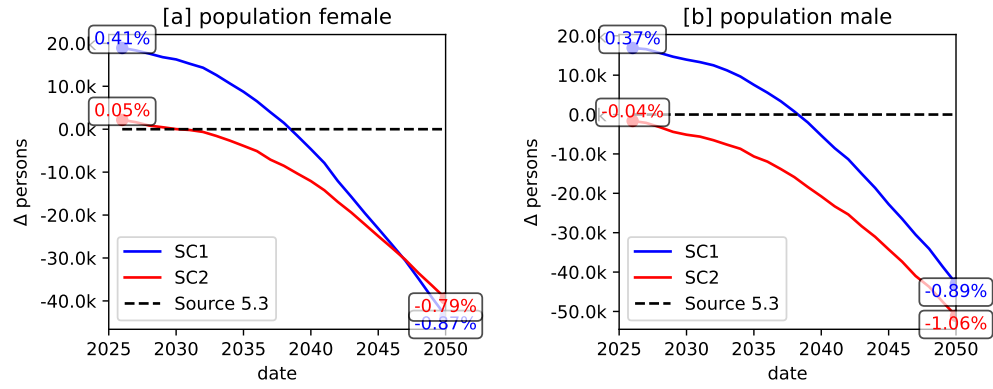


Figure 11: Differences between the male and female population forecast data (Source 5.3) and the two simulation scenarios SC1 and SC2.

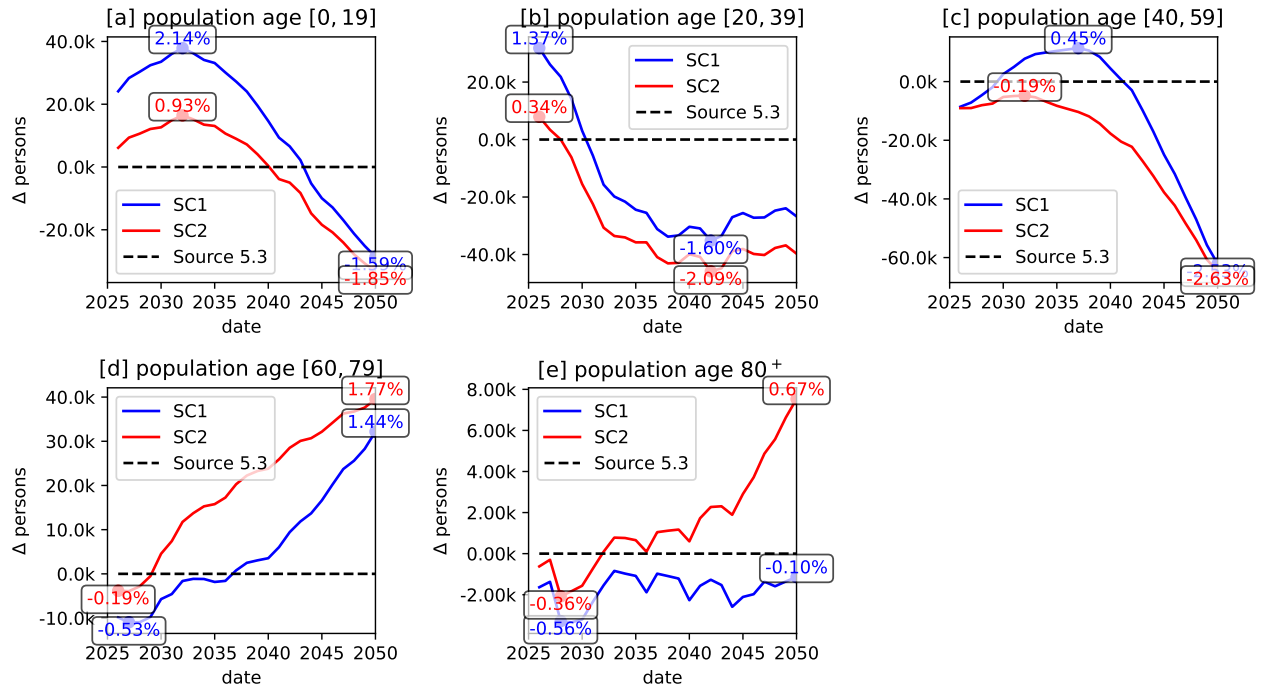


Figure 12: Differences between the population forecast data (Source 5.3) and the two simulation scenarios SC1 and SC2 after aggregation to Austria and 20-year age classes.

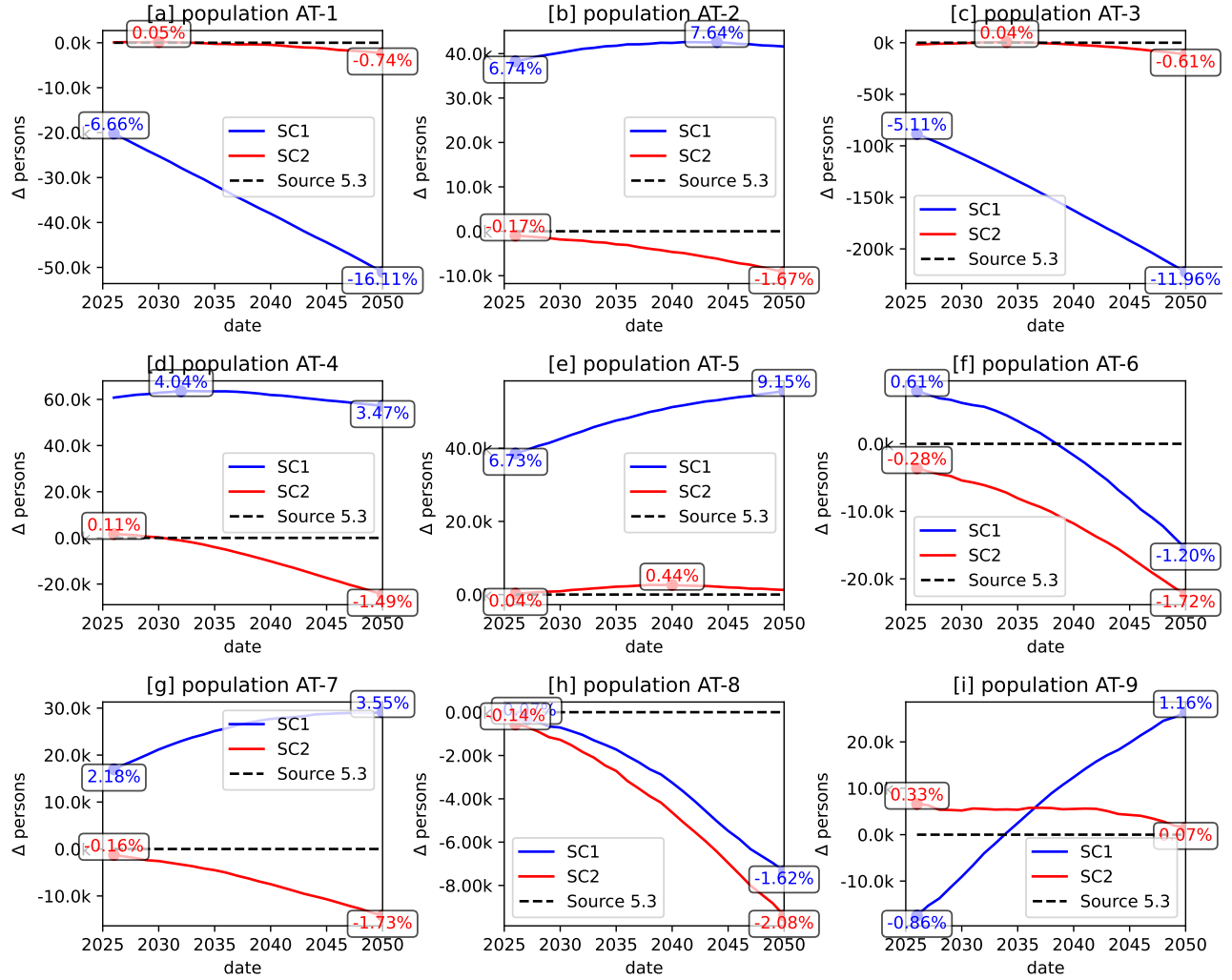


Figure 13: Differences between the population forecast data (Source 5.3) and the two simulation scenarios SC1 and SC2 for all federalstates of Austria.

Table 6: Relative maximum differences between the total population forecast (Source 5.3 and the simulations SC1 and SC2 between 2025 and 2050.

region	sex	age	SC1 (e_{min}, e_{max})		SC2 (e_{min}, e_{max})	
-	-	-	-0.88%	0.39%	-0.92%	0.01%
-	female	-	-0.87%	0.41%	-0.79%	0.05%
-	male	-	-0.89%	0.37%	-1.06%	-0.04%
-	-	[0, 19]	-1.59%	2.14%	-1.85%	0.93%
-	-	[20, 39]	-1.60%	1.37%	-2.09%	0.34%
-	-	[40, 59]	-2.53%	0.45%	-2.63%	-0.19%
-	-	[60, 79]	-0.53%	1.44%	-0.19%	1.77%
-	-	80 ⁺	-0.56%	-0.10%	-0.36%	0.67%
AT-1	-	-	-16.11%	-6.66%	-0.74%	0.05%
AT-2	-	-	6.74%	7.64%	-1.67%	-0.17%
AT-3	-	-	-11.96%	-5.11%	-0.61%	0.04%
AT-4	-	-	3.47%	4.04%	-1.49%	0.11%
AT-5	-	-	6.73%	9.15%	0.04%	0.44%
AT-6	-	-	-1.20%	0.61%	-1.72%	-0.28%
AT-7	-	-	2.18%	3.55%	-1.73%	-0.16%
AT-8	-	-	-1.62%	-0.07%	-2.08%	-0.14%
AT-9	-	-	-0.86%	1.16%	0.07%	0.33%
AT-1	-	[0, 19]	-19.98%	-6.59%	-2.59%	0.56%
AT-2	-	[0, 19]	14.40%	17.09%	-2.38%	0.09%
AT-3	-	[0, 19]	-16.06%	-5.94%	-3.24%	0.46%
AT-4	-	[0, 19]	1.33%	6.68%	-3.63%	0.68%
AT-5	-	[0, 19]	8.06%	10.18%	0.35%	2.66%
AT-6	-	[0, 19]	-0.52%	3.38%	-2.89%	0.35%
AT-7	-	[0, 19]	1.81%	4.25%	-4.86%	-0.14%
AT-8	-	[0, 19]	-4.01%	0.33%	-3.88%	0.34%
AT-9	-	[0, 19]	-1.03%	3.14%	1.21%	3.21%
AT-1	-	[20, 39]	-7.14%	4.34%	-3.87%	-0.44%
AT-2	-	[20, 39]	15.65%	20.71%	-0.77%	0.42%
AT-3	-	[20, 39]	-5.22%	4.01%	-3.44%	-0.29%
AT-4	-	[20, 39]	4.22%	7.05%	-3.20%	-0.14%
AT-5	-	[20, 39]	9.43%	12.58%	-3.16%	0.42%
AT-6	-	[20, 39]	-0.80%	2.36%	-3.61%	0.10%
AT-7	-	[20, 39]	0.99%	3.40%	-2.96%	0.18%
AT-8	-	[20, 39]	-3.48%	-0.51%	-1.78%	-0.16%
AT-9	-	[20, 39]	-10.73%	-9.84%	-0.19%	1.37%
AT-1	-	[40, 59]	-18.26%	-7.06%	-4.42%	0.89%
AT-2	-	[40, 59]	4.71%	9.57%	-2.69%	-0.40%
AT-3	-	[40, 59]	-15.20%	-9.59%	-2.57%	0.27%
AT-4	-	[40, 59]	2.52%	6.29%	-2.85%	0.63%
AT-5	-	[40, 59]	4.90%	11.45%	-1.11%	0.66%
AT-6	-	[40, 59]	-2.75%	1.51%	-3.61%	-0.39%
AT-7	-	[40, 59]	2.05%	4.74%	-2.62%	-0.06%
AT-8	-	[40, 59]	-4.27%	0.07%	-4.16%	-0.05%
AT-9	-	[40, 59]	-0.77%	2.59%	-1.91%	-0.65%
AT-1	-	[60, 79]	-17.06%	-12.91%	-0.05%	4.58%
AT-2	-	[60, 79]	-2.84%	1.53%	-1.31%	-0.20%
AT-3	-	[60, 79]	-11.56%	-6.50%	0.09%	4.36%
AT-4	-	[60, 79]	0.12%	6.26%	0.00%	2.40%
AT-5	-	[60, 79]	2.70%	9.27%	-0.29%	2.24%
AT-6	-	[60, 79]	-2.15%	0.20%	-0.41%	1.47%
AT-7	-	[60, 79]	1.62%	6.09%	-0.33%	1.29%
AT-8	-	[60, 79]	0.76%	2.49%	-0.37%	0.14%
AT-9	-	[60, 79]	7.86%	12.50%	-0.57%	-0.07%
AT-1	-	80 ⁺	-19.53%	-9.04%	0.38%	1.91%
AT-2	-	80 ⁺	-4.50%	-1.41%	-1.27%	-0.01%
AT-3	-	80 ⁺	-11.33%	-6.18%	-0.45%	1.69%
AT-4	-	80 ⁺	-0.26%	0.30%	-0.56%	0.06%
AT-5	-	80 ⁺	0.86%	5.03%	-0.07%	1.64%
AT-6	-	80 ⁺	-3.26%	-1.03%	-0.05%	0.40%
AT-7	-	80 ⁺	1.17%	3.55%	-0.22%	1.28%
AT-8	-	80 ⁺	0.56%	2.82%	-0.62%	0.81%
AT-9	-	80 ⁺	7.34%	19.45%	-2.03%	1.45%

7.2 Births and Deaths

In the next step, we will compare total births and deaths. For both, we utilise Source 5.9.

7.2.1 Comparison with Source 5.9 (Demographische Zeitreihenindikatoren) - Births

Source 5.9 provides a reference for births in the time-frame between 2000 and 2024. Table 1 in Source 5.9 contains the total number of births per sex of the child and per age of the mother on the federal-state level.

Total and sex. As seen in Figure 14, the births of male and female children in scenarios SC1 and SC2 remain within a $\pm 4.7\%$ range around the given data, with the highest deviation for year 2022. In absolute numbers, this corresponds to around 1700 newborn children per year. SC2 performs slightly better than SC1.

Age-classes. Figure 15 shows, how the births correspond to the age of the mother. An unexpectedly large deviation can be observed for the last simulation year 2024, in particular for the 45^+ age class with roughly 30%. This is due to a problem with computing the age-dependent forecast (see Section 6.2): The Gauss distribution used for the forecast is fully symmetric, whereas the actual age-distribution between 2015 and 2024 is skewed to the right (see Figure 3). As a result, the naive birthrates forecast for 2025 is too high for the older age classes and too low for the younger ones. Due to technical reasons (compare with Corollary 3.1), this also influences the final probabilities for 2024.

Federal-states. Table 7 shows summary of the maximum offsets from SC1 and SC2 compared to the total births from Source 5.9. Comparing the federalstates, in particular model SC1 shows high differences due to mentioned role of internal migration for those regions which particularly impacts young families. A figure analogous to the ones for comparing age-classes and sex can be found in the Appendix section (Figure 43).

7.2.2 Comparison with Source 5.9 (Demographische Zeitreihenindikatoren) - Deaths

Source 5.9 provides a reference for deaths in the time-frame between 2000 and 2024. Table 3 in Source 5.9 contains the total number of deaths per sex and age on the federalstate level.

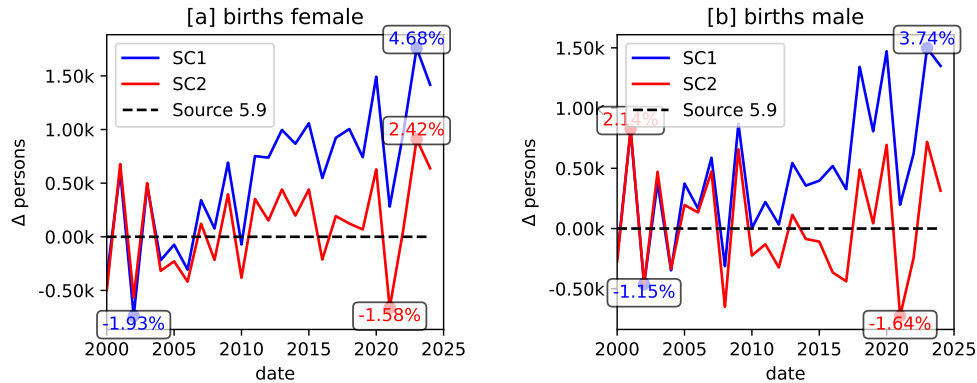


Figure 14: Differences between the male and female births (Source 5.9) and the two simulation scenarios SC1 and SC2.

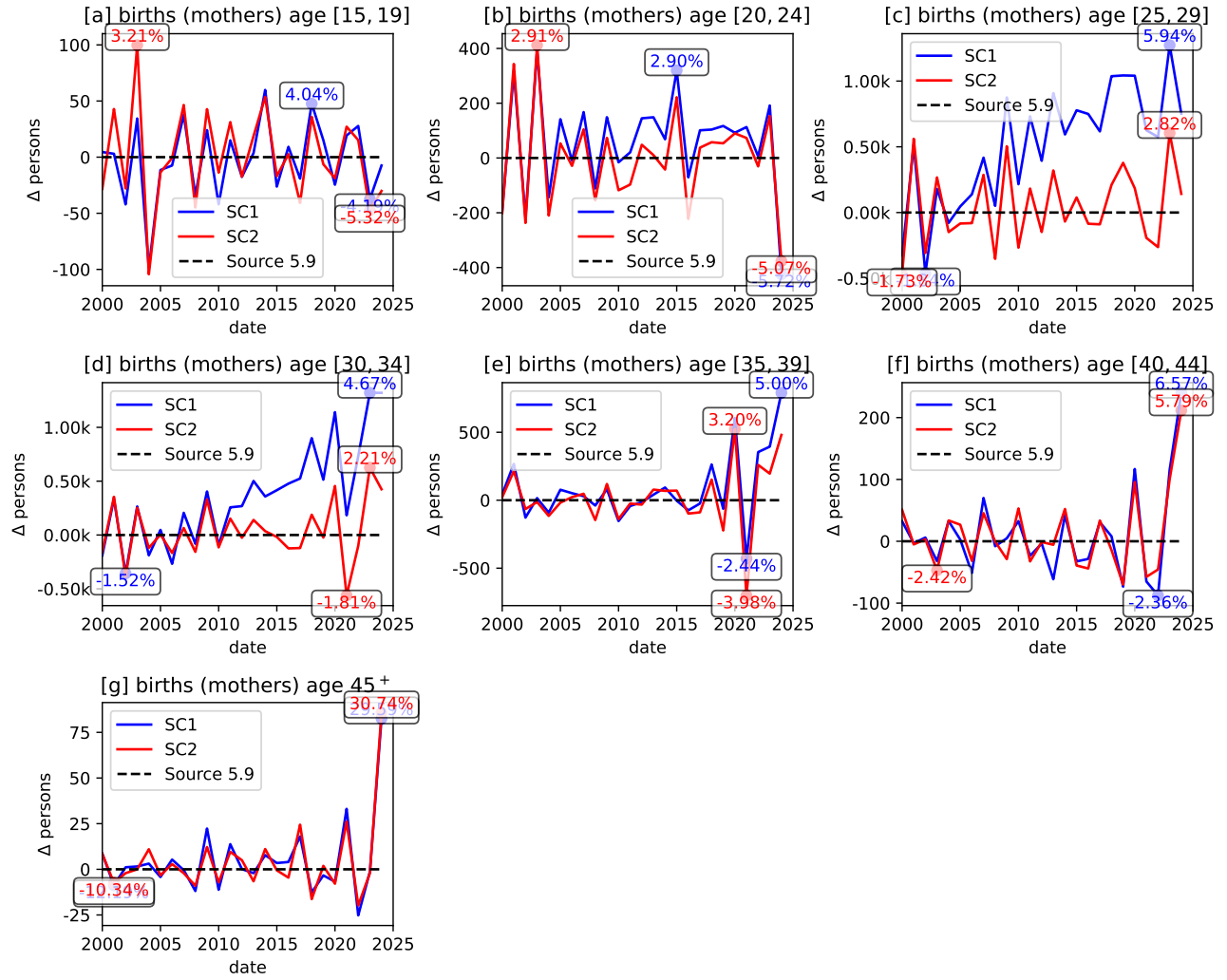


Figure 15: Differences between the total births (Source 5.9) and the two simulation scenarios SC1 and SC2 after aggregation to Austria and 5-year age classes of the mother.

Table 7: Relative maximum differences between the birth data (Source 5.9, Table 1) and the simulations SC1 and SC2 between 2000 and 2024.

region	sex	age (mother)	SC1 (e_{min}, e_{max})		SC2 (e_{min}, e_{max})	
-	-	-	-1.53%	4.20%	-1.61%	2.10%
-	female	-	-1.93%	4.68%	-1.58%	2.42%
-	male	-	-1.15%	3.74%	-1.64%	2.14%
-	-	[15, 19]	-4.19%	4.04%	-5.32%	3.21%
-	-	[20, 24]	-5.72%	2.90%	-5.07%	2.91%
-	-	[25, 29]	-1.84%	5.94%	-1.73%	2.82%
-	-	[30, 34]	-1.52%	4.67%	-1.81%	2.21%
-	-	[35, 39]	-2.44%	5.00%	-3.98%	3.20%
-	-	[40, 44]	-2.36%	6.57%	-2.42%	5.79%
-	-	45 ⁺	-12.19%	29.59%	-10.34%	30.74%
AT-1	-	-	-0.15%	12.50%	-2.58%	7.44%
AT-2	-	-	0.53%	20.65%	-2.98%	2.83%
AT-3	-	-	-2.83%	6.83%	-2.15%	3.02%
AT-4	-	-	-1.15%	11.38%	-1.99%	3.17%
AT-5	-	-	-1.37%	17.10%	-1.49%	7.67%
AT-6	-	-	-1.17%	7.35%	-2.51%	4.23%
AT-7	-	-	-1.50%	5.13%	-2.89%	2.83%
AT-8	-	-	-1.18%	5.66%	-2.20%	3.59%
AT-9	-	-	-13.68%	1.94%	-4.23%	3.56%

Table 8: Relative maximum differences between the death data (Source 5.9, Table 3) and the simulations SC1 and SC2 between 2000 and 2024.

region	sex	age	SC1 (e_{min}, e_{max})		SC2 (e_{min}, e_{max})	
-	-	-	-2.93%	1.98%	-2.77%	2.08%
-	female	-	-3.10%	1.87%	-3.30%	2.05%
-	male	-	-2.74%	2.10%	-2.66%	2.11%
-	-	[0, 19]	-2.74%	5.70%	-1.70%	5.19%
-	-	[20, 39]	-3.54%	4.26%	-3.99%	4.30%
-	-	[40, 59]	-2.28%	2.02%	-2.39%	2.07%
-	-	[60, 79]	-2.24%	2.16%	-1.95%	2.24%
-	-	80 ⁺	-4.13%	1.92%	-3.92%	2.14%
AT-1	-	-	-10.27%	0.92%	-3.26%	3.34%
AT-2	-	-	-5.13%	2.03%	-4.37%	2.76%
AT-3	-	-	-7.38%	0.83%	-2.41%	1.65%
AT-4	-	-	-3.78%	2.65%	-3.94%	2.70%
AT-5	-	-	-4.62%	2.23%	-3.19%	1.75%
AT-6	-	-	-5.72%	1.72%	-5.14%	2.85%
AT-7	-	-	-3.11%	1.98%	-2.64%	1.76%
AT-8	-	-	-3.45%	4.22%	-4.14%	3.53%
AT-9	-	-	-1.35%	9.61%	-3.08%	2.99%

Total and sex. As seen in Figure 16, the male and female deaths in scenarios SC1 and SC2 remain within a $\pm 3.5\%$ range around the given data. In absolute numbers, this corresponds to around 1000 persons per year. No obvious difference between SC1 and SC2 can be found. In general, deaths seem to be slightly underestimated.

Age-classes and federal-states. With respect to age (see Table 8), the models lie within $\pm 6\%$ around the reference data for 20-year age-classes. With respect to federal-states, both models perform similarly well, with the exception of AT-1 and AT-9, where, again, the internal migration effect seems to be important for the validity of the model. Result plots are found in the Appendix (Figures 53 and 51).

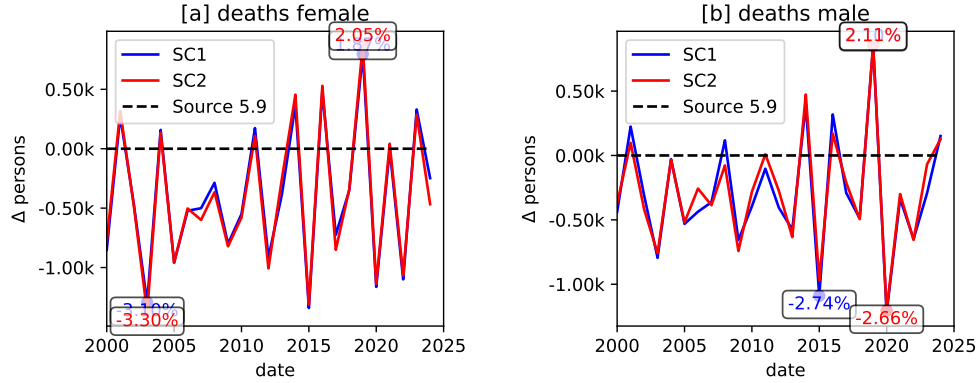


Figure 16: Differences between the male and female deaths (Source 5.9) and the two simulation scenarios SC1 and SC2.

7.2.3 Comparison with Source 5.7 (Bevölkerungsbewegung 1961 bis 2100)

To validate the model beyond 2024, we investigate the migration forecast Source 5.7. The forecast includes a total number of births and deaths per year and federalstate until 2100, without any further age and sex resolution.

Total. Figure 17, panels [a] and [b], show a comparison between the total births and deaths from Source 5.7 and SC1 and SC2. For births and deaths, both models perform roughly equally well and the error is within $\pm 5\%$ around the reference. Births tend to become underestimated, deaths trend towards overestimation.

Federalstates. Focusing on federalstates, the model with internal migration results is more valid with respect to births and deaths than the one without. The highest deviations are found on the end of the simulation period, indicating that certain errors are dynamically accumulating when internal migration is not considered. A summary of the differences is given in Table 9. Figures showing the differences per federal-state are found in the Appendix (Figures 47 and 55).

Table 9: Relative maximum differences between the birth and death forecast (Source 5.7) and the simulations SC1 and SC2 between 2000 and 2049.

region	sex	age	SC1 (e_{min}, e_{max})		SC2 (e_{min}, e_{max})	
births						
-	-	-	-3.15%	2.46%	-4.12%	0.76%
AT-1	-	-	-12.57%	5.98%	-4.41%	0.97%
AT-2	-	-	11.38%	17.07%	-3.22%	0.17%
AT-3	-	-	-10.14%	4.70%	-3.31%	1.75%
AT-4	-	-	2.57%	9.26%	-4.38%	1.95%
AT-5	-	-	5.79%	11.98%	1.87%	4.98%
AT-6	-	-	-1.79%	3.60%	-4.21%	0.60%
AT-7	-	-	-0.43%	4.02%	-5.87%	-0.25%
AT-8	-	-	-6.66%	1.29%	-8.28%	-0.79%
AT-9	-	-	-11.35%	-9.14%	-6.19%	-0.39%
deaths						
-	-	-	-0.41%	4.22%	-0.35%	4.89%
AT-1	-	-	-14.14%	-7.86%	-0.22%	6.23%
AT-2	-	-	-2.09%	1.79%	-1.03%	3.66%
AT-3	-	-	-7.53%	-5.72%	-0.44%	6.29%
AT-4	-	-	-0.32%	5.69%	-0.08%	4.86%
AT-5	-	-	0.88%	11.10%	-1.34%	7.18%
AT-6	-	-	-1.94%	2.62%	-1.19%	5.11%
AT-7	-	-	0.86%	8.87%	-0.06%	5.88%
AT-8	-	-	-0.11%	8.62%	-0.17%	6.53%
AT-9	-	-	7.49%	18.32%	-1.94%	2.88%

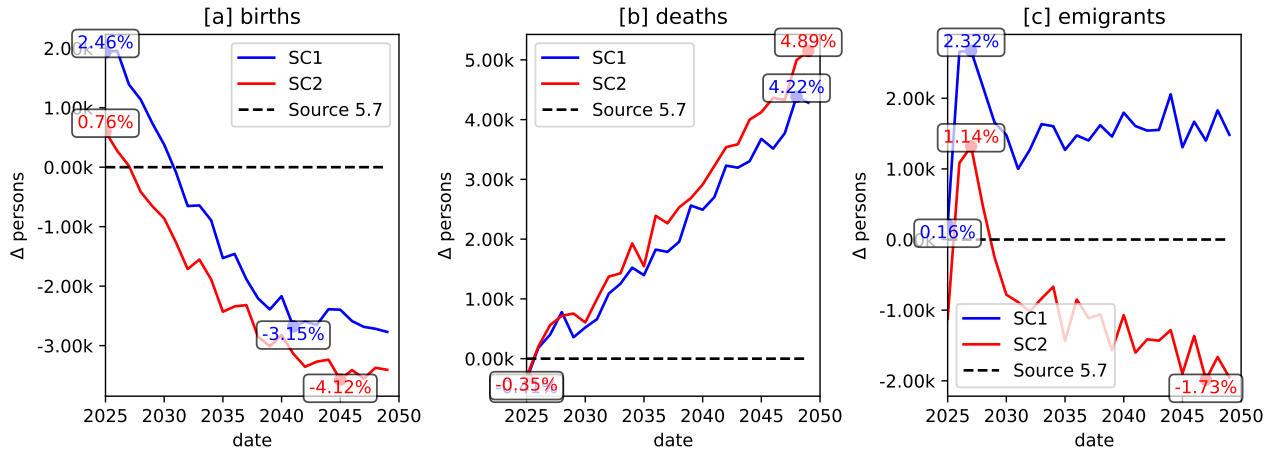


Figure 17: Differences between the total births deaths and emigrants from Source 5.7 and the two simulation scenarios SC1 and SC2.

7.3 Emigrants and Immigrants

With respect to external migration, we only need to validate emigrants. Immigrants are sampled into the model based on processed parameter data, which matches the given raw data on the respective level of aggregation (see calculation of immigration data, Section 6.5). Therefore, alike the initial population, immigrant counts in the model cannot differ from the parameter table (as long as the model’s implementation is correct).

As before, we validate the time period before and after 2024 separately from each other.

7.3.1 Comparison with Source 5.4 (Wanderungen mit dem Ausland ab 2002 nach Alter, Geschlecht und Staatsangehörigkeit)

Source 5.4 contains migration data with sex and single age resolution for the whole of Austria from 2002 to 2024. Hence, we use it to evaluate the simulation results on the country level w.r. sex and age-classes.

Total and sex. As indicated by Figure 18, emigrants is the highest fluctuating quantity in the model with deviations of up to $\pm 8\%$ per sex. Corresponding large errors, however, are only found between 2019 and 2023. Before and after this period, the offsets are in the $\pm 2\%$ region. Having a look at the total numbers, Figure 19 reveals that the comparably high errors originate from a temporary smoothing effect of the model. Hence, the sudden decline of the numbers during the COVID pandemic could not be depicted properly. On average, SC1 seems to slightly underestimate the quantity.

Age-classes. Table 10 indicates that the emigrants for high age cohorts 80^+ stand out with huge errors. The origin of this problem is, again, that emigration itself highly fluctuates between years which are smoothed by the model. Considering the low number of elderly emigrants, we do not consider this problem as very severe.

7.3.2 Comparison with Sources 5.5 and 5.6 (Wanderungen mit dem Ausland von 2002 bis 2014 (ab 2015) nach Altersgruppen, Gemeinde und Staatsangehörigkeit)

Sources 5.5 and 5.6 contain total number of migrants per municipality and five year-age classes between 2002 and 2024. We will use this source for validation w.r. to federalstates. As seen in Table 11, the maximum relative differences are comparably large (up to 15% in AT-1). The large differences are put into perspective when analysing Figure 20, which shows the time-series of the emigrations in the model and in the data for the nine federal-states. The charts show how the target value itself fluctuates and how the model depicts the fluctuations. Even if the behaviour corresponds qualitatively to the data, the model reacts less sensitively and fluctuates with lower amplitude. A plot showing the differences, analogous to the ones displayed before, is found in the Appendix (Figure 59).

Table 10: Relative maximum differences between the emigration data (Source 5.4) and the simulations SC1 and SC2 between 2002 and 2024.

region	sex	age	SC1 (e_{min}, e_{max})		SC2 (e_{min}, e_{max})	
-	-	-	-3.15%	5.05%	-3.75%	5.67%
-	female	-	-5.47%	8.08%	-6.32%	7.66%
-	male	-	-3.27%	5.18%	-2.49%	5.89%
-	-	[0, 19]	-3.21%	16.53%	-6.93%	12.36%
-	-	[20, 39]	-6.51%	3.50%	-2.75%	5.33%
-	-	[40, 59]	-3.81%	9.07%	-6.74%	6.08%
-	-	[60, 79]	-13.39%	11.40%	-12.81%	10.47%
-	-	80^+	-27.05%	46.83%	-27.68%	46.80%

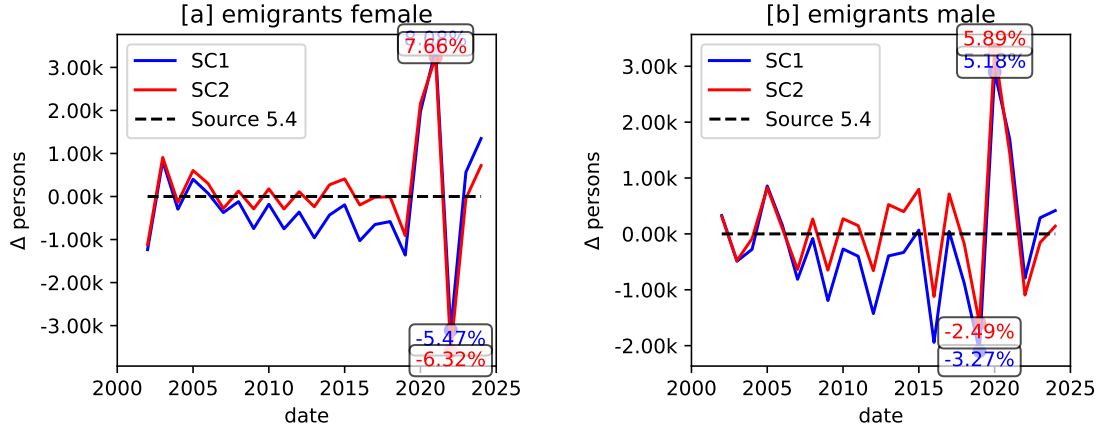


Figure 18: Differences between the male and female emigrants (Source 5.4) and the two simulation scenarios SC1 and SC2.

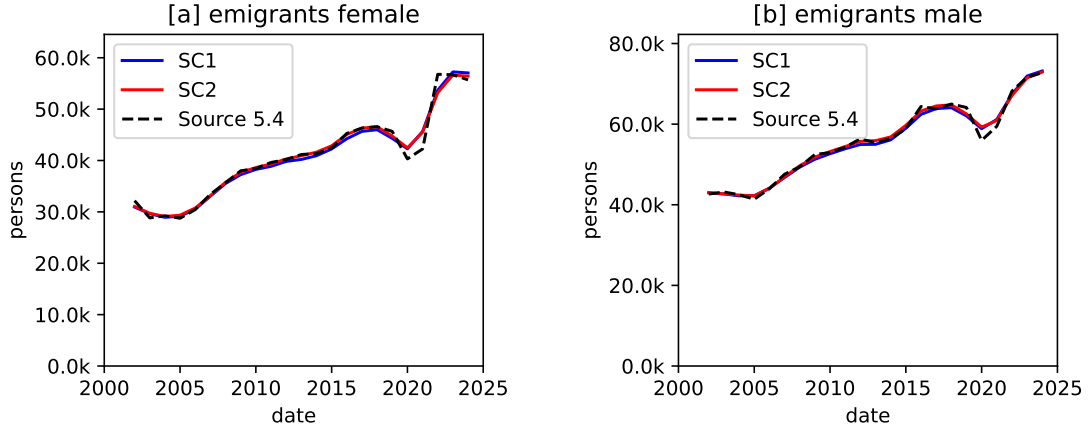


Figure 19: Comparison between emigrant data from Source 5.4 and the simulation scenarios SC1 and SC2 for male and female persons.

Table 11: Relative maximum differences between the emigration data (Sources 5.5 and 5.6) and the simulations SC1 and SC2 between 2002 and 2024.

region	sex	age	SC1 (e_{min}, e_{max})		SC2 (e_{min}, e_{max})	
AT-1	-	-	-9.24%	14.65%	-7.28%	13.65%
AT-2	-	-	-11.56%	15.38%	-13.09%	13.85%
AT-3	-	-	-10.11%	4.33%	-5.67%	9.35%
AT-4	-	-	-2.81%	12.20%	-5.76%	8.84%
AT-5	-	-	-11.03%	11.34%	-11.21%	9.10%
AT-6	-	-	-8.09%	2.44%	-3.65%	6.10%
AT-7	-	-	-6.19%	9.42%	-7.83%	7.51%
AT-8	-	-	-5.17%	6.37%	-5.77%	5.15%
AT-9	-	-	-5.18%	4.55%	-4.32%	5.32%

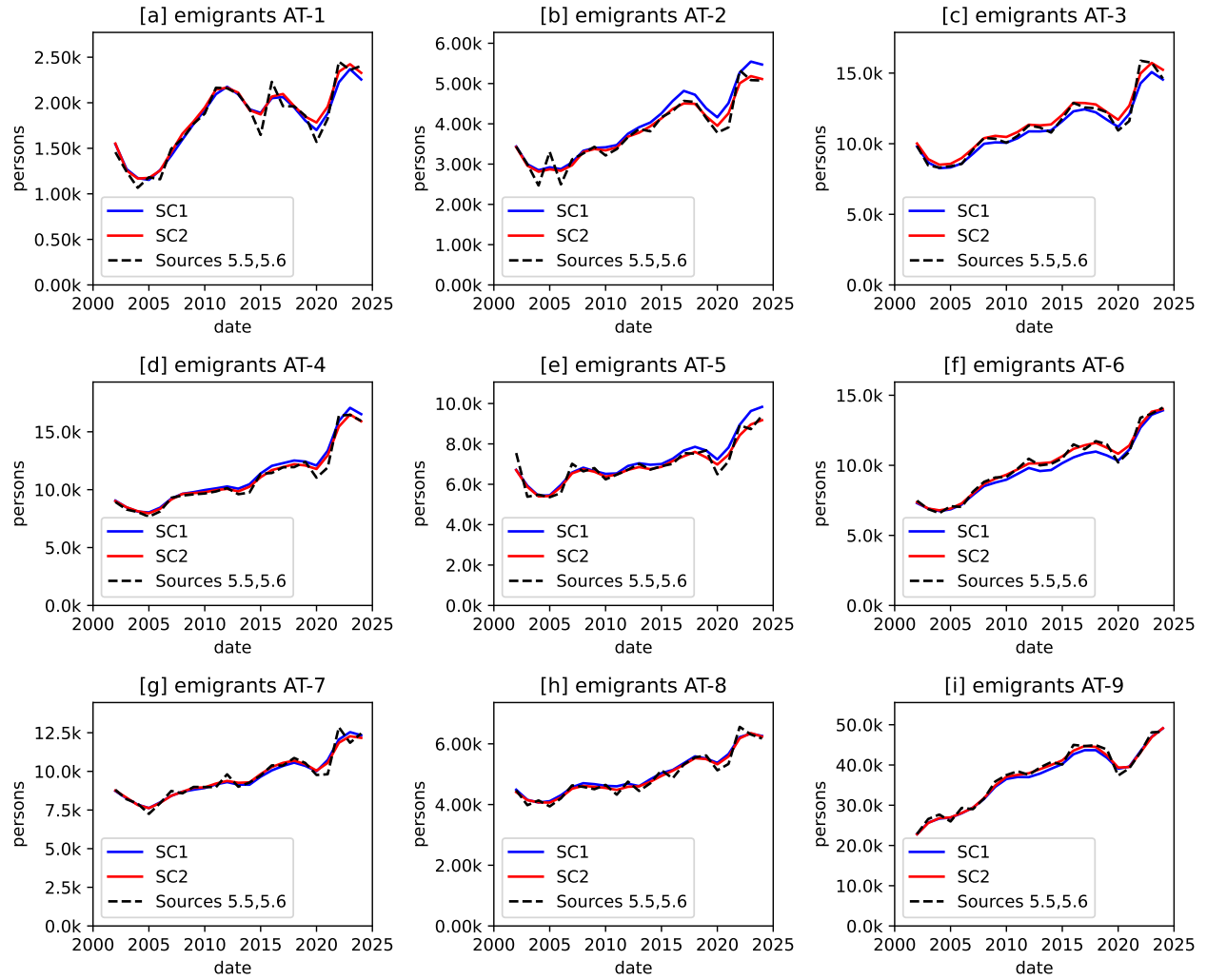


Figure 20: Comparison between the emigrants from Sources 5.5 and 5.6 and the two simulation scenarios SC1 and SC2 for all nine federalstates.

7.3.3 Comparison with Source 5.7 (Bevölkerungsbewegung 1961 bis 2100 nach Bundesland, Bewegungsarten und Szenarien)

To validate the model beyond 2024, we investigate the migration forecast Source 5.7. The forecast includes a total number of emigrants per year and federalstate until 2100 without any further age and sex resolution.

Total. Figure 17 panel [c] shows a comparison between the total emigrants from Source 5.7 and SC1 and SC2. Since the open data forecast does not fluctuate so much anymore, also the differences become more stable for the forecast. Interestingly, the model without internal migration (SC1) shows a slightly different trend than the one with internal migration, which underestimates the data slightly. Up to some extent, SC1 can be considered more accurate here. However, looking at the differences on federal-state level (see below) gives a different picture, so that the comparably good fit of SC1 is estimated to be rather coincidental.

Federal-states. For SC1, numbers fluctuate up to 15% around the data (AT-2,AT-1) in the first few forecasting years (2027), thereafter, the maximum offsets decrease to around 4%. SC2 performs better with differences in the 3% region. A summary of the differences is given in Table 12. Figures showing the differences per federal-state are found in the Appendix (Figure 63).

Table 12: Relative maximum differences between the emigration forecast (Source 5.7) and the simulations SC1 and SC2 between 2025 and 2049.

region	sex	age	SC1 (e_{min}, e_{max})		SC2 (e_{min}, e_{max})	
-	-	-	0.16%	2.32%	-1.73%	1.14%
AT-1	-	-	-12.57%	0.97%	-1.38%	4.78%
AT-2	-	-	8.69%	14.83%	-1.05%	2.54%
AT-3	-	-	-9.69%	-1.87%	-0.01%	3.23%
AT-4	-	-	4.05%	6.34%	-1.40%	2.34%
AT-5	-	-	7.14%	9.84%	-1.76%	0.99%
AT-6	-	-	-0.62%	4.91%	-1.15%	2.30%
AT-7	-	-	2.77%	6.13%	-2.06%	1.29%
AT-8	-	-	-3.70%	0.13%	-1.89%	1.36%
AT-9	-	-	-2.65%	0.74%	-3.29%	0.01%

7.4 Internal Migrants

In the next step, we will evaluate the validity of internal migration processes. Hereby, we will add two additional model scenarios for GEPOC IM using the two other available IM models, i.e. the biregional and the interregional model. The parametrisation is summarised in Table 13.

Table 13: Parametrisation setup for simulation scenarios SC3 and SC4. All other parameters are equal to the SC2 scenario.

		SC3	SC4
internal migration model		biregional	interregional
regional-level for internal migration	r_{ie}	districts_districts	districts_districts
internal emigration probabilities	$\hat{I}E^p$	P.V. 6.14	P.V. 6.14
internal migrants	$\hat{I}I/\hat{O}D$	P.V. 6.15	P.V. 6.13

7.4.1 Comparison with Source 5.1 (Bevölkerungsstand) and 5.3 (Bevölkerung zum Jahresanfang 1952 bis 2101)

Before investigating the impact of the three internal migration models, we first of all investigate if SC3 and SC4 properly depict the population.

Table 14 shows the differences between the model results of SC2-SC4 compared to the population data, Source 5.1 (2000-2025), and the population forecast, Source 5.3 (2025-2050). First of all, all three simulation scenarios can be considered as valid population models, for their maximum deviation from the total population lies around 1%. In general, SC4 shows a slightly different trend than the other two. This is due to the problem that the model does not match the age-distribution of the internal immigrants. As a result, the age-distributions of the total population will start to evolve differently, which, in direct consequence, causes differences in births, deaths and emigrations. The corresponding plots have been left out of the report since they do not provide any new insights.

7.4.2 Comparison with Source 5.15 (BWanderungen innerhalb Österreichs ab 2002)

In the next step, we will evaluate the internal emigrants and immigrants compared to Source 5.15. We will use 20 year age classes and compare the internal emigrants and immigrants between model result and data for all nine federal-states.

Table 15 shows that the offsets from the actual data are very similar to the ones for the external emigration. The offsets vary in the range between -6% to 10% w.r. to sex, age and federal-states. As the number of people is around ten times higher than in the case of external emigration, the percentage deviations are more serious. On the high aggregation level, the largest offset is found for the male internal emigrants with around 6% . As seen in Figure 21, the differences can be considered to be the result of a time-lag effect rather than systematic underestimation. Further result plots are found in the Appendix (Figures 65 to 69).

Comparing the three scenarios SC2 to SC4 with each other, we do not find any larger differences. This was expected, since all three use the same parameters and modelling concept for deciding about internal emigration. We expect this to change when looking into internal immigration.

Table 16 shows the offset-table for the internal immigrants. First of all, the overall deviations for sex and age was left out, since it would be equivalent to the ones from the internal emigration. Even investigating the outcomes on the federal-state level for the destination region, no systematic differences between the model results can

Table 14: Relative maximum differences between the population data and forecast (Sources 5.1 and 5.3) and the simulations SC2-SC4.

region	sex	age	SC2 (e_{min}, e_{max})		SC3 (e_{min}, e_{max})		SC4 (e_{min}, e_{max})	
2000-2024								
-	-	-	-0.12%	0.01%	-0.12%	0.01%	-0.08%	0.45%
-	female	-	-0.13%	0.02%	-0.13%	0.02%	-0.09%	0.47%
-	male	-	-0.12%	0.01%	-0.13%	0.02%	-0.07%	0.43%
-	-	[0, 19]	-0.34%	0.70%	-0.35%	0.71%	0.29%	0.77%
-	-	[20, 39]	-0.09%	0.34%	-0.08%	0.35%	0.04%	1.33%
-	-	[40, 59]	-0.25%	-0.13%	-0.24%	-0.09%	-0.27%	0.18%
-	-	[60, 79]	-0.47%	-0.09%	-0.48%	-0.09%	-0.47%	-0.23%
-	-	80 ⁺	-1.67%	-0.27%	-1.64%	-0.32%	-1.63%	-0.37%
AT-1	-	-	0.05%	0.34%	-0.10%	0.38%	0.08%	0.79%
AT-2	-	-	-0.27%	-0.01%	-0.38%	0.17%	-0.19%	0.44%
AT-3	-	-	-0.09%	0.20%	0.00%	0.23%	-0.04%	0.20%
AT-4	-	-	0.06%	0.25%	-0.18%	0.42%	0.25%	0.61%
AT-5	-	-	-0.01%	0.21%	-0.39%	0.23%	-0.01%	0.31%
AT-6	-	-	-0.46%	-0.25%	-0.42%	0.00%	-0.43%	0.68%
AT-7	-	-	-0.08%	0.37%	-0.38%	0.73%	0.17%	0.42%
AT-8	-	-	-0.15%	0.19%	-0.44%	0.25%	-0.35%	0.18%
AT-9	-	-	-0.47%	-0.10%	-0.72%	0.15%	-0.41%	0.62%
2025-2049								
-	-	-	-0.92%	0.01%	-0.92%	0.01%	0.11%	0.54%
-	female	-	-0.79%	0.05%	-0.83%	0.05%	0.20%	0.58%
-	male	-	-1.06%	-0.04%	-1.00%	-0.04%	0.03%	0.50%
-	-	[0, 19]	-1.85%	0.93%	-1.94%	0.83%	-0.41%	2.14%
-	-	[20, 39]	-2.09%	0.34%	-2.07%	0.37%	-0.81%	1.35%
-	-	[40, 59]	-2.63%	-0.19%	-2.55%	-0.19%	-1.20%	0.79%
-	-	[60, 79]	-0.19%	1.77%	-0.18%	1.78%	-0.39%	2.19%
-	-	80 ⁺	-0.36%	0.67%	-0.39%	0.68%	-0.54%	0.20%
AT-1	-	-	-0.74%	0.05%	-0.56%	0.36%	0.53%	1.69%
AT-2	-	-	-1.67%	-0.17%	-1.94%	-0.72%	-0.06%	0.41%
AT-3	-	-	-0.61%	0.04%	-0.49%	0.31%	-0.19%	0.19%
AT-4	-	-	-1.49%	0.11%	-1.52%	-0.45%	-0.68%	0.52%
AT-5	-	-	0.04%	0.44%	-0.95%	-0.34%	0.12%	0.49%
AT-6	-	-	-1.72%	-0.28%	-1.70%	-0.46%	0.37%	0.76%
AT-7	-	-	-1.73%	-0.16%	-1.46%	-0.23%	-1.11%	0.22%
AT-8	-	-	-2.08%	-0.14%	-1.48%	0.40%	-2.44%	-0.35%
AT-9	-	-	0.07%	0.33%	0.09%	0.95%	1.14%	1.56%

Table 15: Relative maximum differences between the internal emigrants (Source 5.15) and the simulations SC2 to SC4 between 2002 and 2024.

region (origin)	sex	age	SC2 (e_{min}, e_{max})		SC3 (e_{min}, e_{max})		SC4 (e_{min}, e_{max})	
-	-	-	-1.38%	5.24%	-1.49%	5.26%	-1.25%	5.65%
-	female	-	-1.48%	4.68%	-1.64%	4.64%	-1.15%	5.32%
-	male	-	-2.51%	5.74%	-2.51%	5.81%	-2.43%	5.94%
-	-	[0, 19]	-2.68%	6.16%	-2.78%	5.91%	-1.64%	8.32%
-	-	[20, 39]	-0.88%	5.93%	-0.90%	5.97%	-1.36%	5.28%
-	-	[40, 59]	-3.15%	3.08%	-3.34%	3.18%	-2.46%	4.03%
-	-	[60, 79]	-3.83%	9.54%	-3.92%	9.11%	-2.99%	9.14%
-	-	80 ⁺	0.71%	8.85%	0.42%	8.33%	0.41%	8.84%
AT-1	-	-	-2.81%	6.07%	-2.73%	6.17%	1.17%	14.36%
AT-2	-	-	-3.73%	7.89%	-4.04%	7.36%	-1.20%	10.50%
AT-3	-	-	-2.68%	4.48%	-2.71%	4.77%	-1.79%	8.66%
AT-4	-	-	-2.41%	5.79%	-2.36%	5.70%	-2.38%	7.02%
AT-5	-	-	-3.21%	6.15%	-2.89%	5.41%	-2.36%	6.98%
AT-6	-	-	-2.68%	4.97%	-2.62%	4.79%	-2.63%	5.78%
AT-7	-	-	-4.74%	9.48%	-4.61%	9.70%	-4.95%	9.17%
AT-8	-	-	-5.04%	4.73%	-5.44%	5.02%	-5.00%	5.23%
AT-9	-	-	-1.47%	8.16%	-1.42%	8.26%	-4.64%	7.02%

be observed: By aggregation over the origin regions and the age distribution, the potential weaknesses of the interregional and the biregional models do not show. The maximum offsets all vary between $\pm 9\%$ and are very similar to the ones for the internal emigration.

The similarity between the models stops, having a look at the mixed outputs for destination federal-states and age-classes, which are shown in the lower part of Table 16. While it is surprising that the errors for SC3 are smaller than the ones for SC2, in general, SC4 performs by far worse than the other two models.

As an illustrative example, we can point out federal-state AT-1, in which SC4 heavily overestimates for age-class [20, 39] and underestimates for older age-classes (up to -30% off for [60, 79]). This is a direct result of the concept of the interregional model: the age profile of the immigrants into a certain destination region originates from the average of the age-profiles of the internal emigrants of the contributing origin regions, weighted by their contribution to the migration. That means the age profile of the immigrants into AT-1 is primarily caused by the age profile of the emigrants from AT-1 itself (i.e. persons who move within the federal-state) and the ones from AT-3 and AT-9, which, in sum, are responsible for over 75% of all internal immigrations beyond the federal-state border. Figure 22 shows a layover of the internal immigration age-profile of AT-1 and the age-profiles of internal emigrants of AT-1, AT-3 and AT-9. It is clearly visible that the emigration age-profiles are all much “younger” than the immigration age-profile, in particular the one from Vienna (AT-9). Hence, the interregional model can never properly depict the real situation.

The real situation is more complex and requires to make destination regions age-dependent: Primarily older internal emigrants from AT-3 and AT-9 tend to move to AT-1, whereas younger ones usually move within AT-3, AT-9 or someplace else. From the perspective of federal-state AT-1, young internal emigrants tend to leave the region, whereas older ones are more likely to stay.

In summary, the interregional model shows expected weaknesses when it comes to depicting the correct age-resolution of internal immigrants into certain regions, since the age resolution is ignored when choosing a new destination region. In this light, it is up to some extent even surprising that the results for the other quantities (population, births, etc.) are still within a reasonable error margin. However, since a correct regional age distribution is relevant for all other model processes, SC4 leads to the largest errors to the reference data, not only for the population but in general for all other dimensions (deaths, births, etc.).

7.4.3 Comparison with Source 5.14 (Wanderungen innerhalb Österreichs ab 2002)

Finally, we investigate the origin-destination behaviour of the three internal-migration models. Table 17 shows the maximum offsets of the model results to the data. It is well visible that the biregional model (SC3) does not

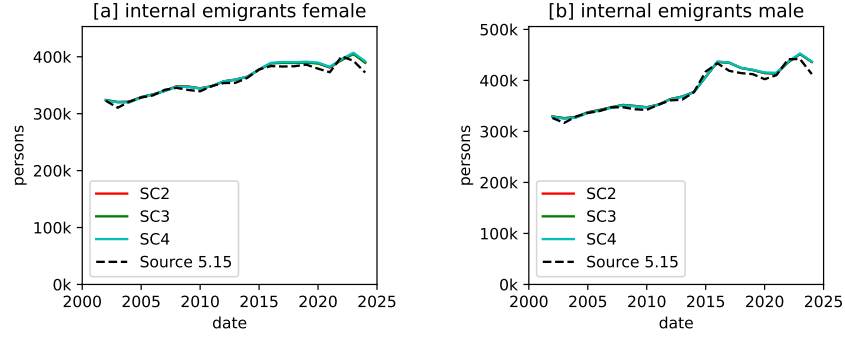


Figure 21: Comparison between internal emigrants data from Source 5.15 and the simulation scenarios SC2-SC4 for male and female persons.

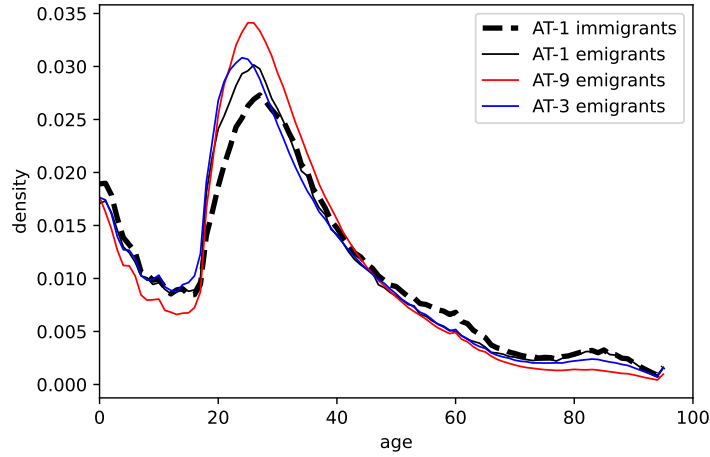


Figure 22: Comparison between the age profiles of internal immigrants into AT-1 and the internal emigrants of the most important origin regions AT-1, AT-3 and AT-9. All data is from Source 5.15 for the years between 2002 and 2024 and normed to an area of one below the curve.

Table 16: Relative maximum differences between the internal immigrants (Source 5.15) and the simulations SC2 to SC4 between 2002 and 2024.

region (destination)	sex	age	SC2 (e_{min}, e_{max})		SC3 (e_{min}, e_{max})		SC4 (e_{min}, e_{max})	
AT-1	-	-	-2.43%	5.86%	-1.98%	5.13%	1.06%	11.32%
AT-2	-	-	-3.05%	7.15%	-1.44%	5.04%	-0.88%	9.52%
AT-3	-	-	-2.77%	4.57%	-1.67%	5.13%	-0.64%	7.34%
AT-4	-	-	-2.30%	5.87%	-1.41%	5.24%	-2.20%	7.01%
AT-5	-	-	-3.15%	5.88%	-1.75%	5.31%	-2.26%	6.78%
AT-6	-	-	-2.38%	4.96%	-1.44%	5.26%	-2.23%	5.88%
AT-7	-	-	-4.46%	9.06%	-1.46%	5.28%	-4.68%	8.77%
AT-8	-	-	-4.68%	4.39%	-1.46%	5.07%	-4.76%	4.88%
AT-9	-	-	-1.41%	6.71%	-1.57%	5.39%	-3.68%	5.66%
AT-1	-	[0, 19]	-5.37%	6.49%	-3.33%	5.51%	-6.10%	6.73%
AT-2	-	[0, 19]	-4.33%	12.96%	-2.89%	5.22%	-2.73%	15.79%
AT-3	-	[0, 19]	-3.89%	6.05%	-2.65%	6.04%	-6.80%	5.36%
AT-4	-	[0, 19]	-3.36%	5.83%	-2.68%	5.84%	-4.93%	7.12%
AT-5	-	[0, 19]	-4.26%	7.72%	-3.73%	6.33%	-1.14%	14.22%
AT-6	-	[0, 19]	-2.88%	6.93%	-2.78%	5.80%	-2.83%	11.38%
AT-7	-	[0, 19]	-5.41%	9.24%	-3.13%	6.32%	-4.98%	10.71%
AT-8	-	[0, 19]	-5.25%	4.94%	-2.38%	6.25%	-6.87%	5.23%
AT-9	-	[0, 19]	-3.24%	6.06%	-3.08%	5.91%	1.74%	10.77%
AT-1	-	[20, 39]	-1.38%	7.43%	-1.17%	6.08%	7.31%	36.76%
AT-2	-	[20, 39]	-2.75%	7.20%	-0.79%	6.11%	2.73%	14.36%
AT-3	-	[20, 39]	-1.37%	5.28%	-1.05%	5.94%	3.53%	18.18%
AT-4	-	[20, 39]	-1.52%	6.99%	-0.89%	5.99%	0.16%	9.50%
AT-5	-	[20, 39]	-3.60%	6.35%	-0.95%	5.83%	-2.93%	6.29%
AT-6	-	[20, 39]	-2.19%	6.10%	-0.57%	6.11%	-2.30%	6.41%
AT-7	-	[20, 39]	-3.14%	6.13%	-0.84%	5.94%	-3.92%	4.56%
AT-8	-	[20, 39]	-4.52%	4.86%	-1.00%	5.64%	-4.08%	5.46%
AT-9	-	[20, 39]	-1.01%	7.16%	-1.00%	5.98%	-11.66%	-0.08%
AT-1	-	[40, 59]	-4.65%	4.95%	-3.51%	3.63%	-15.46%	2.55%
AT-2	-	[40, 59]	-4.52%	5.53%	-3.57%	3.38%	-8.39%	1.82%
AT-3	-	[40, 59]	-4.51%	2.38%	-3.39%	3.02%	-10.54%	1.42%
AT-4	-	[40, 59]	-2.54%	4.88%	-3.21%	2.76%	-3.73%	4.05%
AT-5	-	[40, 59]	-3.00%	3.20%	-2.90%	3.25%	-4.92%	2.34%
AT-6	-	[40, 59]	-3.61%	3.23%	-4.05%	3.18%	-2.98%	3.89%
AT-7	-	[40, 59]	-5.29%	12.30%	-2.95%	3.69%	-4.73%	13.39%
AT-8	-	[40, 59]	-4.91%	3.27%	-2.39%	3.95%	-4.81%	4.71%
AT-9	-	[40, 59]	-3.04%	6.09%	-3.33%	3.29%	1.85%	13.00%
AT-1	-	[60, 79]	-4.20%	7.52%	-4.78%	13.34%	-30.92%	-7.46%
AT-2	-	[60, 79]	-4.64%	17.22%	-3.19%	9.00%	-17.06%	11.63%
AT-3	-	[60, 79]	-6.23%	6.38%	-4.30%	9.79%	-18.34%	-0.50%
AT-4	-	[60, 79]	-7.50%	14.00%	-3.43%	8.96%	-10.48%	11.51%
AT-5	-	[60, 79]	-3.80%	8.89%	-3.89%	9.12%	-6.49%	6.93%
AT-6	-	[60, 79]	-5.00%	6.22%	-3.71%	8.94%	-7.76%	4.80%
AT-7	-	[60, 79]	-12.21%	32.33%	-4.09%	9.04%	-12.57%	33.91%
AT-8	-	[60, 79]	-8.74%	13.15%	-4.23%	6.19%	-8.01%	16.67%
AT-9	-	[60, 79]	-5.13%	5.72%	-4.00%	8.76%	11.89%	34.23%
AT-1	-	80+	0.28%	13.92%	-0.50%	10.64%	-20.93%	6.60%
AT-2	-	80+	-4.41%	17.71%	-0.62%	9.48%	-10.68%	11.44%
AT-3	-	80+	1.63%	9.13%	-0.27%	8.11%	-13.84%	-3.08%
AT-4	-	80+	-2.24%	11.01%	-0.42%	8.06%	-6.36%	8.44%
AT-5	-	80+	-2.32%	13.88%	0.51%	8.97%	-8.42%	11.35%
AT-6	-	80+	-0.48%	11.88%	-0.39%	8.29%	-3.24%	9.80%
AT-7	-	80+	-1.59%	14.58%	0.23%	10.77%	-2.55%	12.08%
AT-8	-	80+	-3.29%	16.04%	-0.15%	10.26%	-2.24%	16.78%
AT-9	-	80+	0.72%	14.00%	2.33%	10.75%	28.05%	55.46%

at all represent a valid behaviour in this regard. The maximum differences are beyond any reasonable margin. As a result, the model should only be used if the correct representation of regional flows are irrelevant. The other two models nicely depict the flows between the federal-states with rare outliers up to $\pm 20\%$ deviation (e.g. between AT-1 and AT-8), which can be explained by small numbers of migrants. Figure 23 shows the total number of internal migrants between the regions between 2002 and 2024. It is easily seen that the internal migrants within the same region are the most relevant ones w.r. to absolute numbers. Flows between the eastern federalstates AT-3, AT-9, AT-4 and AT-6 are also by several margins larger than flows between western federalstates. It is nicely seen in panel [c] with the biregional model that all rows and all columns of the origin-destination matrix are a multiple of each other. This is a result of the concept that the bidirectional model samples all destinations independent of the origin regions.

Figure 24 puts the numbers of SC2 and SC4 in relation with Source 5.14 (SC3 was left out, since it does not validly represent the situation). Both panels show the mentioned systematic underestimation of the internal emigrants (discussed before). Part [b] of the figure reveals bias with respect to the internal migrants from Vienna (AT-9) and Burgenland (AT-1) by the interregional model SC4. We estimate that this is a result of the problem of the interregional model to validly depict the dynamics of the age structure of the population, in particular in large cities.

7.4.4 Comparison with Source 5.7 (Bevölkerungsbewegung 1961 bis 2100)

For the forecast, we compare with Source 5.7, which contains the absolute number of internal immigrants and emigrants per federal-state. Since the numbers exclusively show migration beyond the federal-state border, we need to filter the simulation scenarios accordingly.

Table 18 shows the maximum positive and negative relative differences between the simulation scenarios and the forecasts. Clearly, SC3 cannot be validated with this strategy, since the biregional model does not preserve the origin-destination flows (which includes the flow from one region into itself).

Table 17: Relative maximum differences between the origin destination data (Source 5.14) and the simulations SC2 to SC4 between 2002 and 2024.

region (origin)	sex	region (destination)	SC2 (e_{min}, e_{max})	SC3 (e_{min}, e_{max})	SC4 (e_{min}, e_{max})
AT-1	-	AT-1	-2.53%	-96.60%	1.26%
AT-1	-	AT-2	-7.41%	673.09%	0.29%
AT-1	-	AT-3	-4.02%	72.18%	0.24%
AT-1	-	AT-4	-4.30%	805.26%	-3.90%
AT-1	-	AT-5	-17.92%	965.59%	-6.45%
AT-1	-	AT-6	-4.35%	146.76%	0.16%
AT-1	-	AT-7	-10.75%	832.37%	-3.43%
AT-1	-	AT-8	-14.81%	484.18%	-3.03%
AT-1	-	AT-9	-3.92%	64.78%	0.78%
AT-2	-	AT-1	-13.84%	384.06%	-4.94%
AT-2	-	AT-2	-3.33%	-93.07%	-0.76%
AT-2	-	AT-3	-4.61%	431.26%	-3.75%
AT-2	-	AT-4	-7.30%	355.64%	-5.01%
AT-2	-	AT-5	-5.62%	314.73%	-3.01%
AT-2	-	AT-6	-5.16%	142.32%	-2.89%
AT-2	-	AT-7	-6.00%	319.00%	-1.56%
AT-2	-	AT-8	-6.80%	358.03%	-5.49%
AT-2	-	AT-9	-4.35%	382.76%	-2.87%
AT-3	-	AT-1	-2.14%	24.81%	-0.72%
AT-3	-	AT-2	-4.23%	530.81%	-2.63%
AT-3	-	AT-3	-3.05%	-79.73%	-1.78%
AT-3	-	AT-4	-2.61%	394.47%	-2.81%
AT-3	-	AT-5	-4.71%	776.94%	-2.94%
AT-3	-	AT-6	-2.27%	634.89%	-0.91%
AT-3	-	AT-7	-4.04%	1153.56%	-2.36%
AT-3	-	AT-8	-7.67%	823.94%	-3.43%
AT-3	-	AT-9	-1.92%	70.61%	-2.01%
AT-4	-	AT-1	-6.28%	703.62%	-5.93%
AT-4	-	AT-2	-5.66%	704.75%	-2.82%
AT-4	-	AT-3	-3.37%	477.78%	-3.16%
AT-4	-	AT-4	-2.50%	-82.63%	-2.38%
AT-4	-	AT-5	-3.01%	139.69%	-2.01%
AT-4	-	AT-6	-3.26%	636.90%	-1.80%
AT-4	-	AT-7	-4.93%	523.74%	-3.11%
AT-4	-	AT-8	-5.53%	995.88%	-4.72%
AT-4	-	AT-9	-1.65%	448.04%	-2.26%
AT-5	-	AT-1	-4.93%	395.21%	-5.56%
AT-5	-	AT-2	-5.15%	171.12%	-4.37%
AT-5	-	AT-3	-4.83%	745.28%	-5.17%
AT-5	-	AT-4	-3.82%	70.81%	-3.30%
AT-5	-	AT-5	-3.27%	-92.89%	-2.30%
AT-5	-	AT-6	-3.33%	352.83%	-2.40%
AT-5	-	AT-7	-4.18%	205.77%	-2.56%
AT-5	-	AT-8	-4.63%	624.34%	-3.46%
AT-5	-	AT-9	-3.09%	418.89%	-2.55%
AT-6	-	AT-1	-4.83%	143.06%	-3.67%
AT-6	-	AT-2	-3.96%	140.58%	-4.20%
AT-6	-	AT-3	-5.98%	776.42%	-5.38%
AT-6	-	AT-4	-6.69%	590.99%	-7.57%
AT-6	-	AT-5	-3.36%	481.71%	-5.82%
AT-6	-	AT-6	-2.44%	-84.51%	-2.34%
AT-6	-	AT-7	-4.22%	1047.38%	-5.47%
AT-6	-	AT-8	-5.09%	998.91%	-6.73%
AT-6	-	AT-9	-3.31%	518.10%	-3.76%
AT-7	-	AT-1	-7.26%	580.98%	-6.07%
AT-7	-	AT-2	-3.96%	450.64%	-3.57%
AT-7	-	AT-3	-5.15%	1319.39%	-5.21%
AT-7	-	AT-4	-3.68%	822.28%	-4.16%
AT-7	-	AT-5	-2.75%	345.72%	-3.75%
AT-7	-	AT-6	-6.84%	771.46%	-5.24%
AT-7	-	AT-7	-4.79%	-91.43%	-5.03%
AT-7	-	AT-8	-3.09%	169.65%	-4.06%
AT-7	-	AT-9	-4.21%	742.74%	-2.85%
AT-8	-	AT-1	-8.59%	911.23%	-9.09%
AT-8	-	AT-2	-10.08%	1155.17%	-7.00%
AT-8	-	AT-3	-4.12%	2242.18%	-4.40%
AT-8	-	AT-4	-10.22%	1679.42%	-9.55%
AT-8	-	AT-5	-4.54%	1178.51%	-6.11%
AT-8	-	AT-6	-6.10%	1247.75%	-6.23%
AT-8	-	AT-7	-4.94%	198.49%	-4.11%
AT-8	-	AT-8	-5.04%	-95.20%	-5.00%
AT-8	-	AT-9	-3.44%	626.82%	-3.11%
AT-9	-	AT-1	-1.30%	75.29%	-4.44%
AT-9	-	AT-2	-2.89%	746.32%	-4.94%
AT-9	-	AT-3	-1.69%	39.89%	-4.59%
AT-9	-	AT-4	-2.17%	877.14%	-5.05%
AT-9	-	AT-5	-0.74%	807.26%	-5.56%
AT-9	-	AT-6	-1.85%	847.95%	-5.36%
AT-9	-	AT-7	-1.82%	1281.58%	-5.23%
AT-9	-	AT-8	-2.53%	740.15%	-6.50%
AT-9	-	AT-9	-1.46%	-70.14%	-4.65%

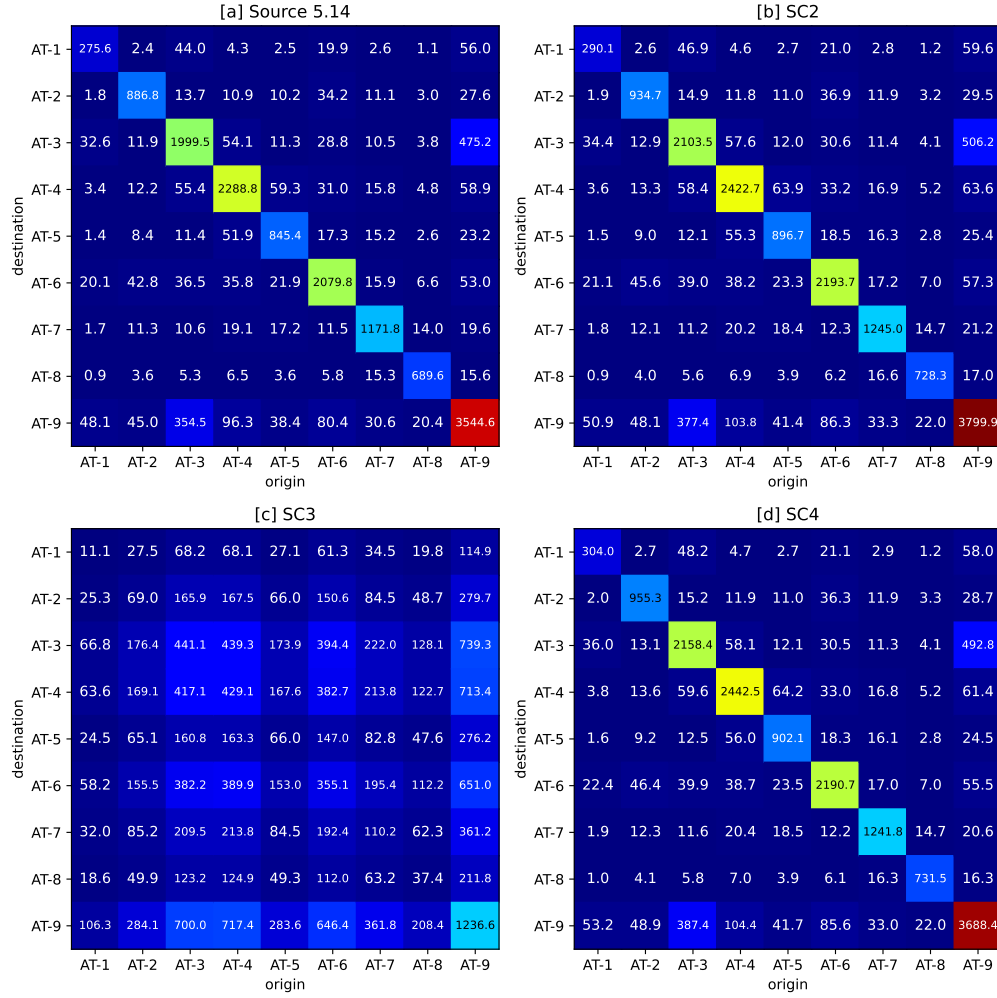


Figure 23: Comparison between origin-destination flow data between 2002 and 2024 (Source 5.14) and the simulation scenarios SC2-SC4. The numbers are in units of thousand migrants over the whole period.

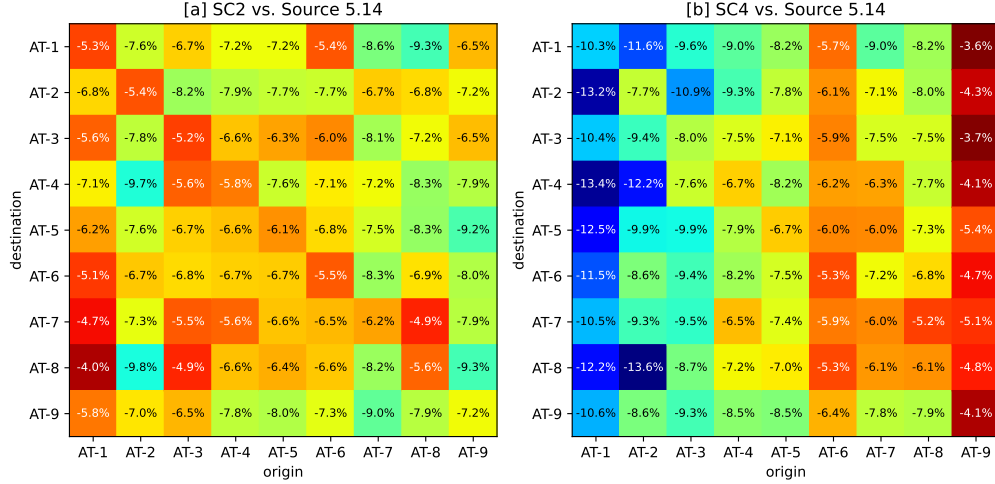


Figure 24: Differences between the total internal migrants between 2002 and 2024 (Source 5.14) and the simulation scenarios SC2 and SC4.

Table 18: Relative maximum differences between the internal migration forecasts (Source 5.7) and the simulations SC2 to SC4 between 2025 and 2049. Only internal migrants into different regions are counted.

region (origin)	sex	region (destination)	SC2 (e_{min}, e_{max})		SC3 (e_{min}, e_{max})		SC4 (e_{min}, e_{max})	
-	-	-	-0.50%	6.43%	383.39%	428.51%	0.66%	7.67%
AT-1	-	-	-2.85%	2.96%	215.92%	230.58%	7.42%	11.63%
AT-2	-	-	-0.27%	21.02%	424.57%	539.93%	3.67%	26.25%
AT-3	-	-	-1.42%	3.07%	246.24%	265.24%	3.63%	7.64%
AT-4	-	-	-1.29%	11.04%	565.43%	657.30%	0.00%	13.71%
AT-5	-	-	-3.30%	13.53%	314.26%	395.55%	-1.00%	16.14%
AT-6	-	-	-1.85%	12.33%	602.45%	699.32%	-0.02%	13.10%
AT-7	-	-	-1.22%	6.79%	734.73%	787.40%	-1.15%	6.27%
AT-8	-	-	-1.41%	2.65%	1060.59%	1103.24%	-2.43%	2.04%
AT-9	-	-	0.48%	2.81%	286.22%	303.82%	-3.15%	-0.02%
AT-1	-	-	0.08%	5.09%	160.44%	191.19%	0.51%	5.31%
AT-2	-	-	-2.66%	7.48%	411.59%	497.17%	-1.64%	6.98%
AT-3	-	-	0.79%	4.53%	193.21%	228.69%	-1.02%	3.06%
AT-4	-	-	-3.60%	5.97%	594.18%	633.66%	-2.18%	7.73%
AT-5	-	-	-2.27%	8.52%	374.30%	419.50%	-1.19%	9.03%
AT-6	-	-	-2.00%	7.43%	572.75%	626.84%	-0.07%	9.83%
AT-7	-	-	-3.13%	7.62%	929.89%	1039.90%	-1.87%	8.12%
AT-8	-	-	-1.97%	7.43%	1056.88%	1138.04%	-2.15%	7.34%
AT-9	-	-	-0.08%	7.50%	336.21%	373.57%	3.76%	11.56%

7.5 Additional Validation Scenarios

In this final validation step, we will define several scenarios to evaluate very specific features about GEPOC ABM IM, using SC2 as reference scenario. First of all, SC5 will use [2000,2100] as simulation time-frame. This will increase the simulation duration from 50 to 100 years and we expect larger errors to the validation data. SC6 will be run with monthly steps (instead of yearly). We expect that, due to the concept with planning for birth-day to birth-day, the influence of the time-step parameter is minimal. Finally, SC7 will use a scale factor $\sigma = 0.1$ instead of 1. We expect that the (up-scaled) mean values still matches with the data. The scenarios are summarised in Table 19.

Table 19: Parametrisation setup for simulation scenarios SC5 - SC7. All other parameters are equal to the SC2 scenario.

		SC2	SC5	SC6	SC7
sim start-time	t_0	2000-01-01	2000-01-01	2000-01-01	2000-01-01
time-step lengths		years	years	months	years
sim end-time	t_e	2050-01-01	2100-01-01	2050-01-01	2050-01-01
scale	σ	1.0	1.0	1.0	0.1
Monthe-Carlo runs		9	9	9	27

7.5.1 Simulation Scenario SC5 (Long Run).

The long simulation scenario between 2003 and 2100 shows a reasonably good fit with the population data until about year 2060. Thereafter, the simulated numbers underestimate the Statistics Austria forecast from Source 5.3 (see Figure 25). In 2100, the male and female population lie around 5% below the reference. The underestimation originates from an underestimated net-balance between births, deaths and emigrants between 2030 and 2100. As seen in Figure 26, births are underestimated by up to 10% (around 8500 births per year), deaths are overestimated by up to 11% (about 10000 deaths in 2080), and emigrants are underestimated by up to 6% (about 8000 emigrants in 2100). While deviations from deaths and emigrants roughly cancel out, the underestimated number of births causes the population to continuously drop below the reference. The origin of this problem lies in the assumptions made for the computation of the forecast from the low-resolution data.

7.5.2 Simulation Scenarios SC6 (Monthly Steps) and SC7 (Scale 0.1).

For both simulation scenarios, SC6 and SC7, we expect that the numbers only deviate minimally from SC2, if aggregated/scaled accordingly. That means, births, deaths, etc. must be aggregated from months to years in SC6, and numbers have to be scaled up by $1/\sigma = 10$ in SC7. Figure 27 shows the differences between SC2, SC6 and SC7 to the reference data and forecast for the population for the nine federal-states of Austria. The upscaled results from scenario SC7 hardly differ from SC2. Since the base version of GEPOC ABM does not involve interaction between the agents, i.e. the agents can be expressed as independent stochastic processes, this behaviour was expected by direct application of the Law of Large Numbers. Scenario SC6, which uses different time-step lengths, shows small differences to SC2. This is a result of how newly created agents are added in GEPOC. Both immigrants and newborns are added to the dynamics of the model at the end of the discrete timesteps, independent of their scheduled immigration-/birth-date. Given the results shown in Figure 27, there is no clear answer to the question of whether reduced time-step lengths improve or harm the validity of the results. However, this confirms the idea that the quality of the GEPOC ABM (IM) simulations does not (much) depend on the step-lengths.

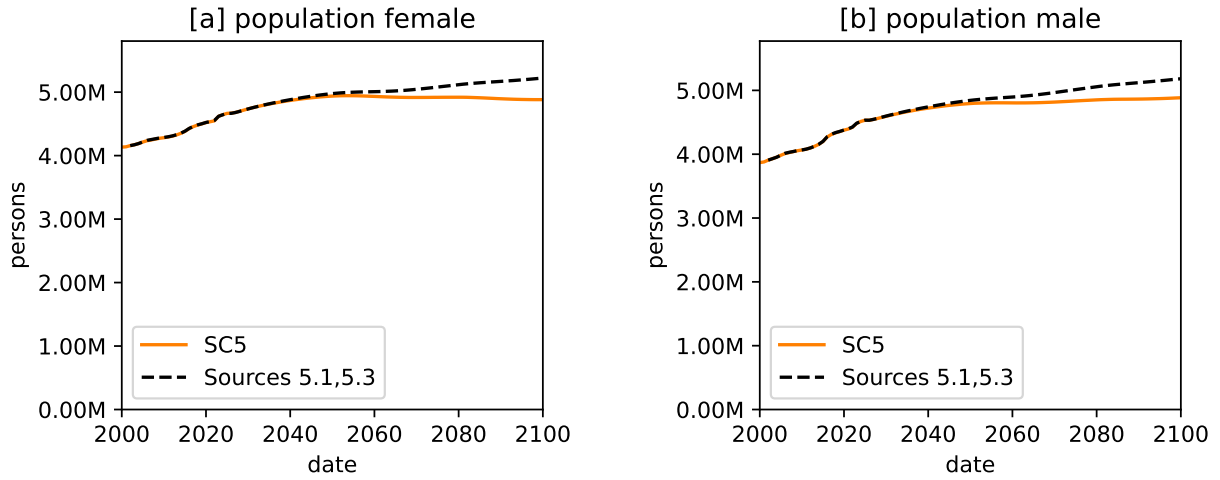


Figure 25: Comparison between the population data and forecast (Sources 5.1 and 5.3) and the simulation scenario SC5.

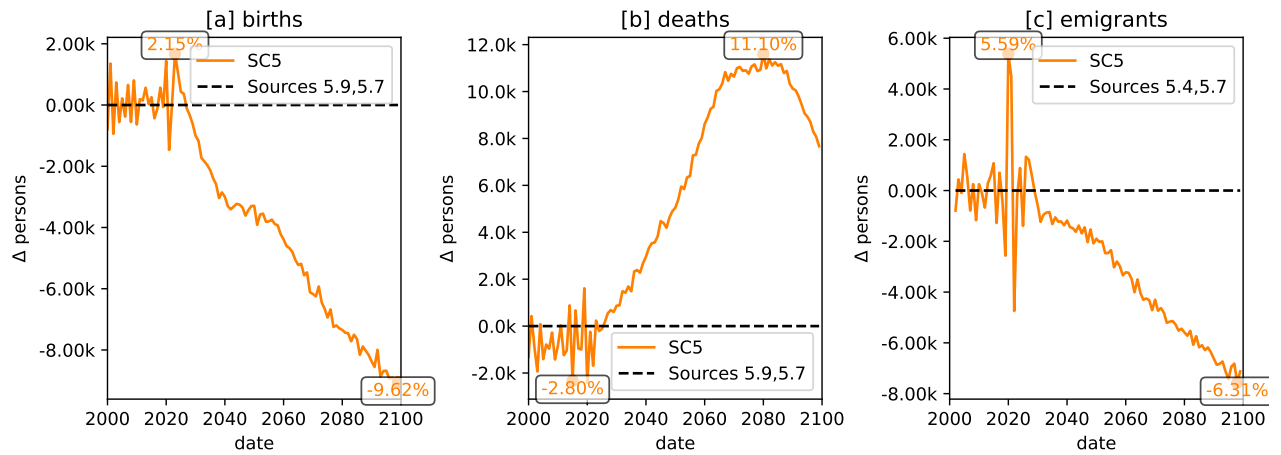


Figure 26: Differences between the different data and forecasts for births, deaths and emigrants and the corresponding values in simulation scenario SC5.

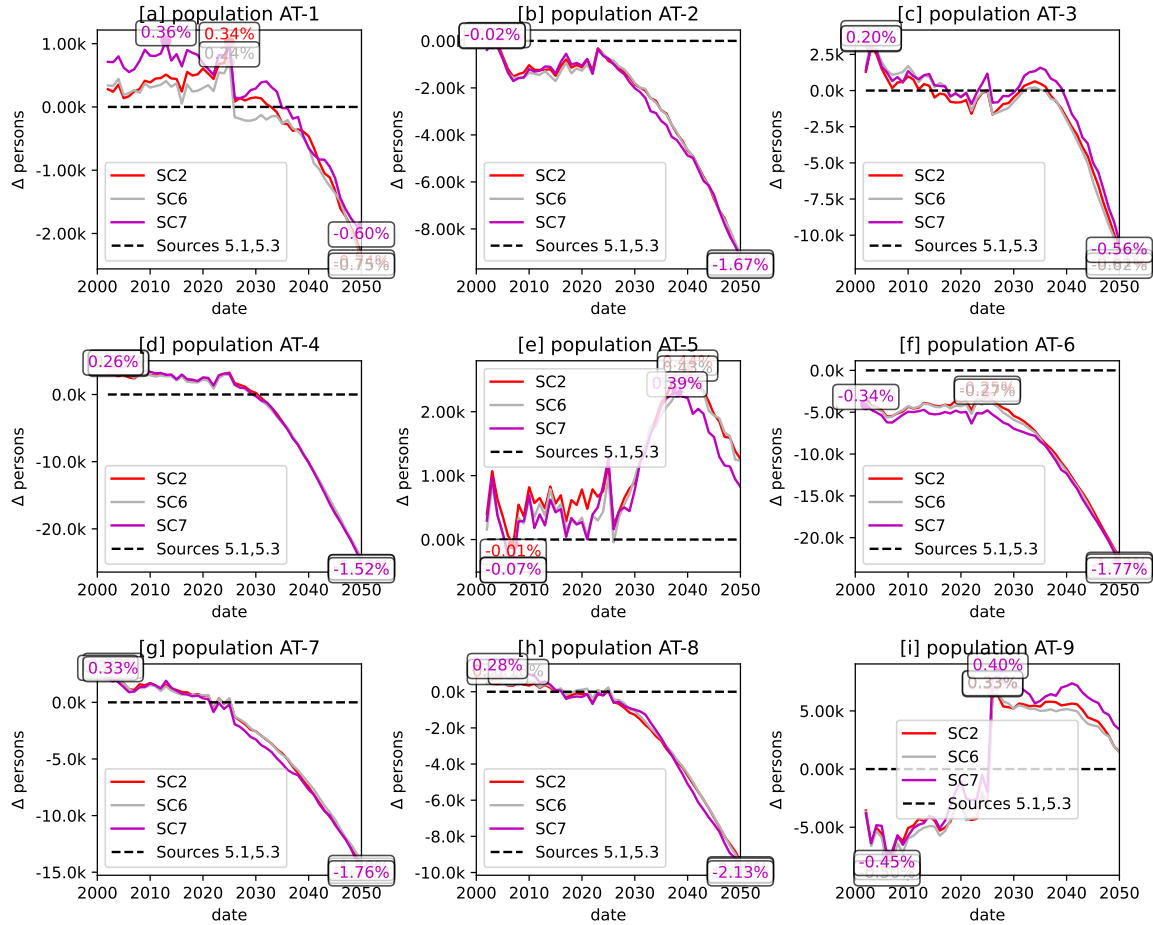


Figure 27: Differences between the population data/forecast (Source 5.1, 5.3) and simulation scenarios SC2, SC6 and SC7 for the nine federal-states of Austria.

Funding

The methods presented are the result of many years of research in the field of population modelling and its application (e.g. epidemiology, logistics, etc.). We would therefore like to express our gratitude to several funding agencies, whose support for research projects has enabled important contributions to the methods presented in this work:

- Austrian Research Promotion Agency (FFG), projects POPSICLE (FO999918405, 2024-2026), CIDS (881665, 2020-2021), AundO (867178, 2018-2020), DEXHELPP (843550, 2014-2018)
- Austrian Science Fund (FWF), project DynOptTestControl (I 5908-G, 2022-2026)
- Vienna Science and Technology Fund (WWTF), projects BETTER (EICOV20028, 2023-2026), HPVVienna (2025-2027), Syd19 (COV20-035, 2020)
- Society for Medical Decision Making (SMDM), project TAV-COVID (gbmf9634, 2020-2021)

References

- [1] Michel L Balinski and H Peyton Young. *Fair representation: meeting the ideal of one man, one vote*. Rowman & Littlefield, 2010.
- [2] Michel Louis Balinski and H Peyton Young. On Huntington methods of apportionment. *SIAM Journal on Applied Mathematics*, 33(4):607–618, 1977. Publisher: SIAM.
- [3] Martin Bicher, Dominik Brunmeir, Claire Rippinger, Christoph Urach, Maximilian Viehauser, Daniele Giannandrea, Hannah Kastinger, and Niki Popper. Gepoc abm, generic population concept – agent-based model, version 2.2, 2025.
- [4] Martin Bicher, Barbara Glock, Florian Miksch, Günter Schneckenreither, and Niki Popper. Definition, Validation and Comparison of Two Population Models for Austria. In *Proceedings of 4th UBT Annual International Conference on Business, Technology and Innovation*, Durres, Albania, 2015. UBT - Higher Education Institution.
- [5] Martin Bicher and Niki Popper. Mean-Field Approximation of a Microscopic Population Model for Austria. In *Proceedings of the 9th EUROSIM Congress on Modelling and Simulation*, pages 544–545, Oulu, Finland, 2016.
- [6] Martin Bicher, Christoph Urach, and Niki Popper. GEPOC ABM: A Generic Agent-Based Population Model for Austria. In *Proceedings of the 2018 Winter Simulation Conference*, pages 2656–2667, Gothenburg, Sweden, 2018. IEEE.
- [7] Martin Bicher, Matthias Wastian, Dominik Brunmeir, Matthias Rößler, and Niki Popper. Review on Monte Carlo Simulation Stopping Rules: How Many Samples Are Really Enough? In *Proceedings of the 10th EUROSIM Congress on Modelling and Simulation*, Logrono, Spain, July 2019. In Print.
- [8] L Paul Chew. Constrained delaunay triangulations. In *Proceedings of the third annual symposium on Computational geometry*, pages 215–222, 1987.
- [9] W Edwards Deming and Frederick F Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

- [10] William Farr. On the Construction of Life-Tables, Illustrated by a New Life-Table of the Healthy Districts of England. *Philosophical Transactions of the Royal Society of London*, 149(1859):837–878, 1859.
- [11] Statistics Finland. Probability of death.
- [12] A. Hanika and J. Klotz. Sterbetafel und leibrententafel 2010/2012 für Österreich. *Statistische Nachrichten*, 2, 2014.
- [13] Chong-Wei Huang and Tian-Yuan Shih. On the complexity of point-in-polygon algorithms. *Computers & Geosciences*, 23(1):109–118, 1997.
- [14] J. Klotz and J. Klimont. Lebenserwartung in gesundheit. zeitliche entwicklung 1978 bis 2014. *Statistische Nachrichten*, 4, 2016.
- [15] Frances Robinson. The alabama paradox. *Teaching Mathematics and its Applications: An International Journal of the IMA*, 1(2):69–72, 1982.

A Appendix

A.1 Validation Plots - Population

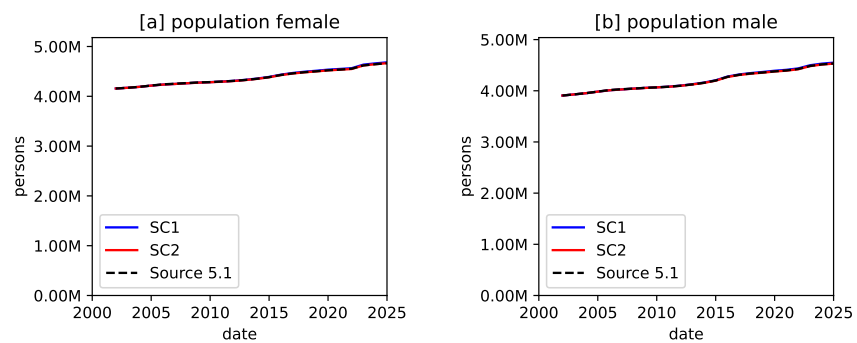


Figure 28: Comparison between population data from Source 5.1 and the simulation scenarios SC1 and SC2 for male and female persons.

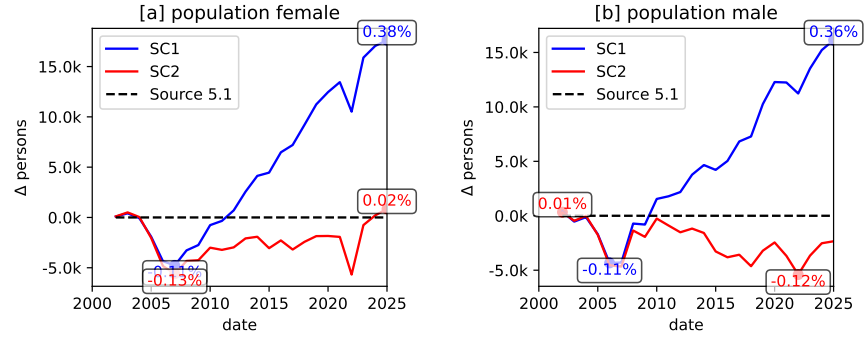


Figure 29: Differences between population data from Source 5.1 and the simulation scenarios SC1 and SC2 for male and female persons.

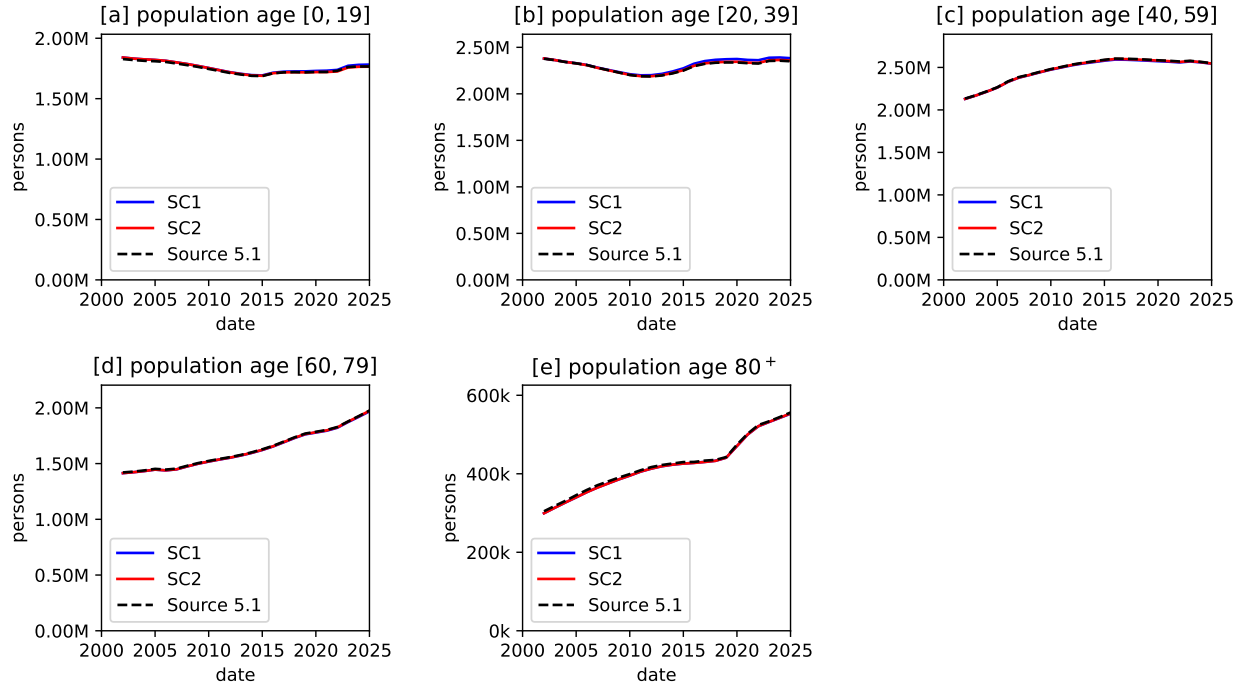


Figure 30: Comparison between population data from Source 5.1 and the simulation scenarios SC1 and SC2 for different age cohorts.

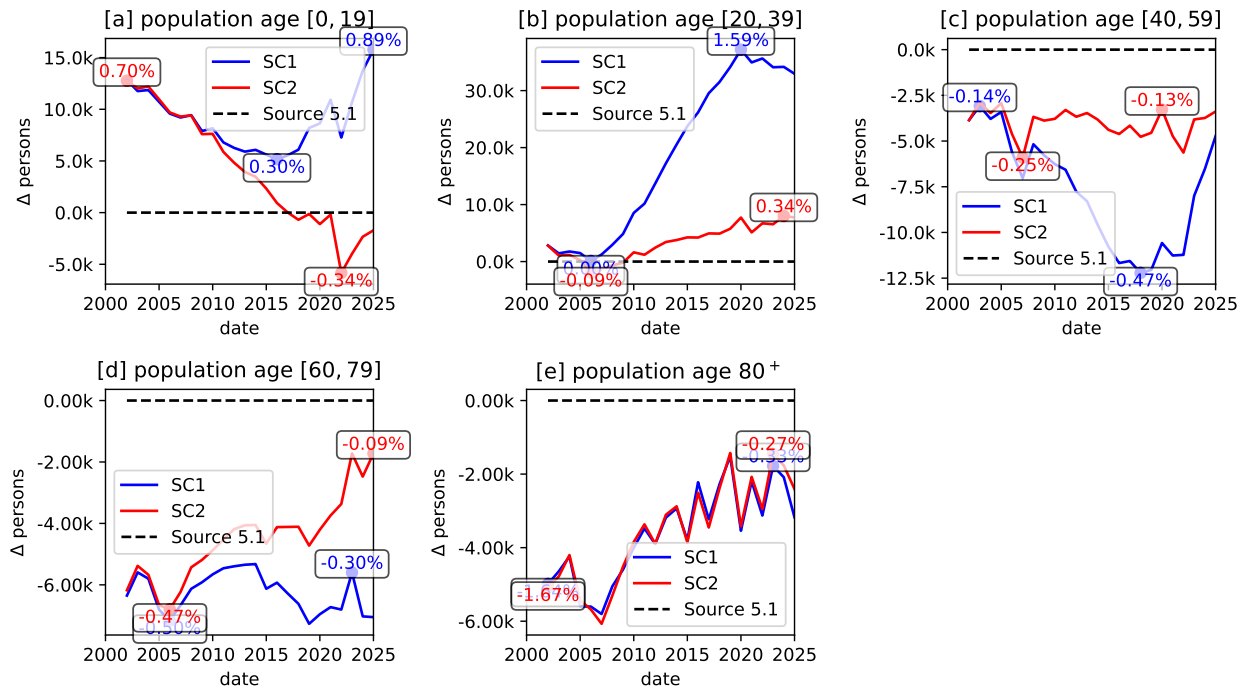


Figure 31: Differences between population data from Source 5.1 and the simulation scenarios SC1 and SC2 for different age cohorts.

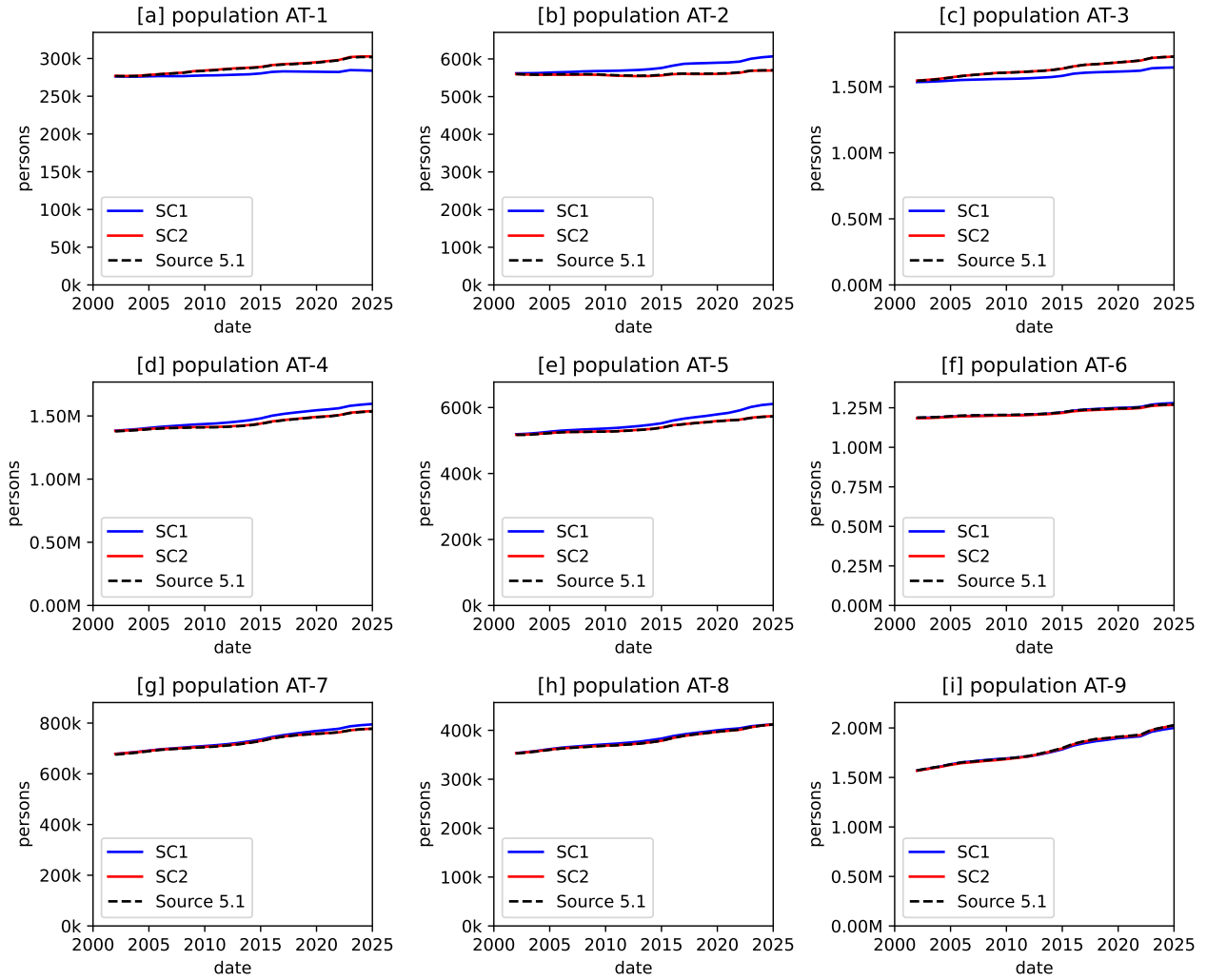


Figure 32: Comparison between population data from Source 5.1 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

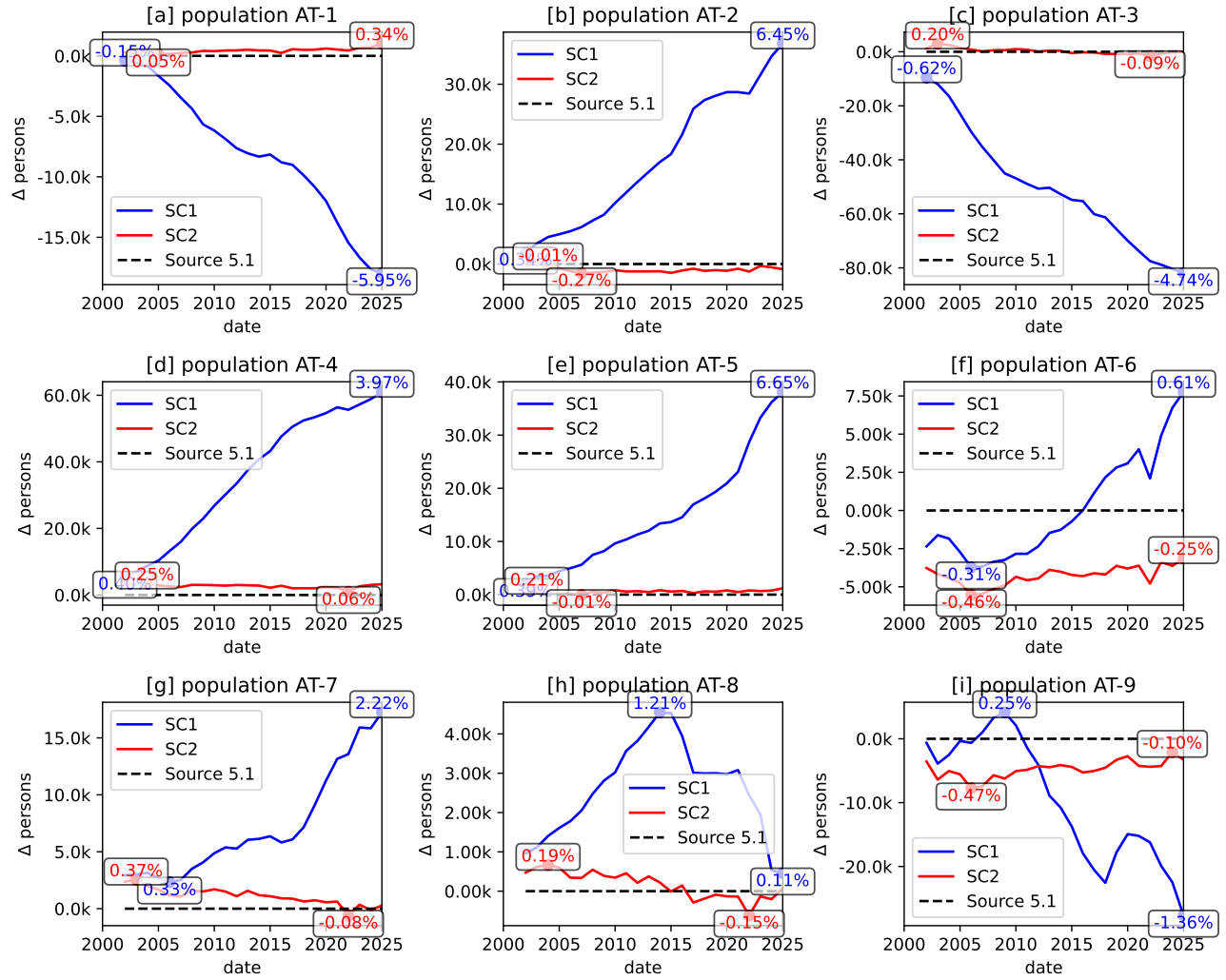


Figure 33: Differences between population data from Source 5.1 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

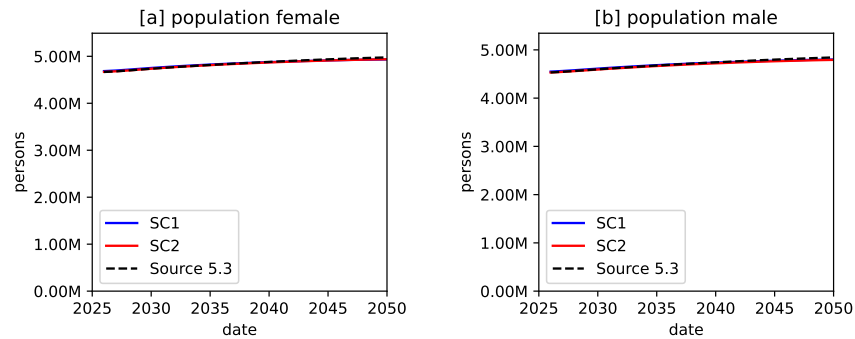


Figure 34: Comparison between population forecast from Source 5.3 and the simulation scenarios SC1 and SC2 for male and female persons.

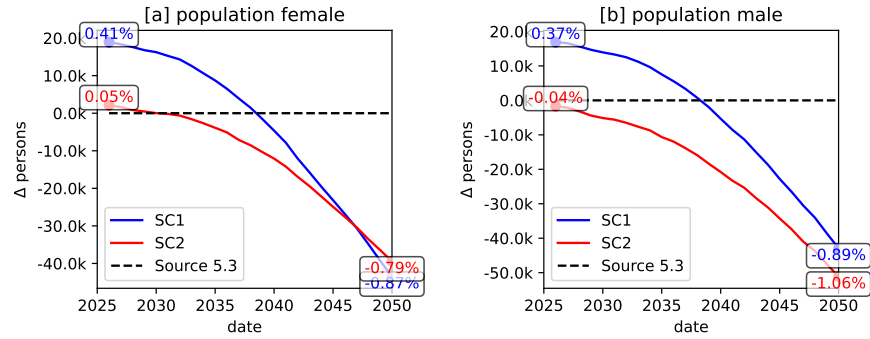


Figure 35: Differences between population forecast from Source 5.3 and the simulation scenarios SC1 and SC2 for male and female persons.

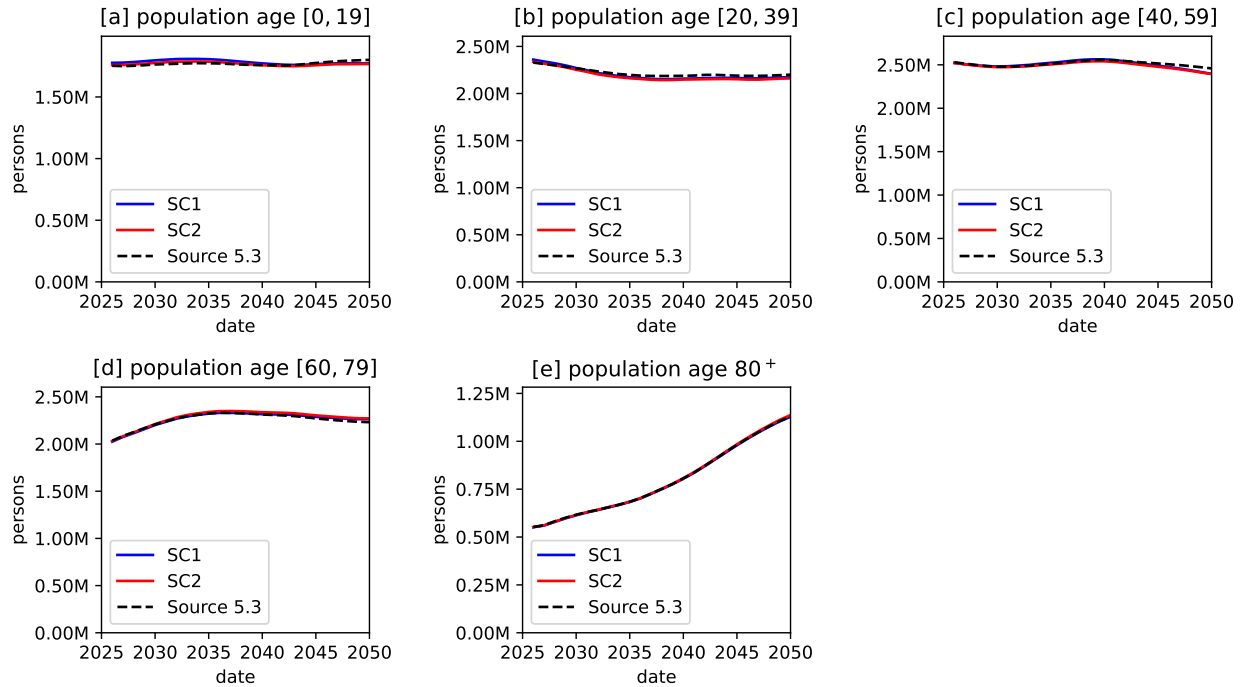


Figure 36: Comparison between population forecast from Source 5.3 and the simulation scenarios SC1 and SC2 for different age cohorts.

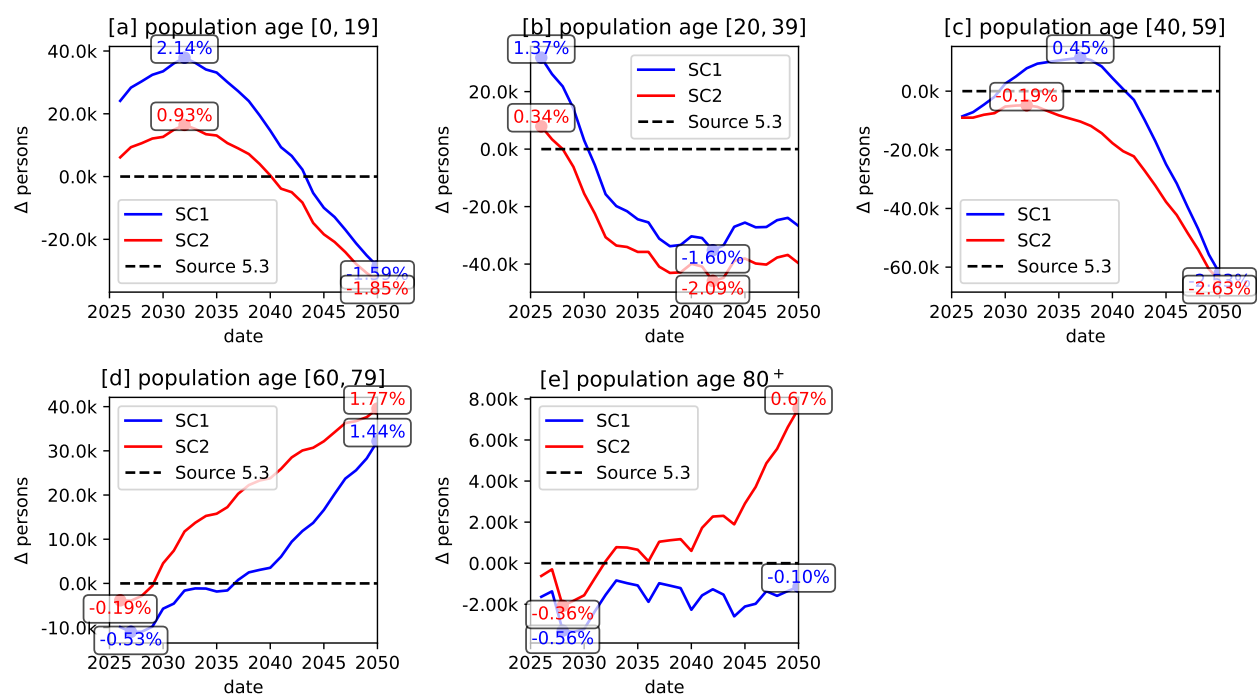


Figure 37: Differences between population forecast from Source 5.3 and the simulation scenarios SC1 and SC2 for different age cohorts.

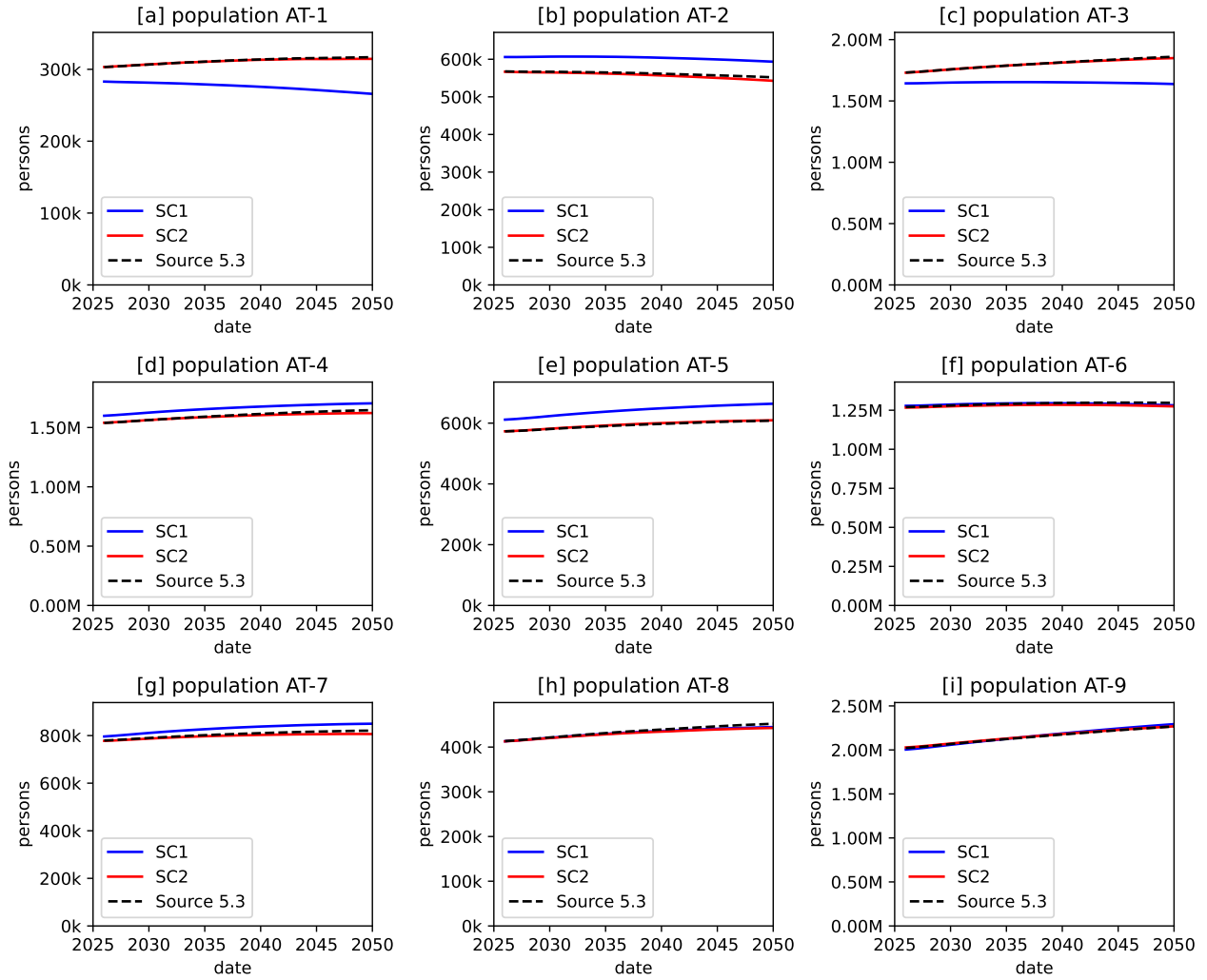


Figure 38: Comparison between population forecast from Source 5.3 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

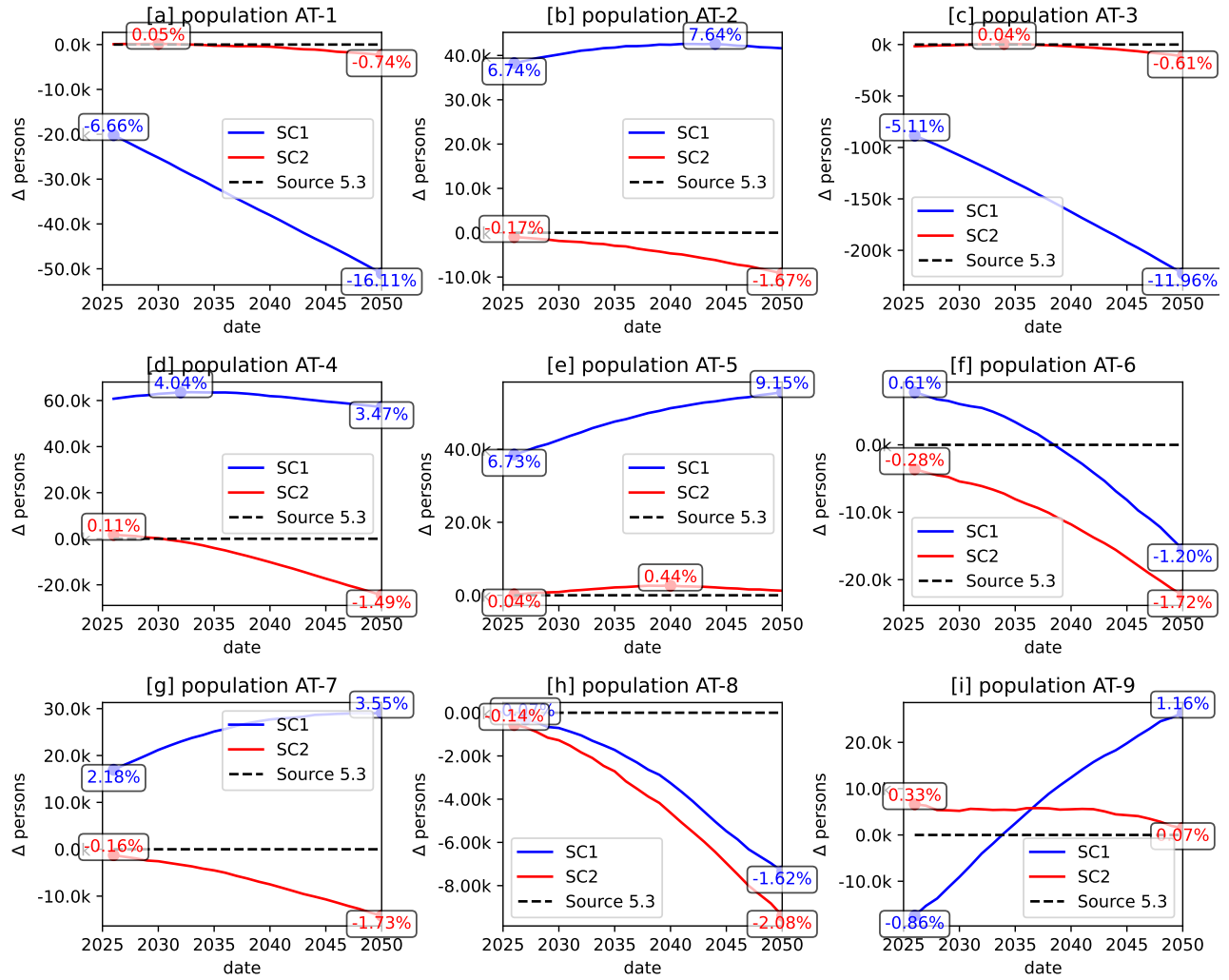


Figure 39: Differences between population forecast from Source 5.3 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

A.2 Validation Plots - Births

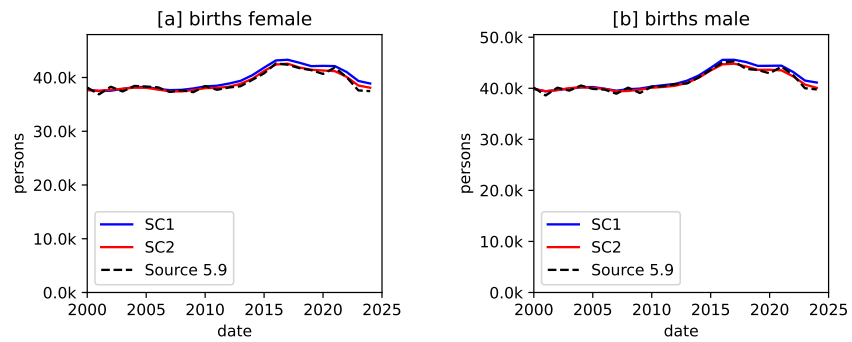


Figure 40: Comparison between births data from Source 5.9 and the simulation scenarios SC1 and SC2 for male and female persons.

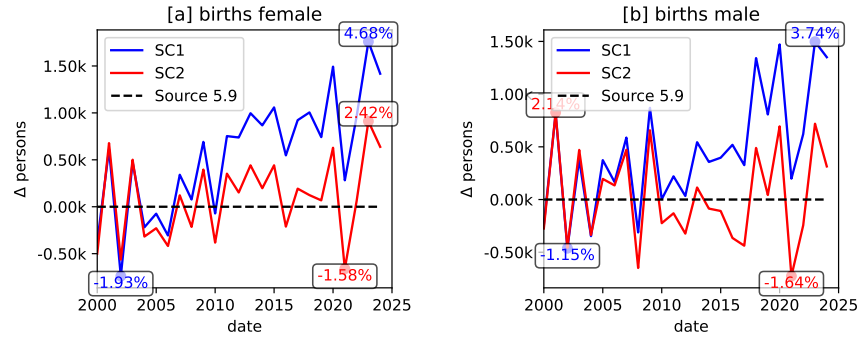


Figure 41: Differences between births data from Source 5.9 and the simulation scenarios SC1 and SC2 for male and female persons.

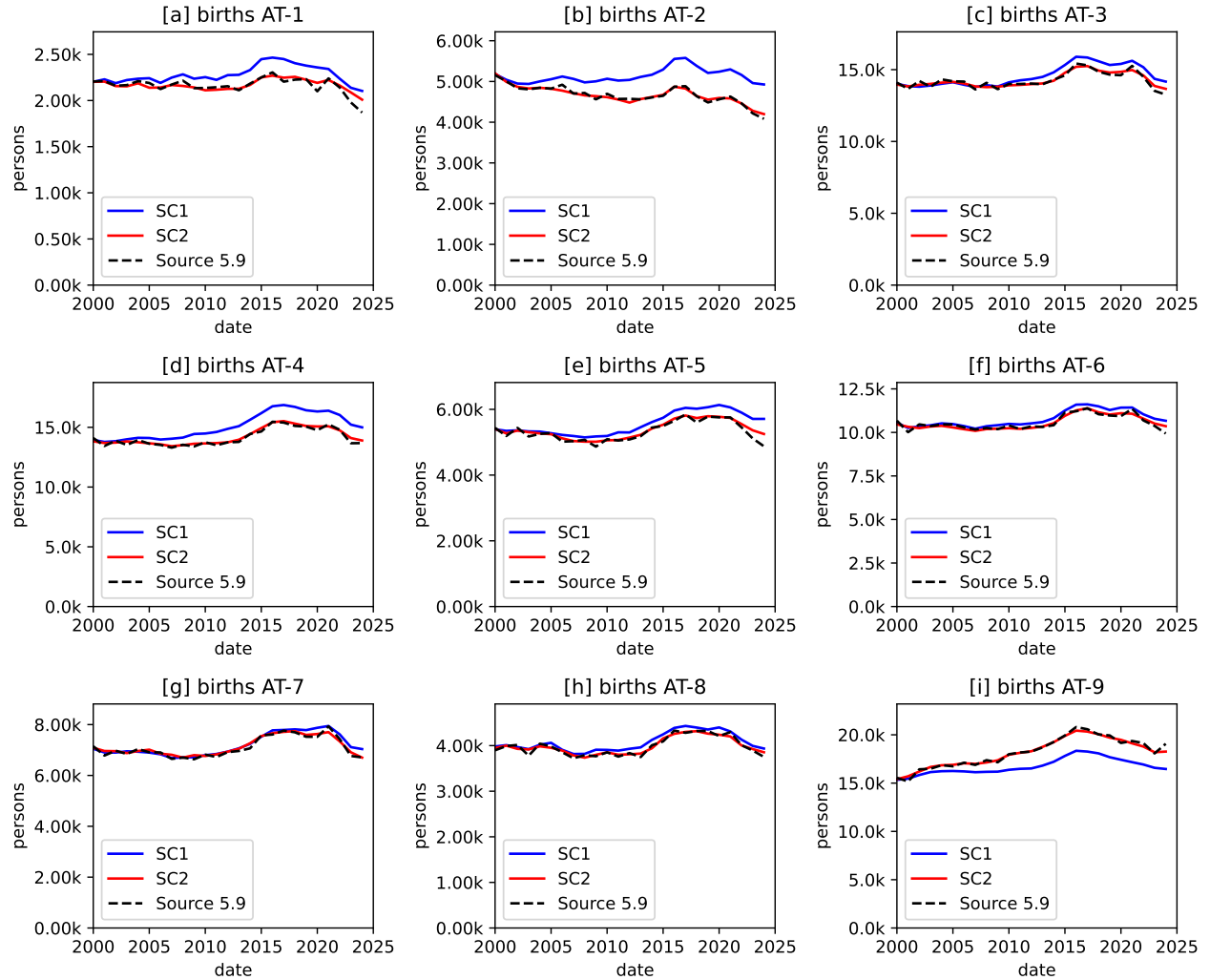


Figure 42: Comparison between births data from Source 5.9 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

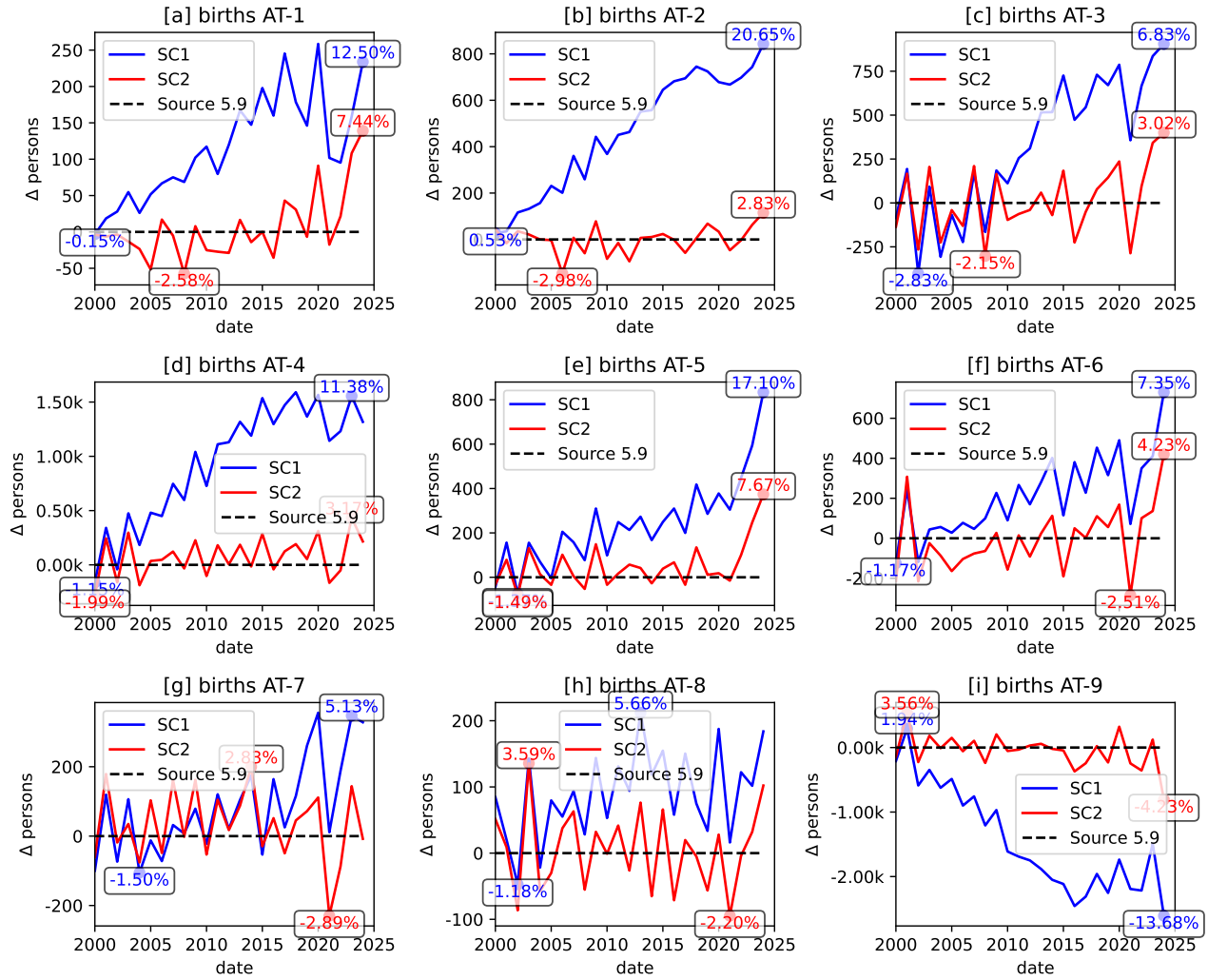


Figure 43: Differences between births data from Source 5.9 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

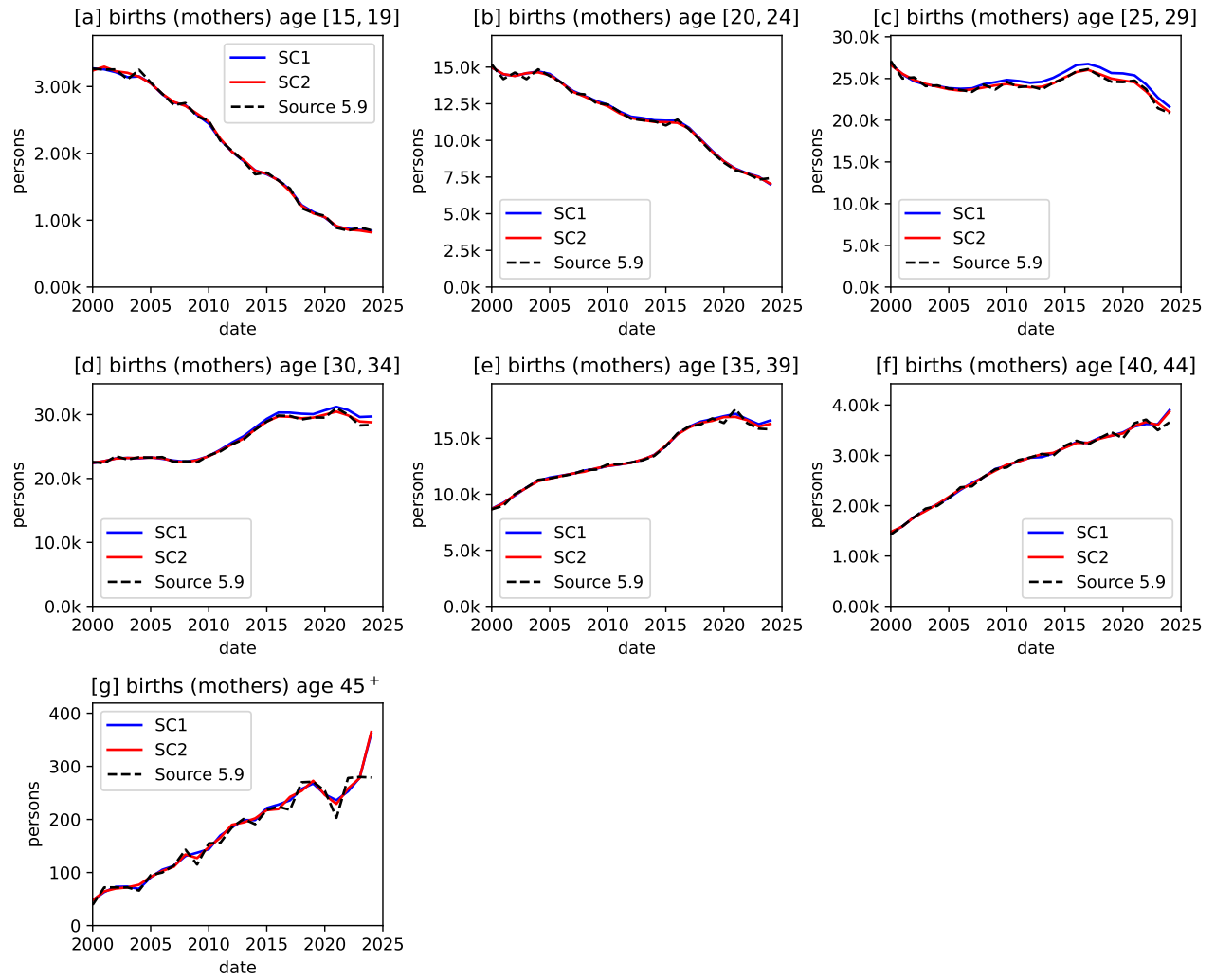


Figure 44: Comparison between births data from Source 5.9 and the simulation scenarios SC1 and SC2 for different age cohorts.

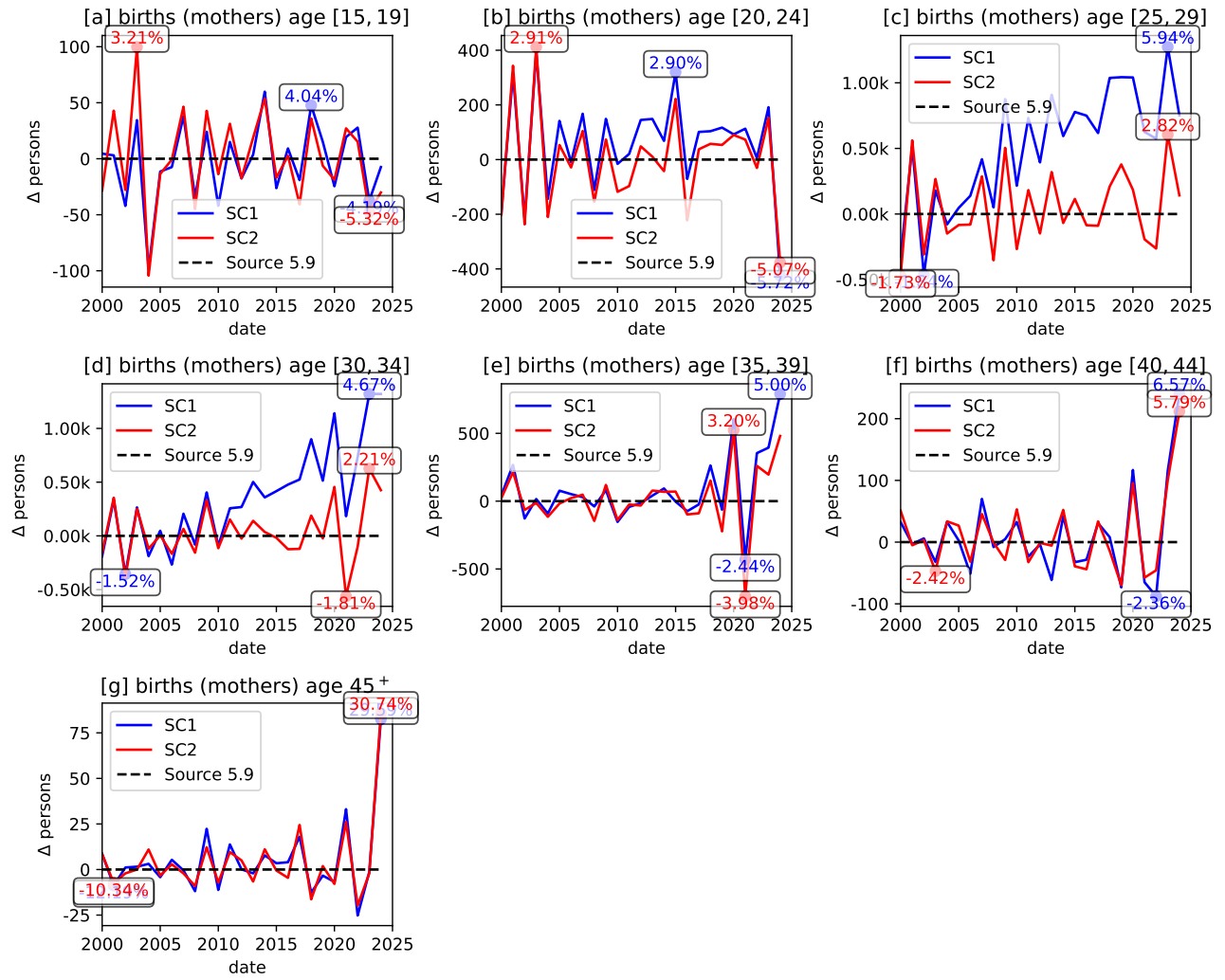


Figure 45: Differences between births data from Source 5.9 and the simulation scenarios SC1 and SC2 for different age cohorts.

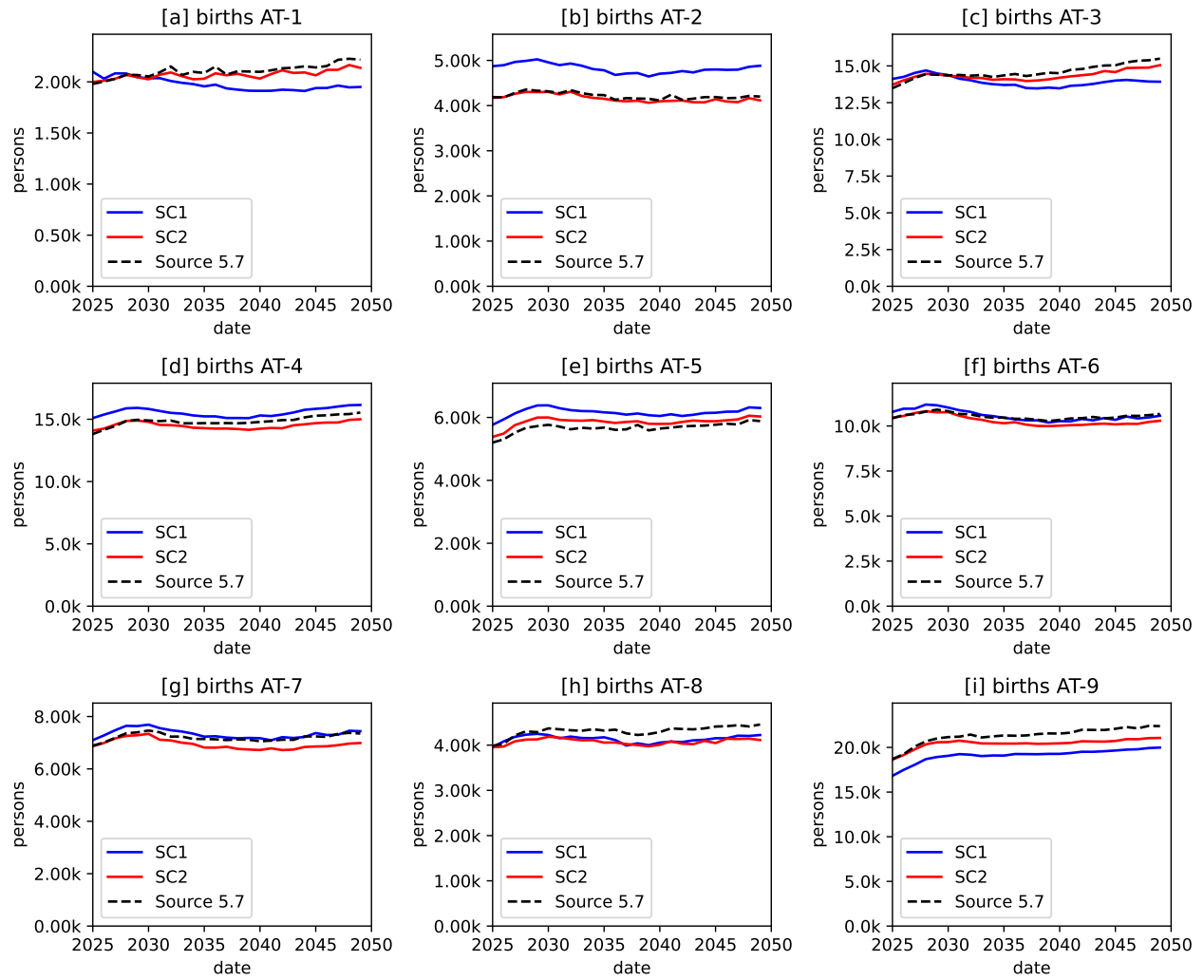


Figure 46: Comparison between births forecast from Source 5.7 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

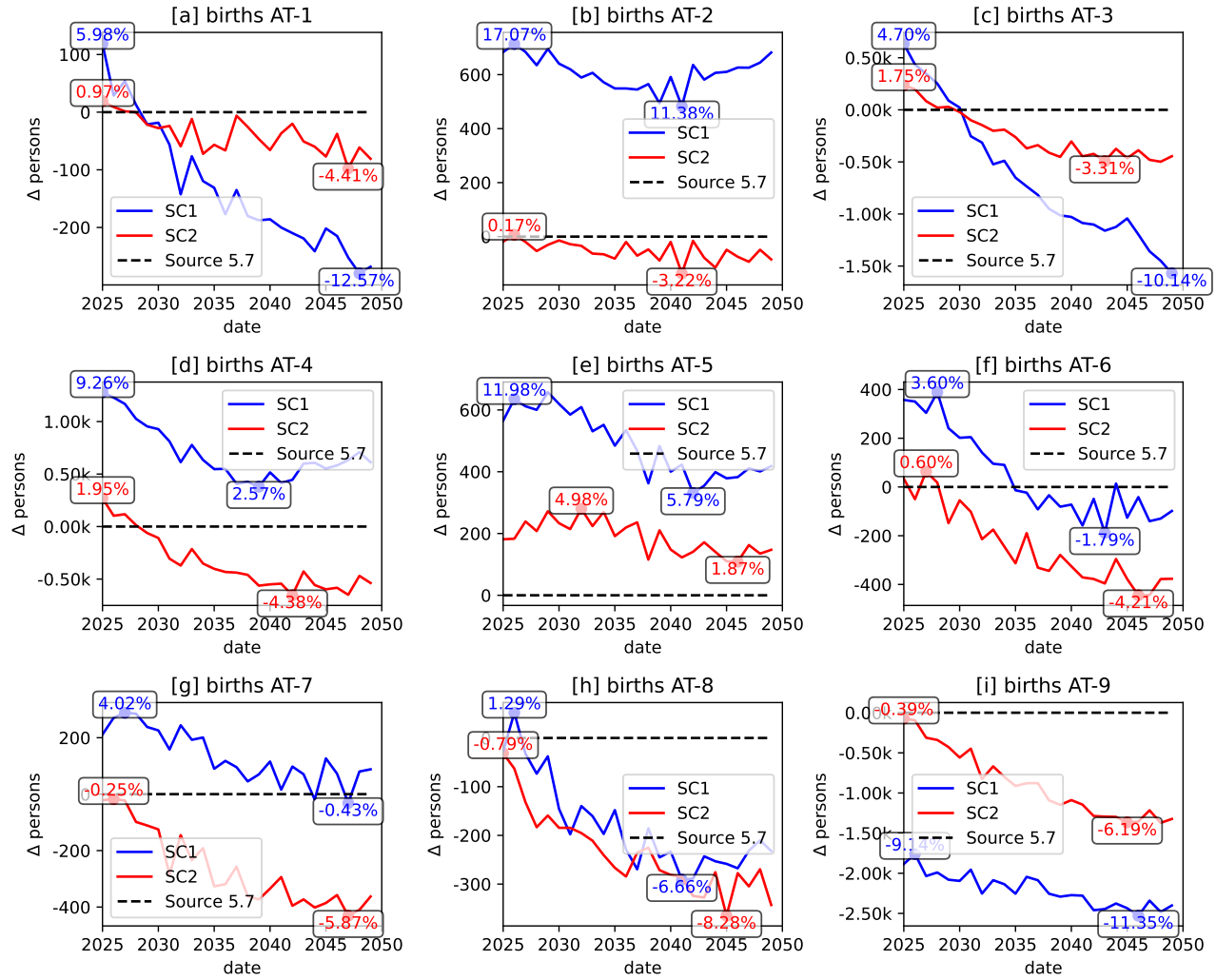


Figure 47: Differences between births forecast from Source 5.7 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

A.3 Validation Plots - Deaths

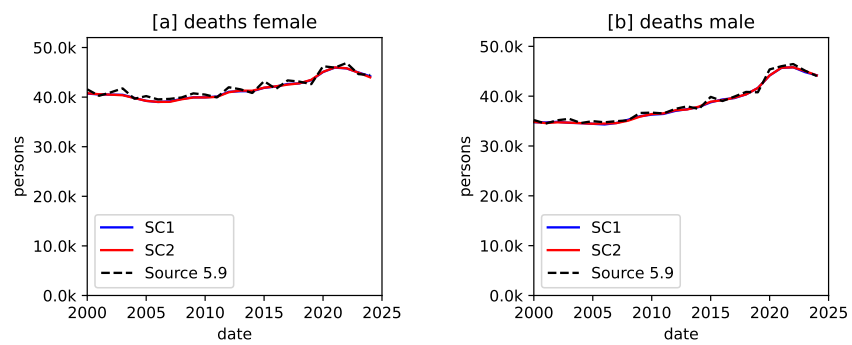


Figure 48: Comparison between deaths data from Source 5.9 and the simulation scenarios SC1 and SC2 for male and female persons.

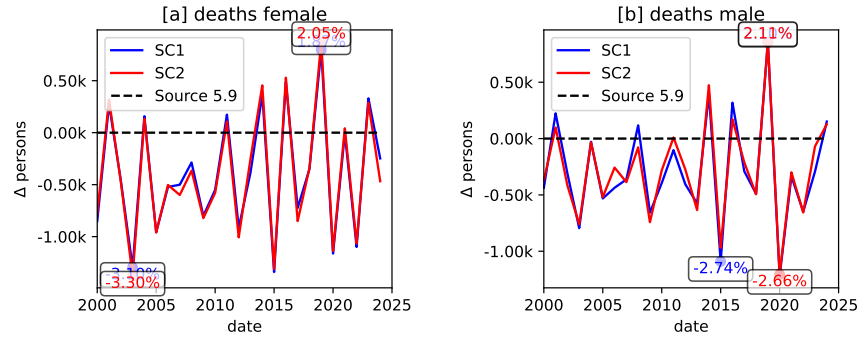


Figure 49: Differences between deaths data from Source 5.9 and the simulation scenarios SC1 and SC2 for male and female persons.

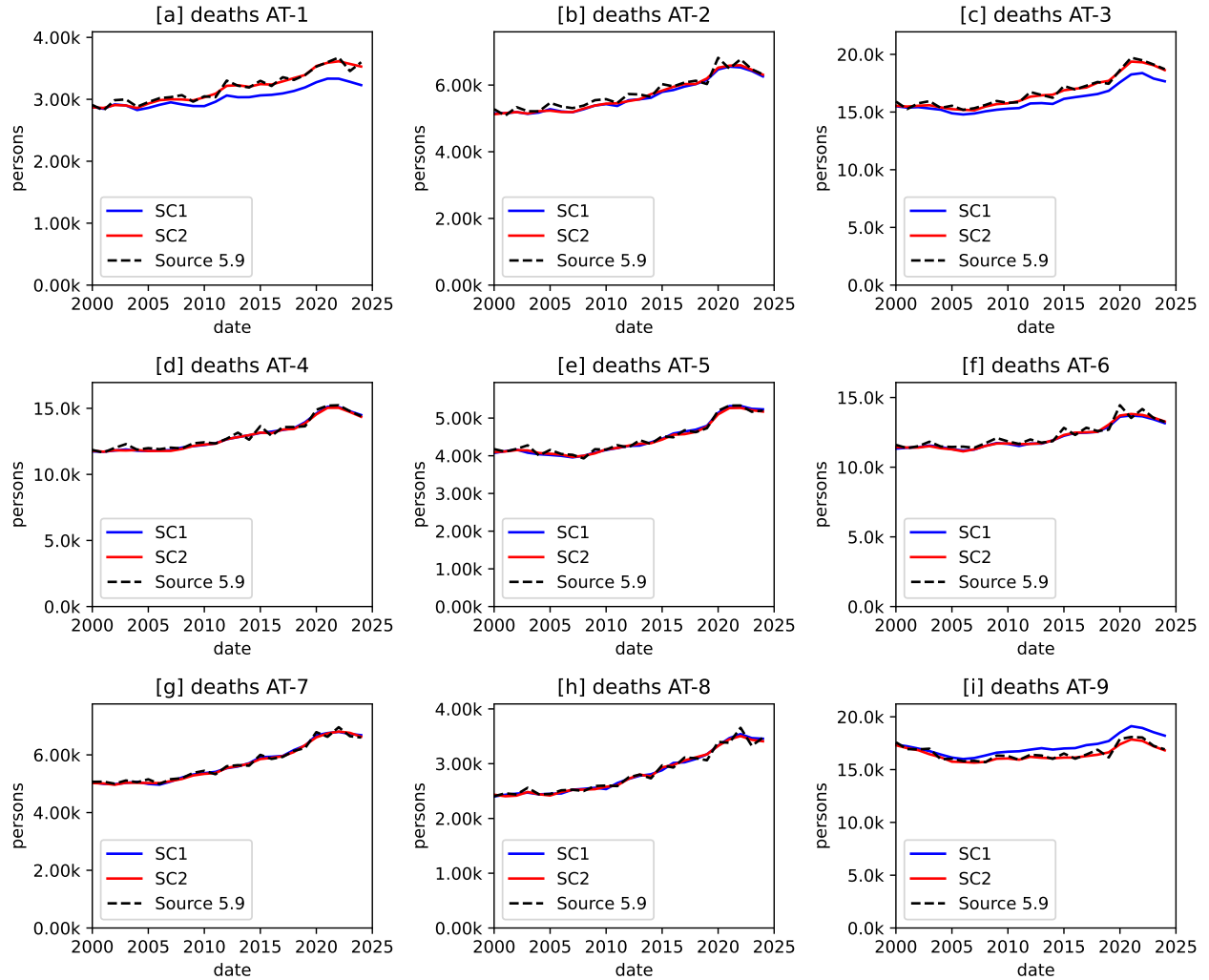


Figure 50: Comparison between deaths data from Source 5.9 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

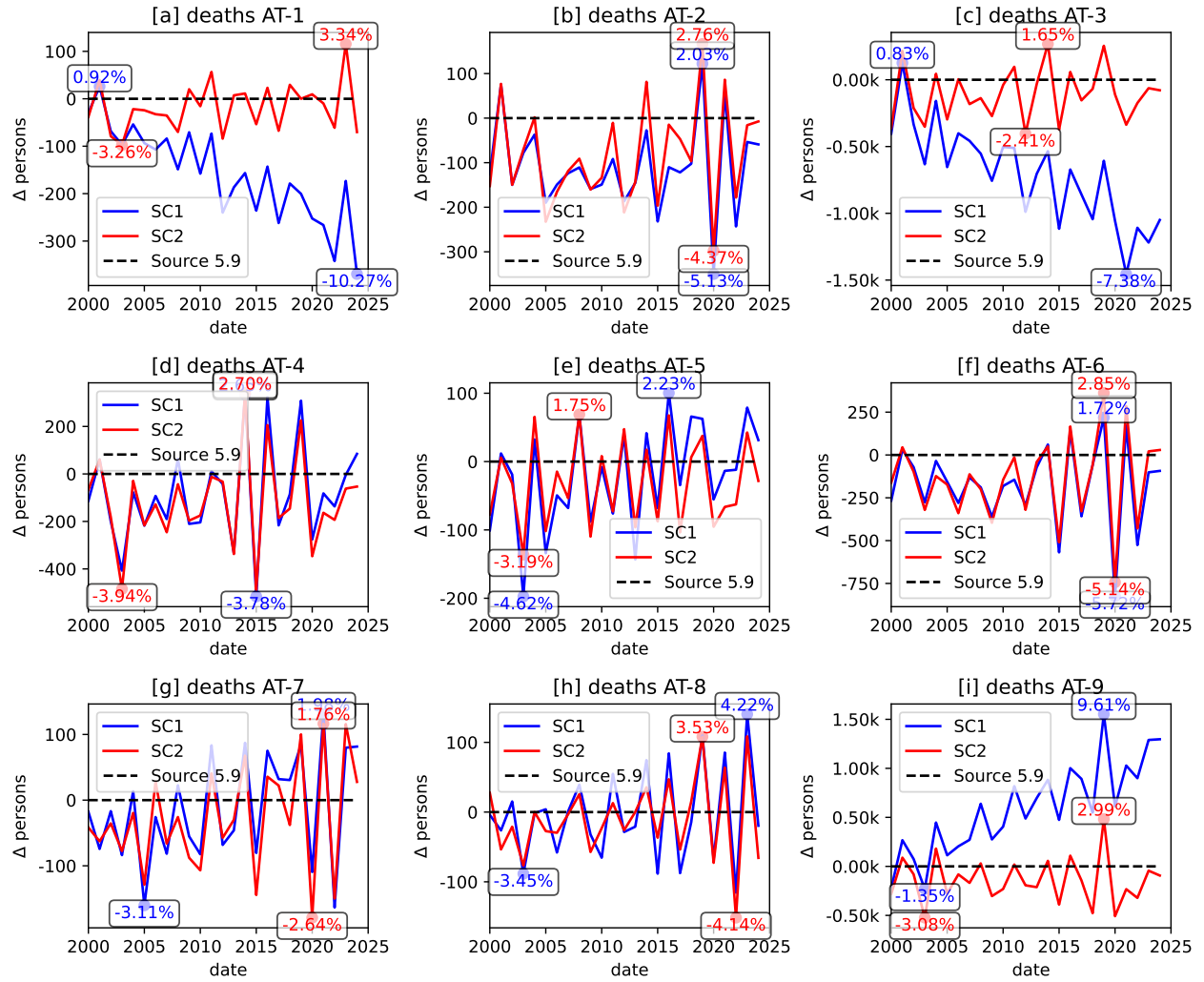


Figure 51: Differences between deaths data from Source 5.9 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

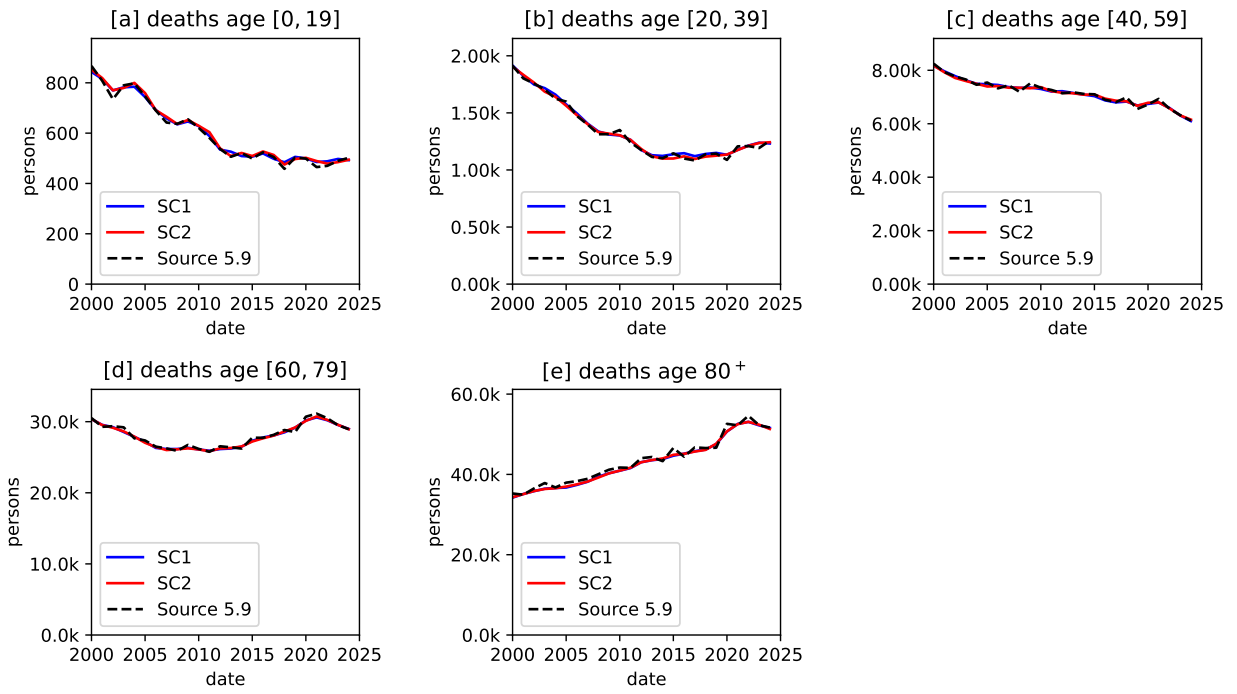


Figure 52: Comparison between deaths data from Source 5.9 and the simulation scenarios SC1 and SC2 for different age cohorts.

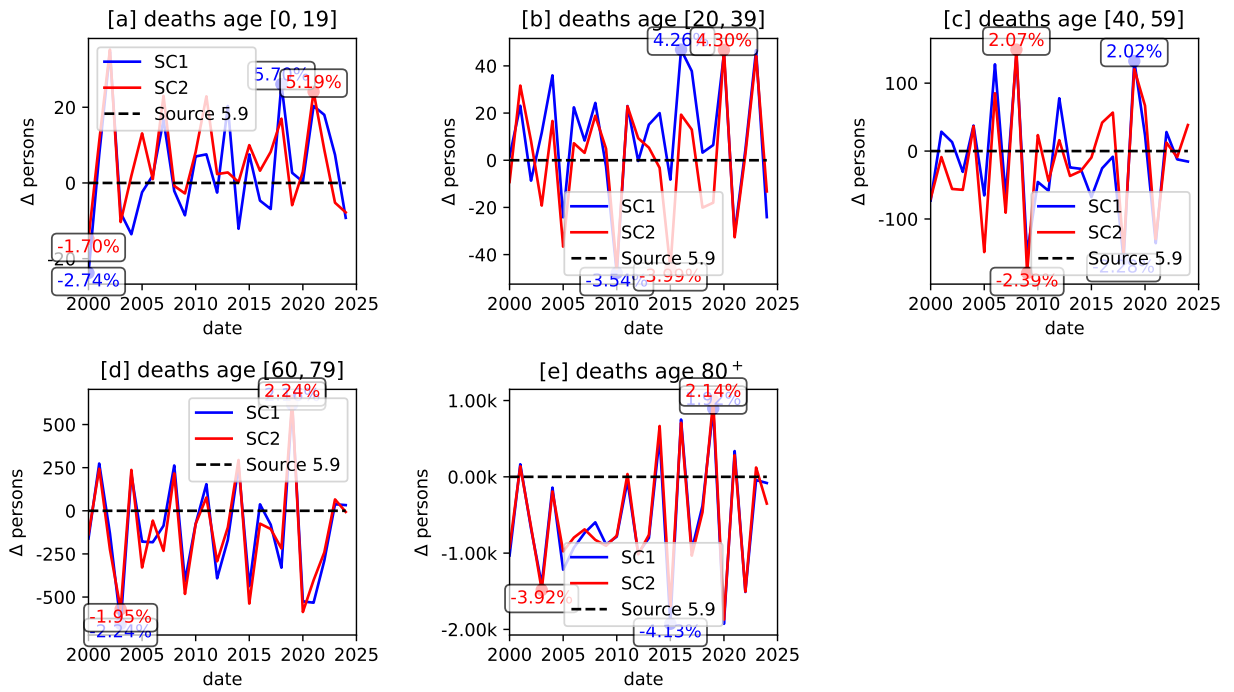


Figure 53: Differences between deaths data from Source 5.9 and the simulation scenarios SC1 and SC2 for different age cohorts.

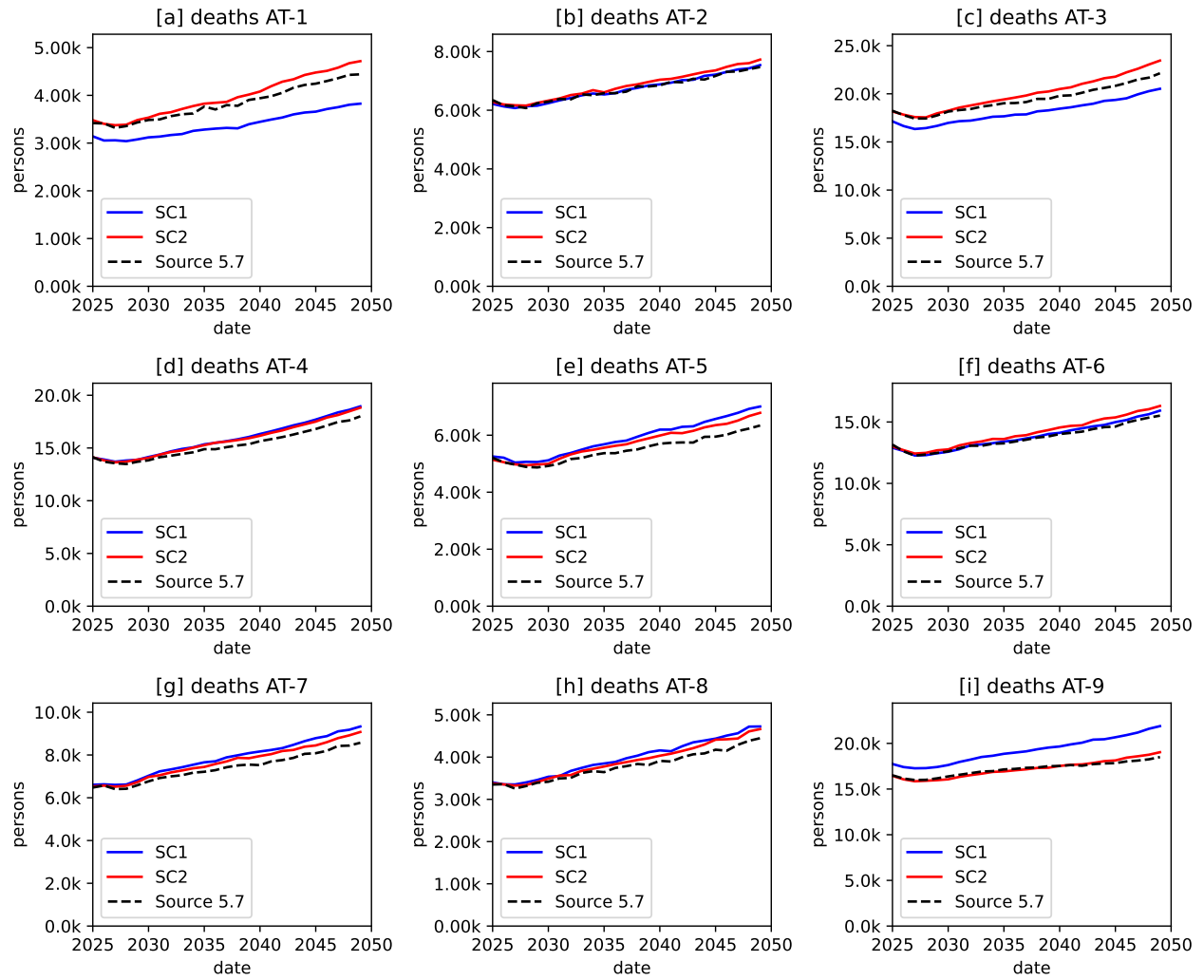


Figure 54: Comparison between deaths forecast from Source 5.7 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

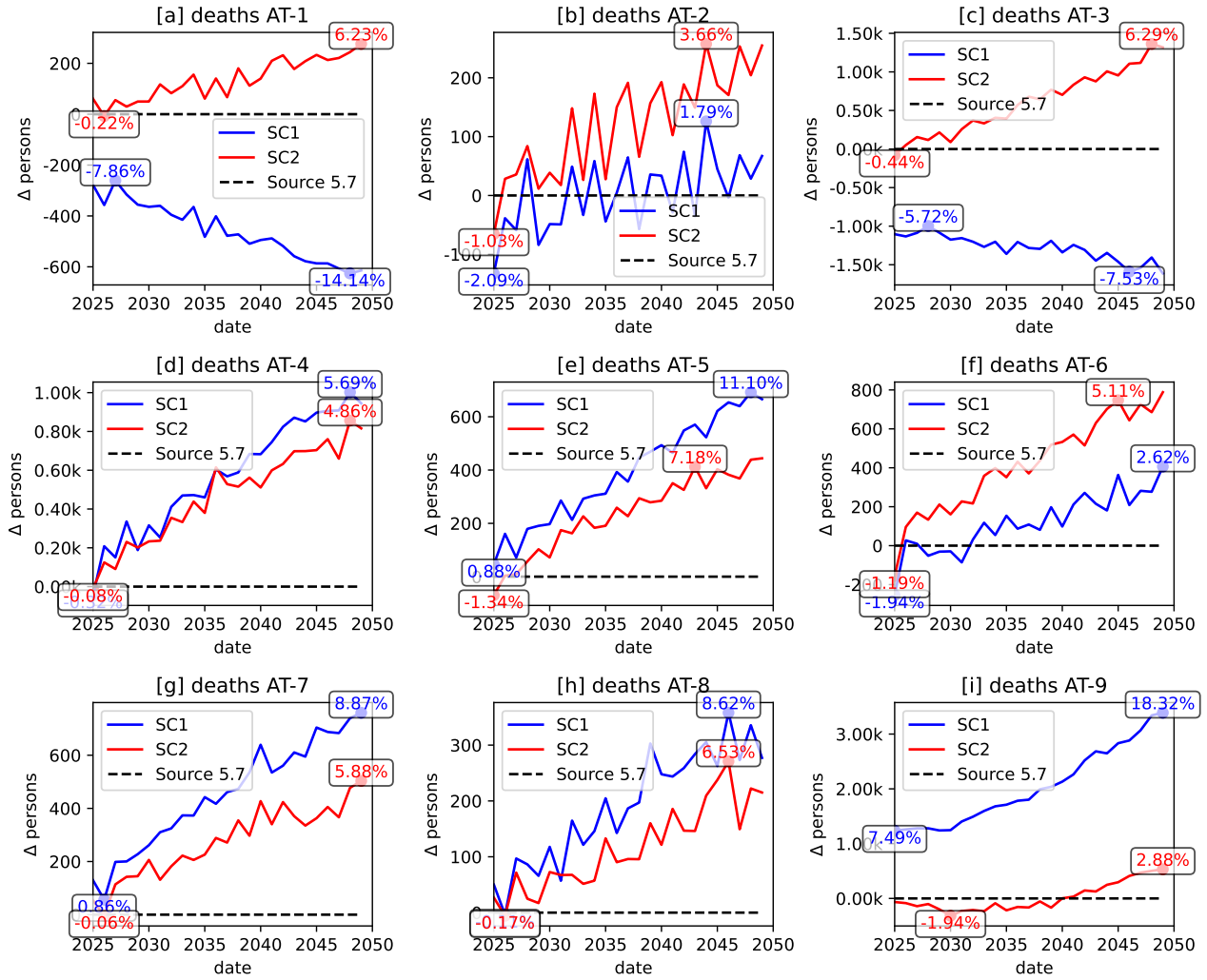


Figure 55: Differences between deaths forecast from Source 5.7 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

A.4 Validation Plots - Emigrants

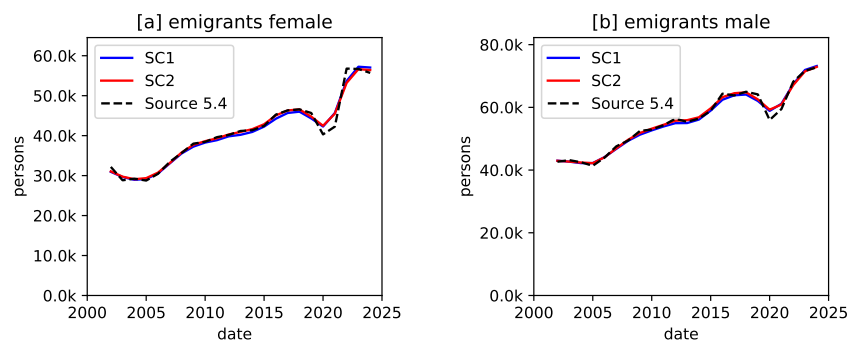


Figure 56: Comparison between emigrant data from Source 5.4 and the simulation scenarios SC1 and SC2 for male and female persons.

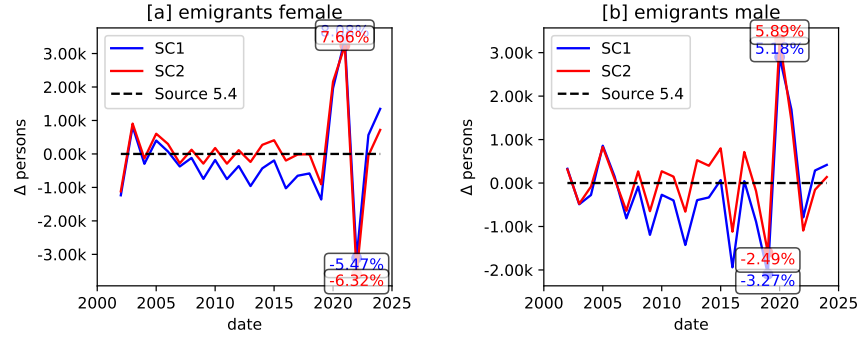


Figure 57: Differences between emigrants data from Source 5.4 and the simulation scenarios SC1 and SC2 for male and female persons.

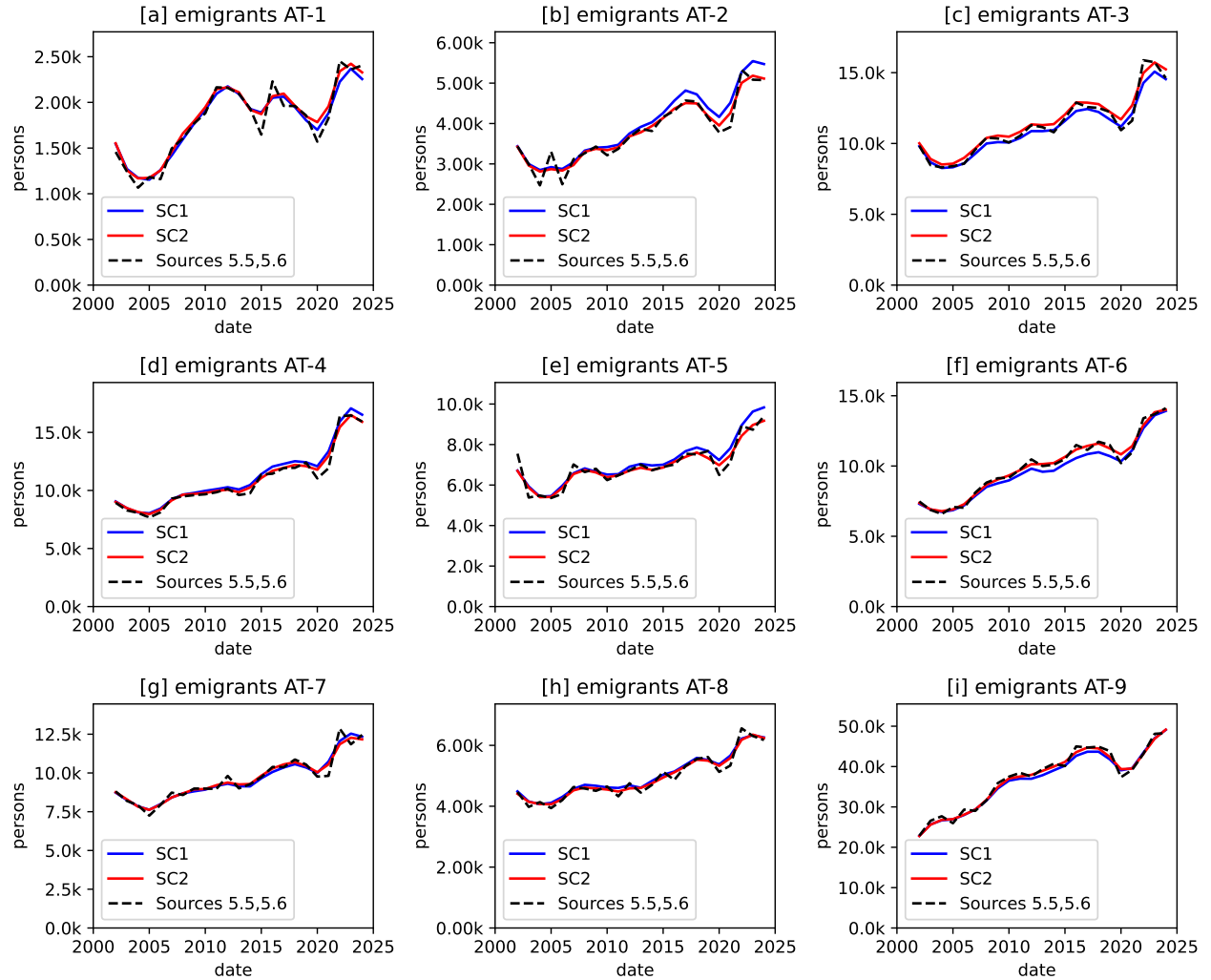


Figure 58: Comparison between emigrants data from Sources 5.5, 5.6 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

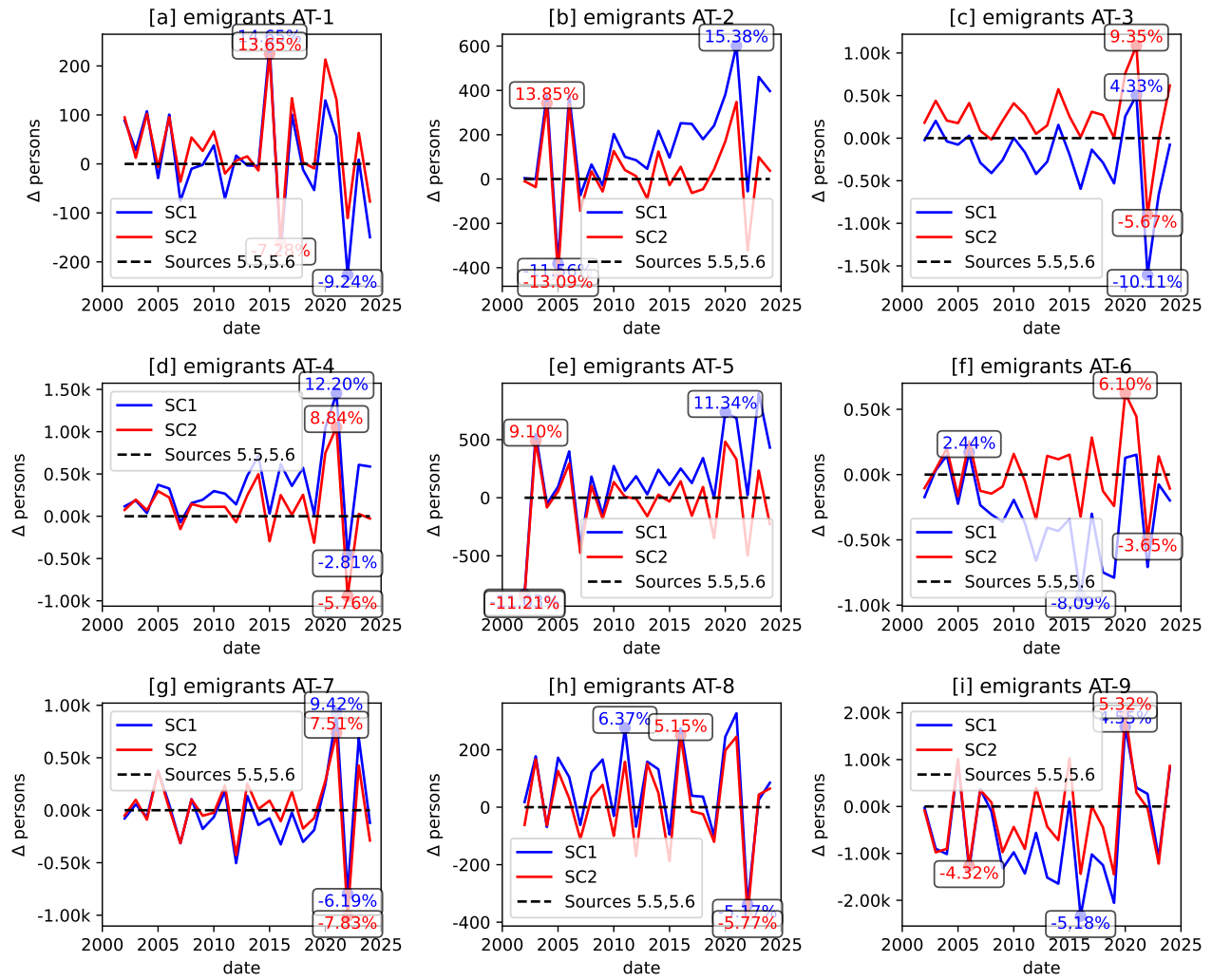


Figure 59: Differences between emigrants data from Sources 5.5, 5.5 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

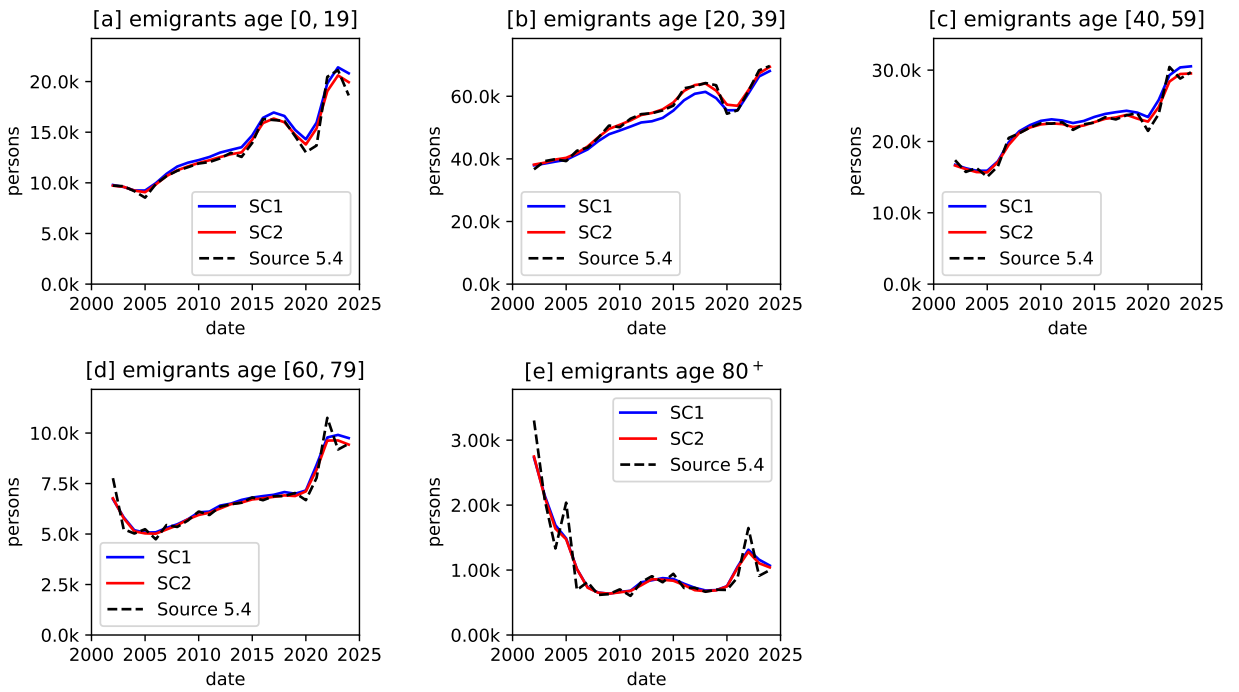


Figure 60: Comparison between emigrants data from Source 5.4 and the simulation scenarios SC1 and SC2 for different age cohorts.

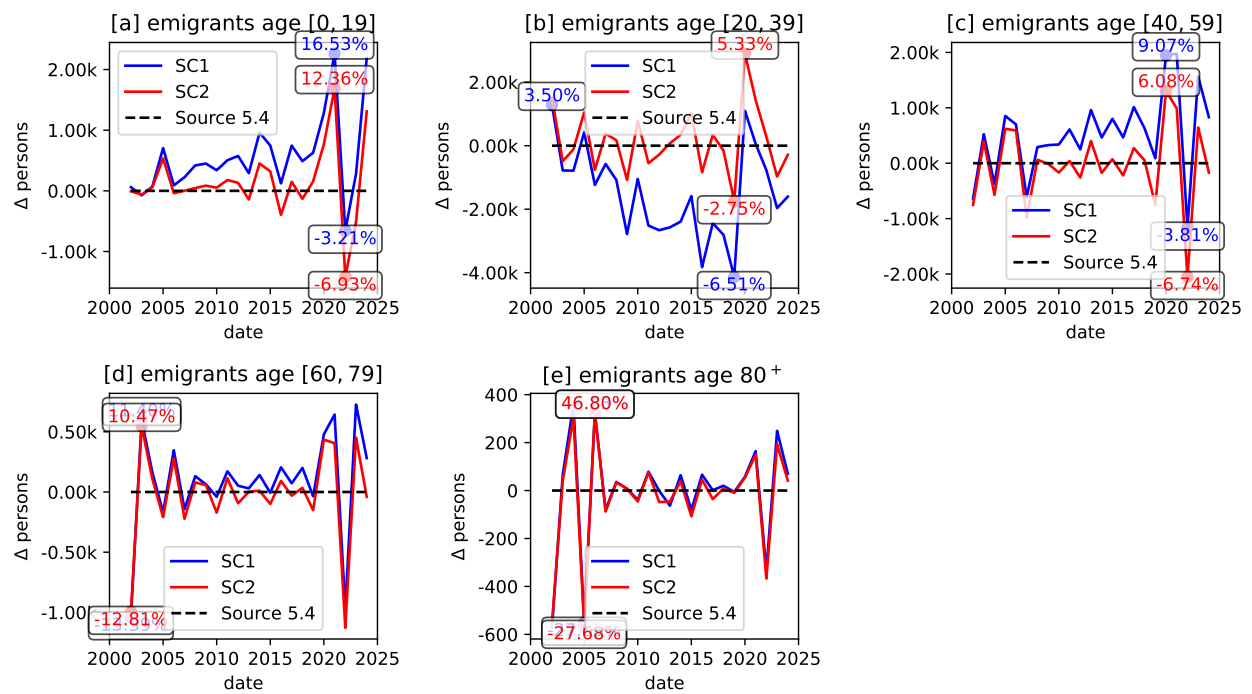


Figure 61: Differences between emigrants data from Source 5.4 and the simulation scenarios SC1 and SC2 for different age cohorts.

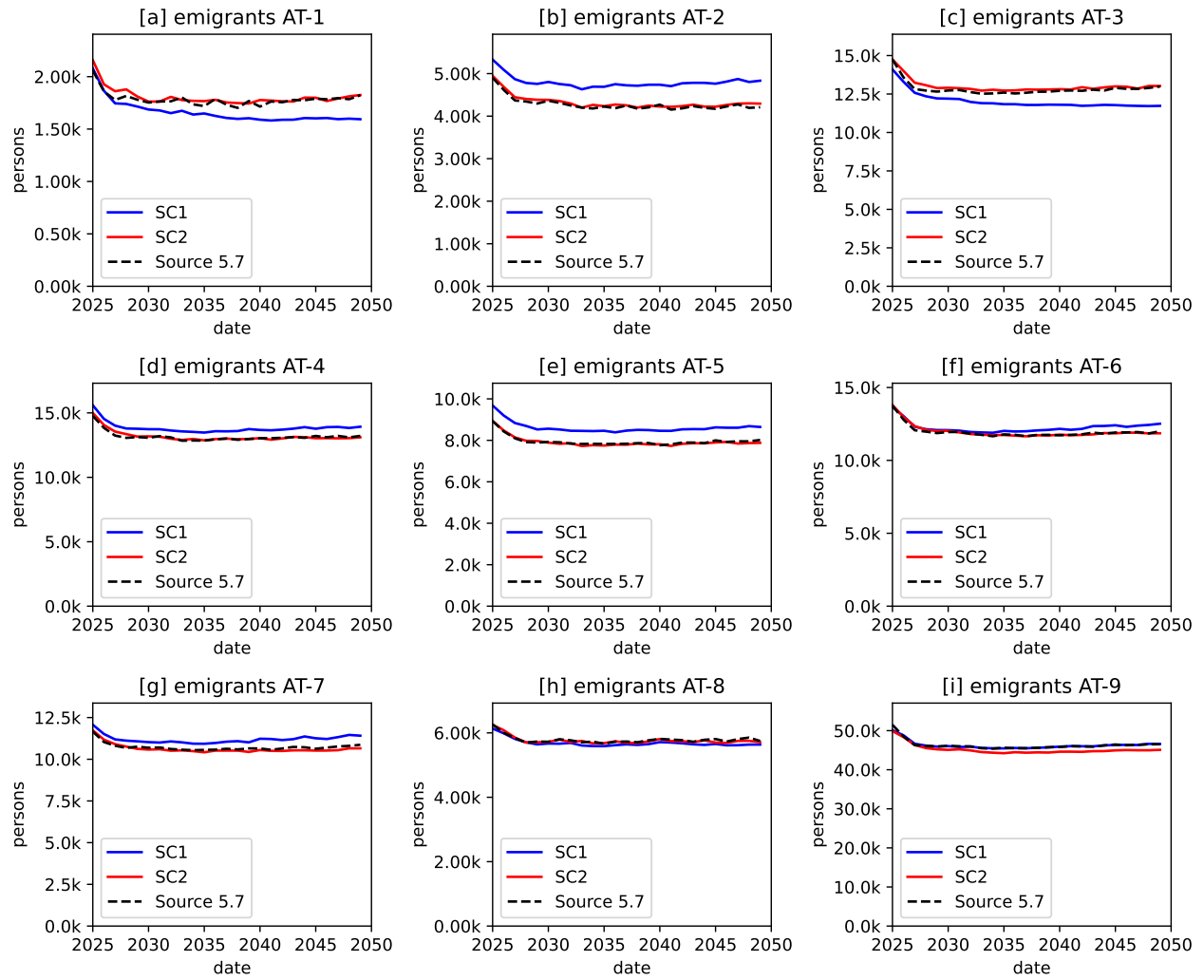


Figure 62: Comparison between emigrants forecast from Source 5.7 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

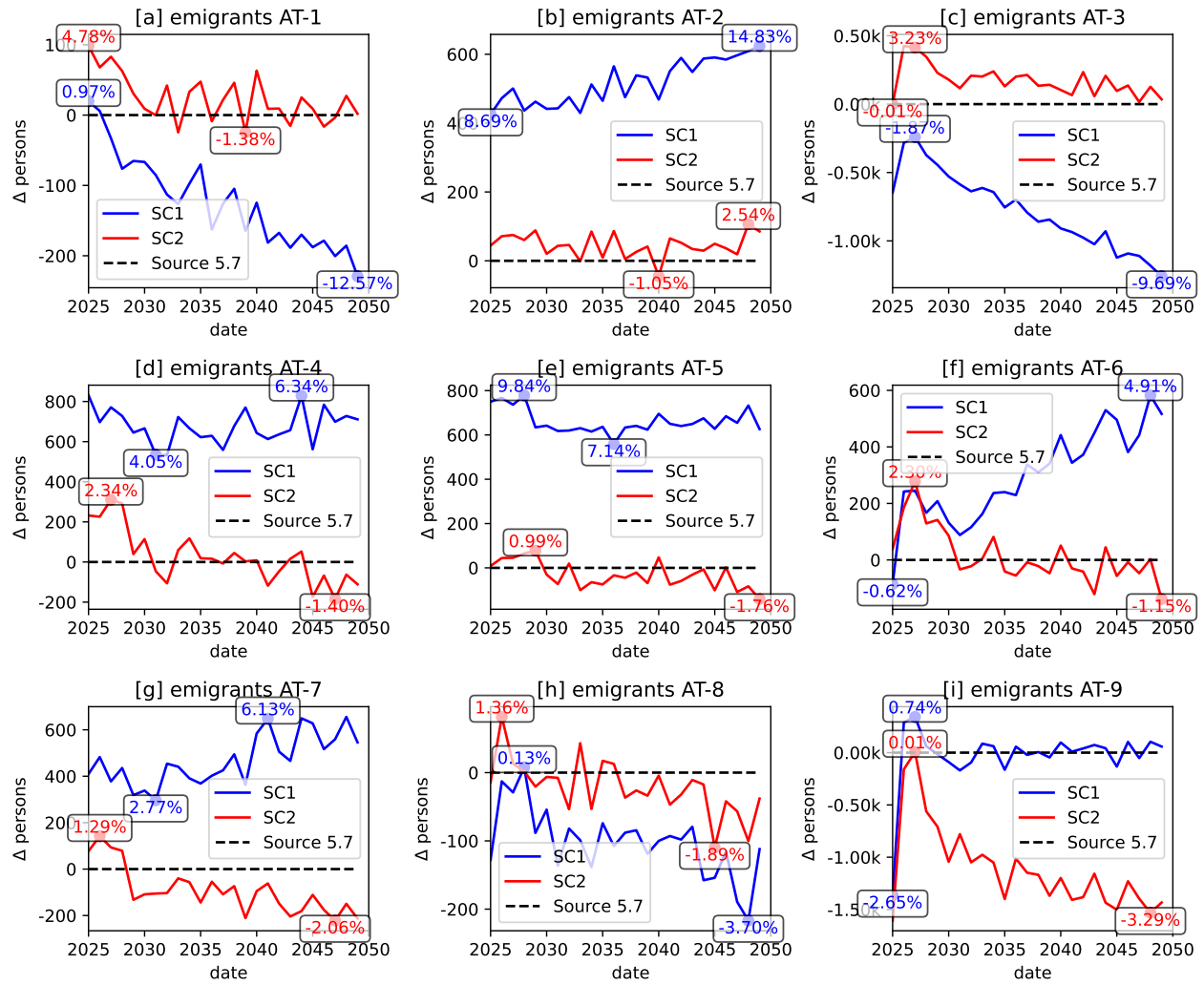


Figure 63: Differences between emigrants forecast from Source 5.7 and the simulation scenarios SC1 and SC2 for the nine federal-states of Austria.

A.5 Validation Plots - Internal Migration

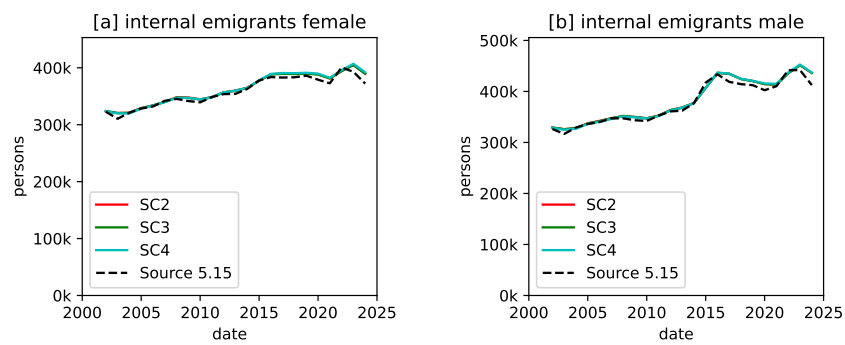


Figure 64: Comparison between internal emigrants data from Source 5.15 and the simulation scenarios SC2-SC4 for male and female persons.

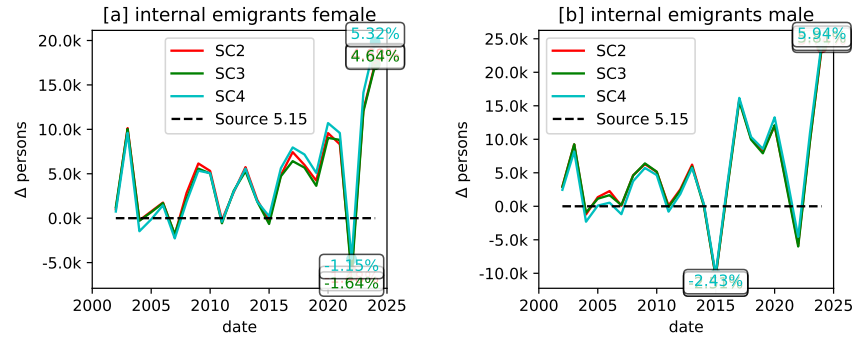


Figure 65: Differences between internal emigrants data from Source 5.15 and the simulation scenarios SC2-SC4 for male and female persons.

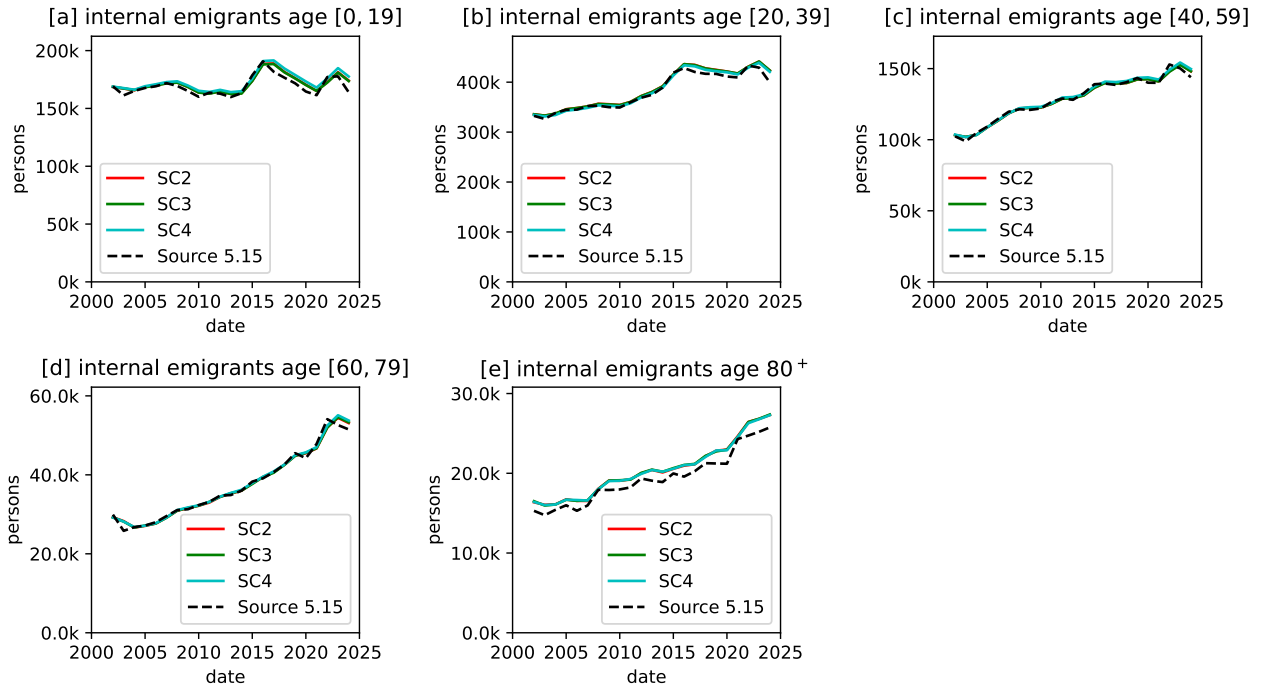


Figure 66: Comparison between internal emigrants data from Source 5.15 and the simulation scenarios SC2-SC4 for different age cohorts.

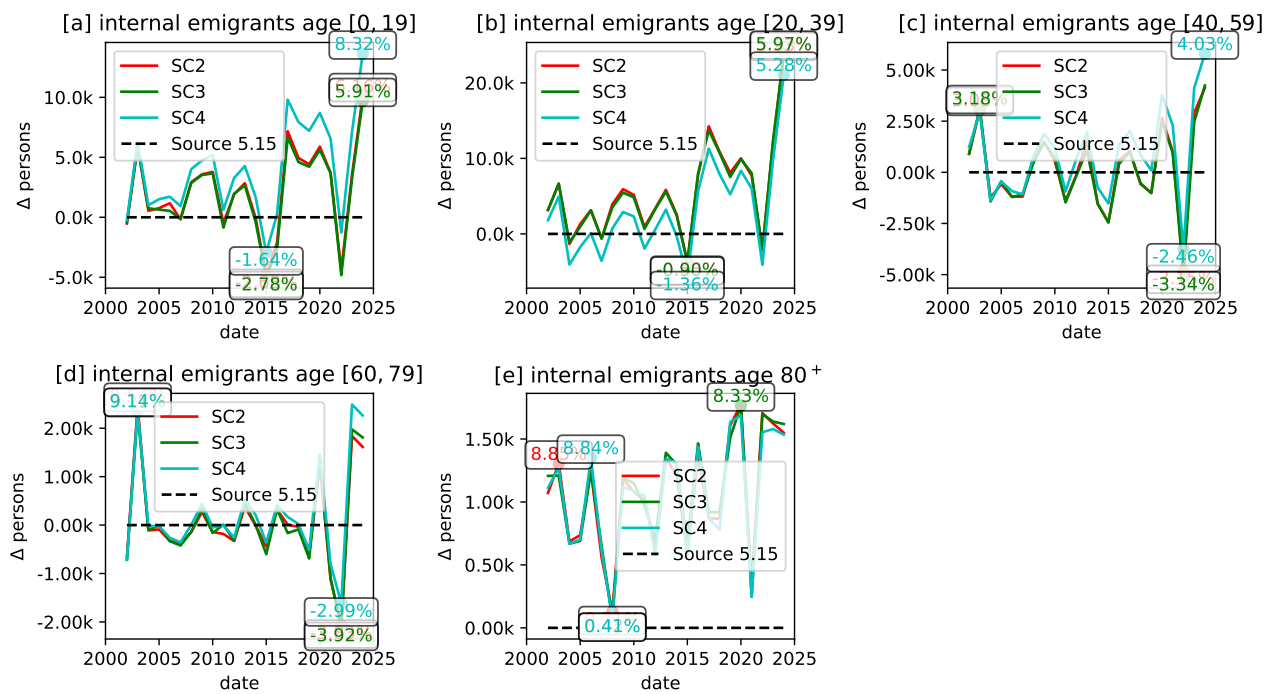


Figure 67: Differences between internal emigrants data from Source 5.15 and the simulation scenarios SC2-SC4 for different age cohorts.

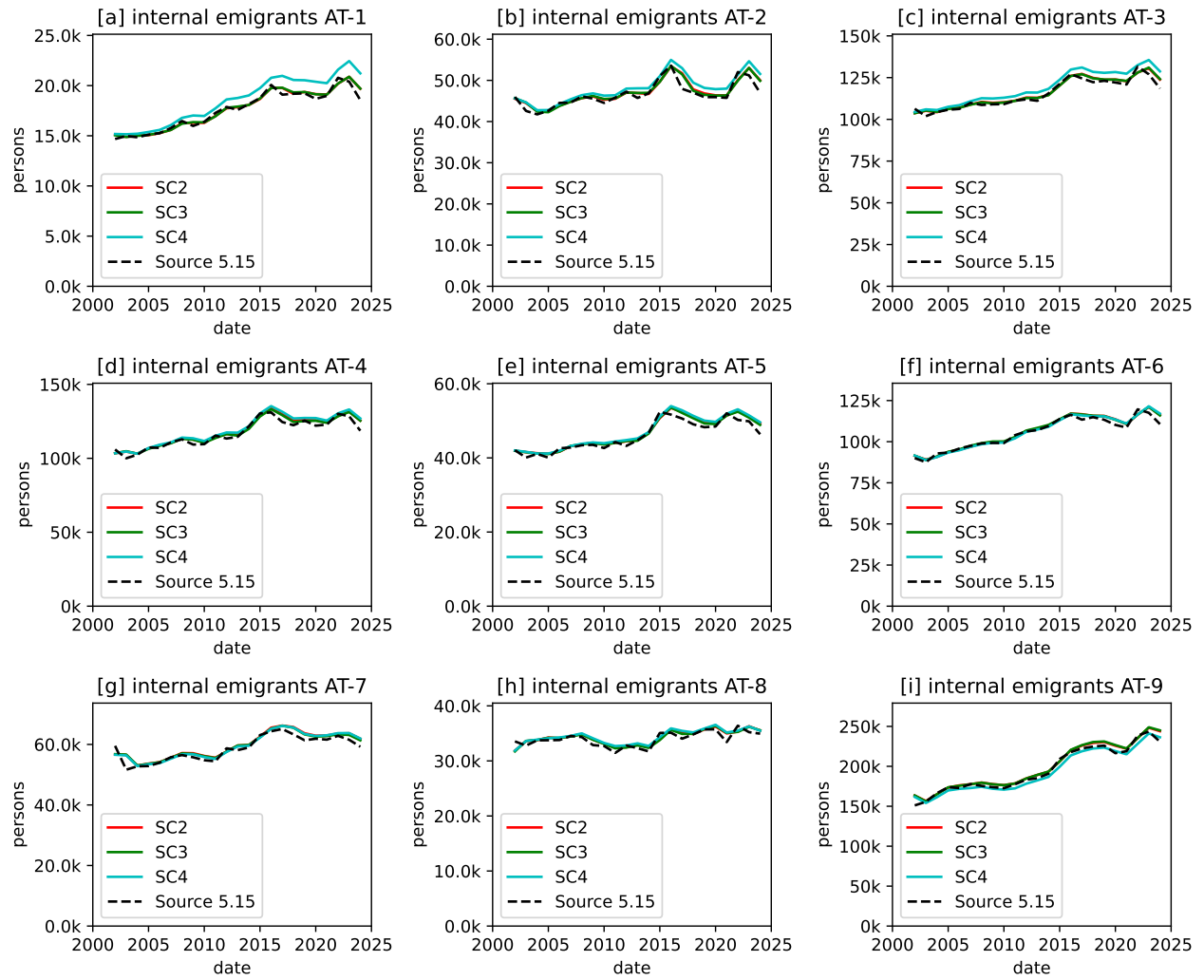


Figure 68: Comparison between internal emigrants data from Source 5.15 and the simulation scenarios SC2-SC4 for the nine federal-states of Austria.

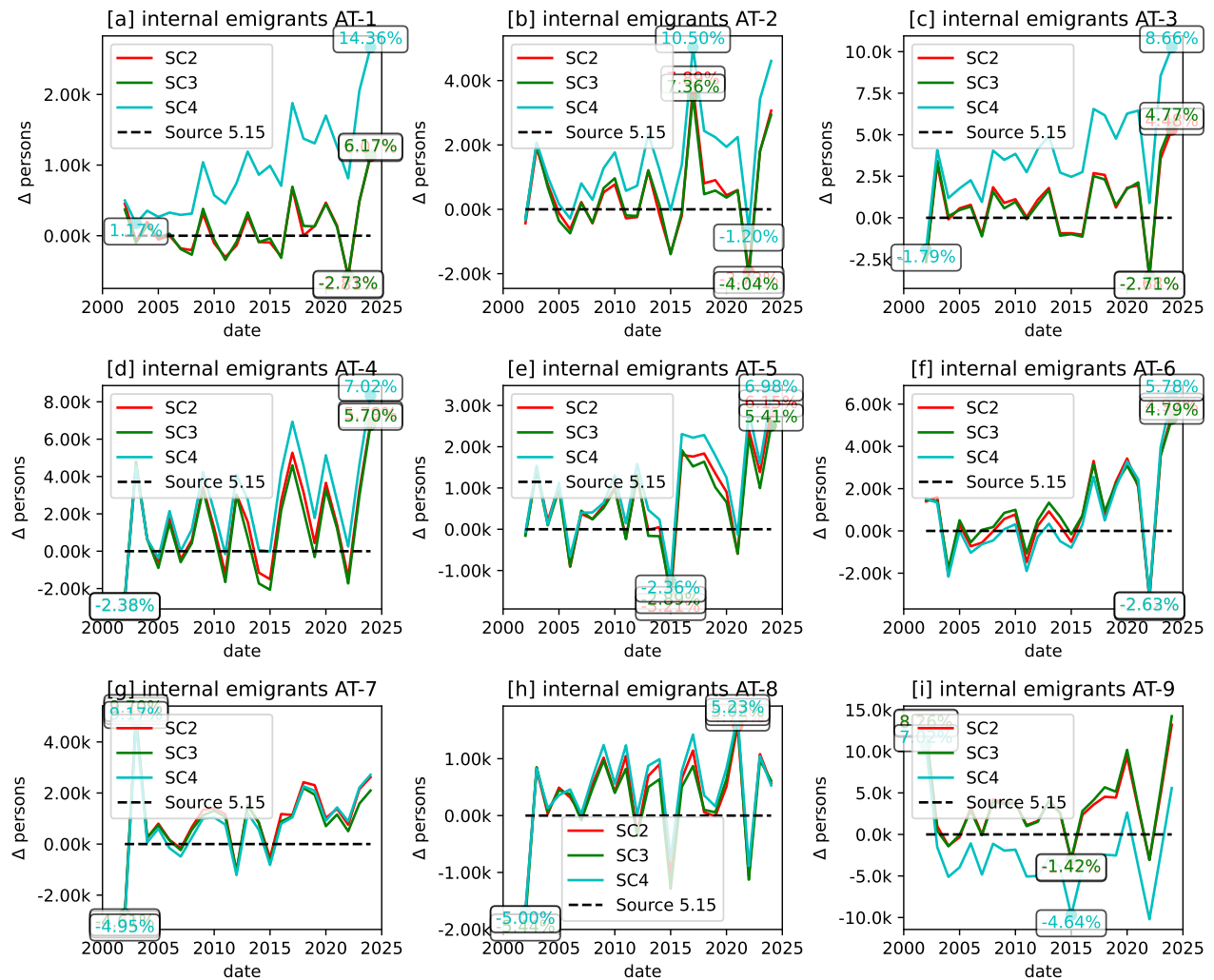


Figure 69: Differences between internal emigrants data from Source 5.15 and the simulation scenarios SC2-SC4 for the nine federal-states of Austria.

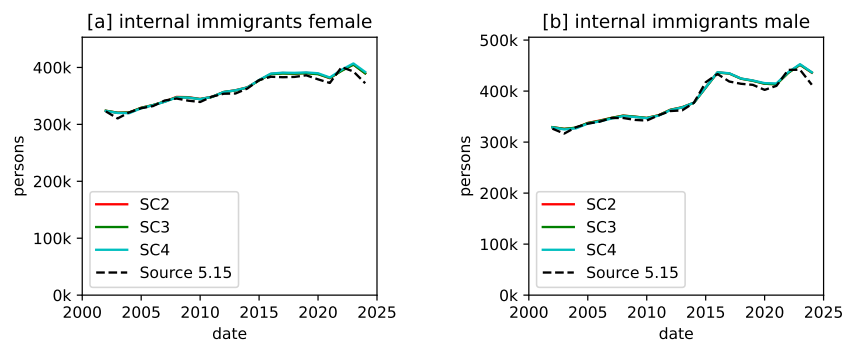


Figure 70: Comparison between internal immigrants data from Source 5.15 and the simulation scenarios SC2-SC4 for male and female persons.

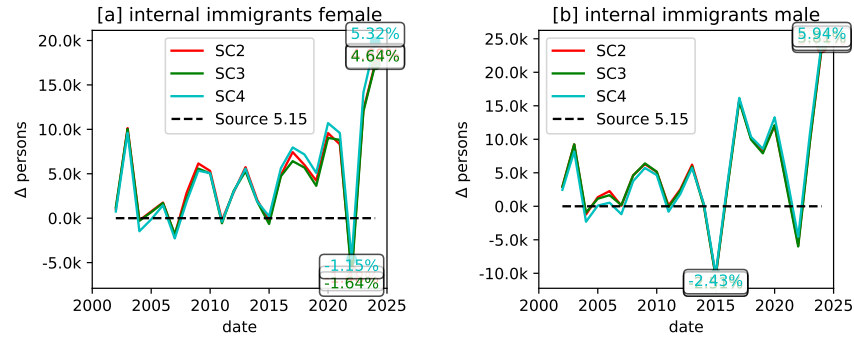


Figure 71: Differences between internal immigrants data from Source 5.15 and the simulation scenarios SC2-SC4 for male and female persons.

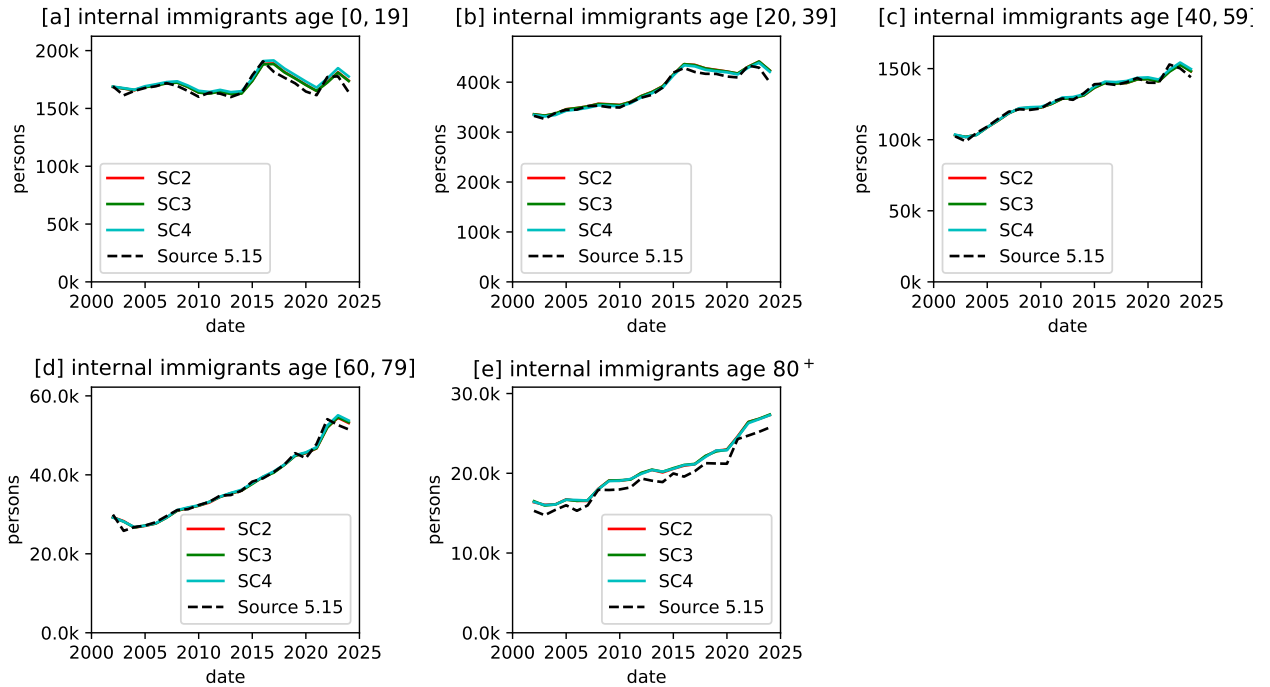


Figure 72: Comparison between internal immigrants data from Source 5.15 and the simulation scenarios SC2-SC4 for different age cohorts.

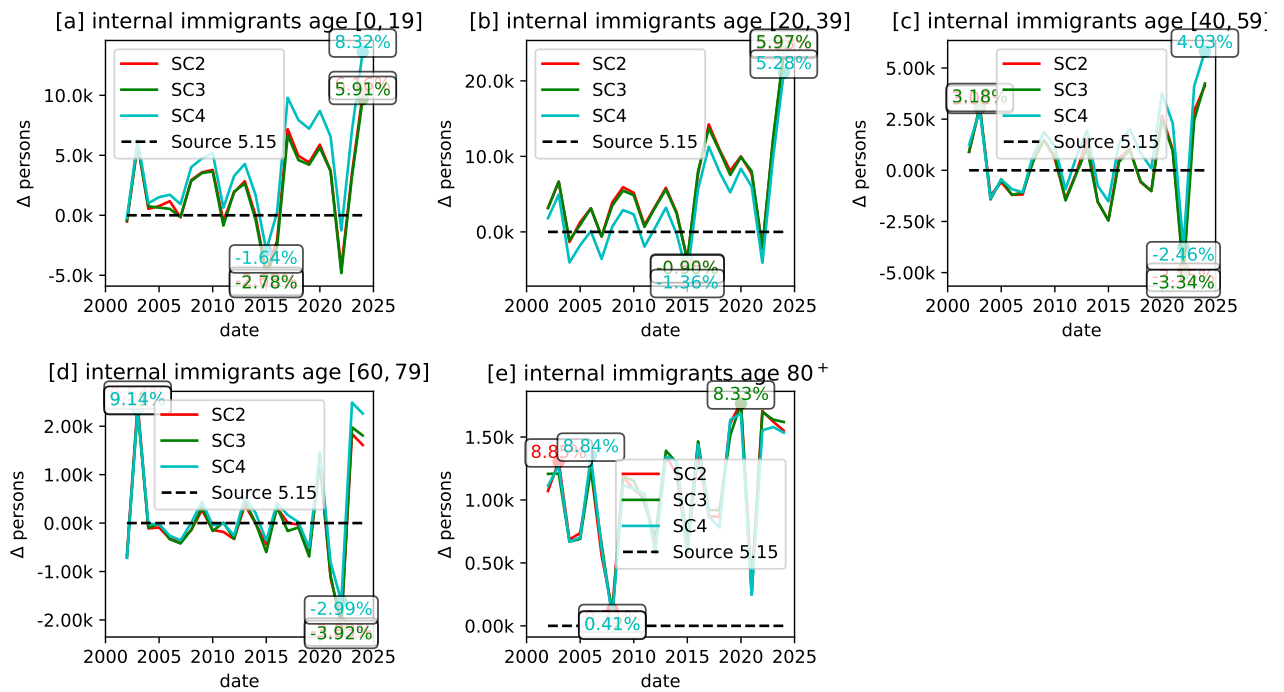


Figure 73: Differences between internal immigrants data from Source 5.15 and the simulation scenarios SC2-SC4 for different age cohorts.

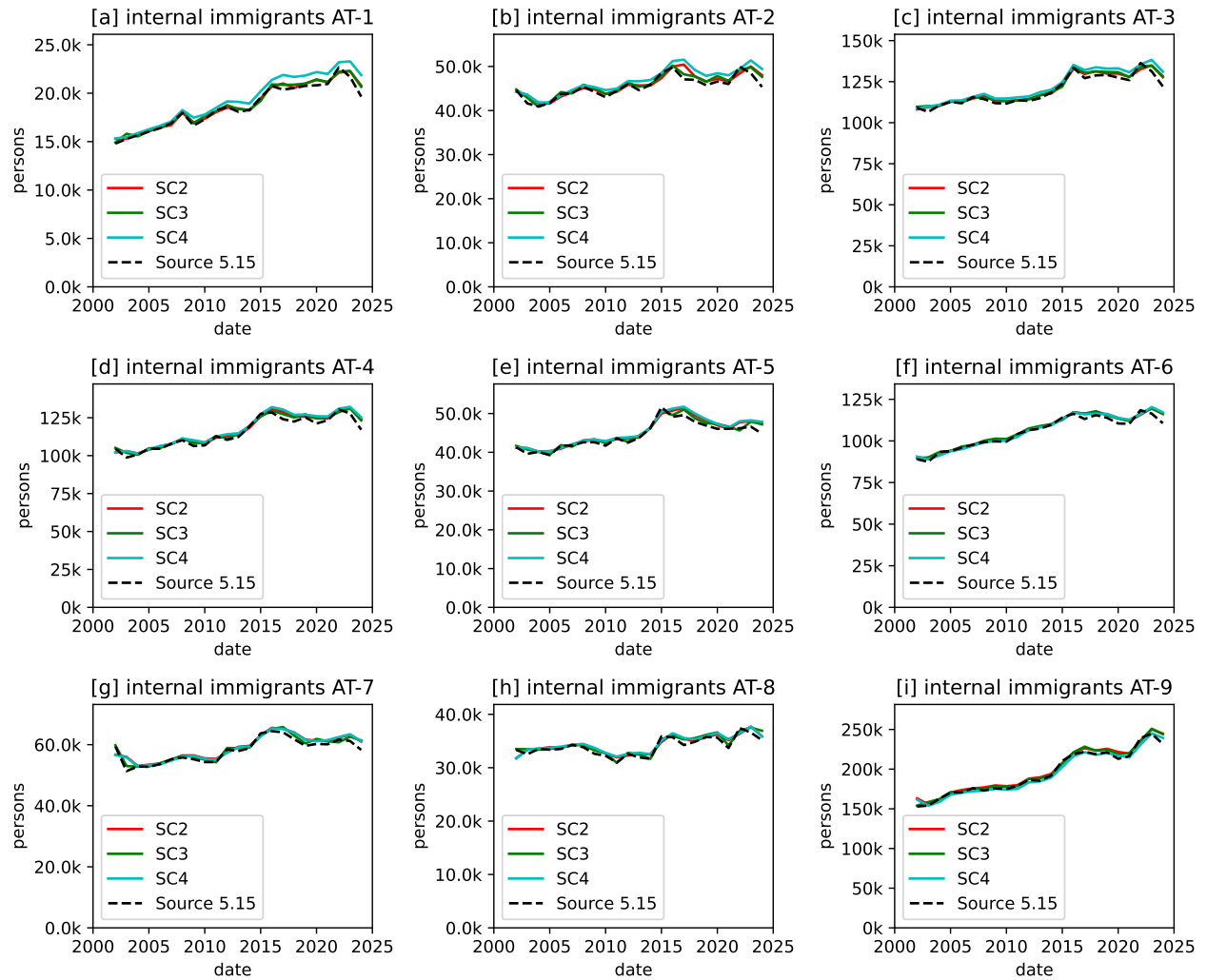


Figure 74: Comparison between internal immigrants data from Source 5.15 and the simulation scenarios SC2-SC4 for the nine federal-states of Austria.

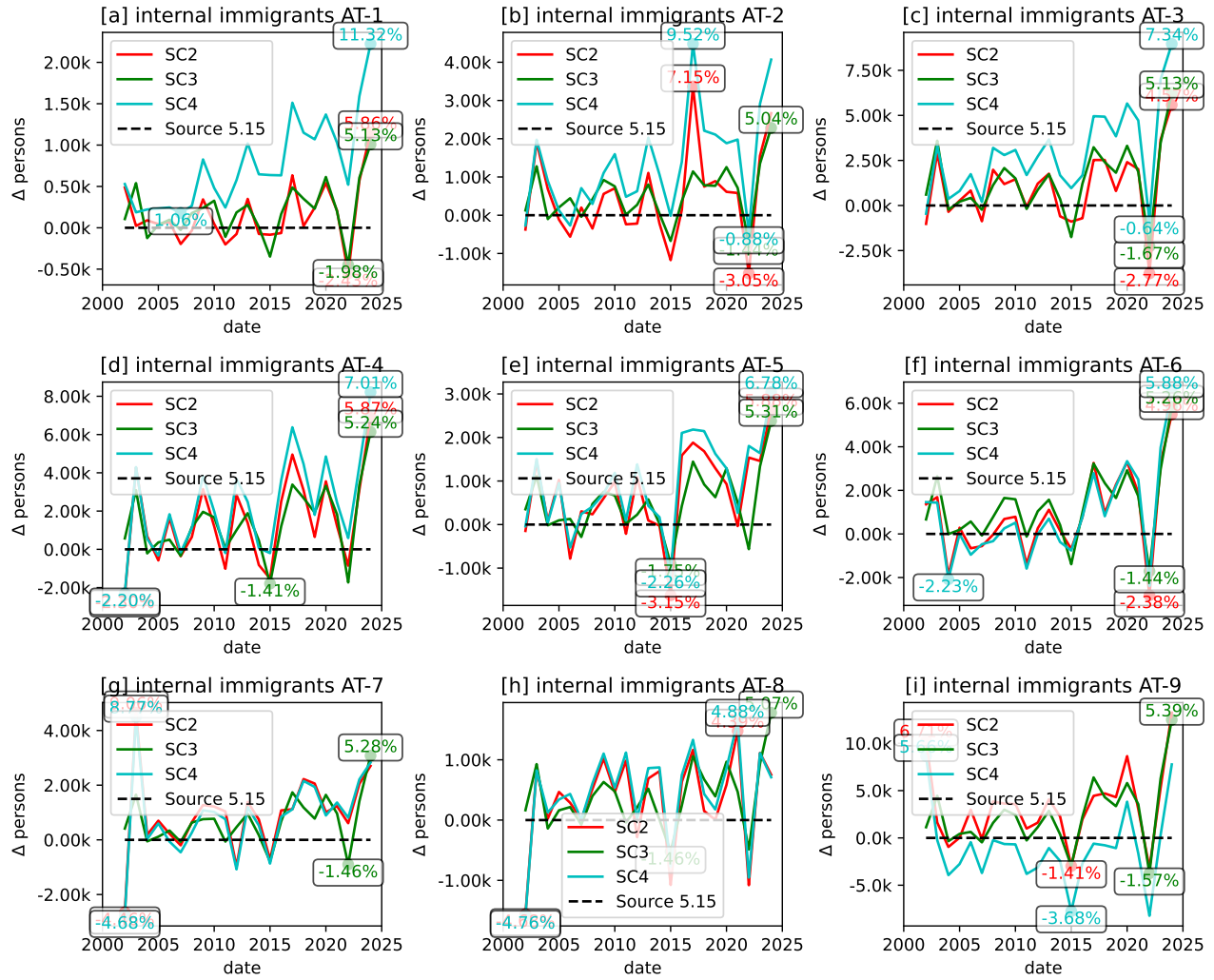


Figure 75: Differences between internal immigrants data from Source 5.15 and the simulation scenarios SC2-SC4 for the nine federal-states of Austria.